INTERNATIONAL STANDARD

ISO/IEC 19795-2

First edition
2007-02-01

Information technology — Biometric performance testing and reporting —

Part 2: Testing methodologies for technology and scenario evaluation

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

Partie 2: Méthodologies d'essai pour l'évaluation des technologies et du scénario

Reference number
ISO/IEC 19795-2:2007(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword iv

Introduction v

1 Scope 1

2 Conformance 1

3 Normative references 1

4 Terms and definitions 2
4.1 Biometric data 2
4.2 Components of a biometric system 2
4.3 User interaction with a biometric system 2
4.4 Performance measures 3

5 Overview of technology evaluations and scenario evaluations 3

6 Technology evaluation 6
6.1 Test design 6
6.2 Assembling an appropriate test corpus 8
6.3 Performance measurement 11
6.4 Reporting 16

7 Scenario evaluation 18
7.1 Test design 18
7.2 Test crew 23
7.3 Performance measurement 24
7.4 Reporting 26

8 Other issues applicable to technology and scenario evaluations 29
8.1 Parties to a test 29
8.2 Fairness 29
8.3 Basis for inclusion of test systems 29
8.4 Use of Frequently Asked Questions 30
8.5 Legal issues 30
8.6 Release of test source code 30
8.7 Supplier comment on test report 30

Annex A (informative) Phases and activities for primary technology test types 31

Annex B (informative) Relationship between presentations, attempts, and transactions 37

Annex C (informative) Reporting effort levels 38

Annex D (informative) Client-server testing 40

Annex E (informative) Comparing results across systems in multi-system tests 41

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. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC 19795-2 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 37, *Biometrics*.

ISO/IEC 19795 consists of the following parts, under the general title *Information technology — Biometric performance testing and reporting*:

— *Part 1: Principles and framework*

— *Part 2: Testing methodologies for technology and scenario evaluation*

The following parts are under preparation:

— *Part 3: Modality-specific testing* [Technical Report]

— *Part 4: Performance and interoperability testing of data interchange formats*

— *Part 5: Performance of biometric access control systems*

Introduction

This part of ISO/IEC 19795 addresses two specific biometric performance testing methodologies: technology and scenario evaluation. The large majority of biometric tests are of one of these two generic evaluation types. Technology evaluations evaluate enrolment and comparison algorithms by means of previously collected corpuses, while scenario evaluations evaluate sensors and algorithms by processing of samples collected from Test Subjects in real time. The former is intended for generation of large volumes of comparison scores and candidate lists indicative of the fundamental discriminating power of an algorithm. The latter is intended for measurement of performance in modeled environments, inclusive of Test Subject-system interactions.

This part of ISO/IEC 19795 builds on requirements and best practices specified in ISO/IEC 19795-1, which addresses specific philosophies and principles that can be applied over a broad range of test conditions.

This part of ISO/IEC 19795 is meant to provide biometric system developers, deployers and end users with mechanisms for design, execution and reporting of biometric performance tests in a fashion that allows meaningful benchmarking of biometric performance within and across technologies, usage scenarios and environments.

Information technology — Biometric performance testing and reporting —

Part 2:
Testing methodologies for technology and scenario evaluation

1 Scope

This part of ISO/IEC 19795 provides requirements and recommendations on data collection, analysis and reporting specific to two primary types of evaluation: technology evaluation and scenario evaluation.

This part of ISO/IEC 19795 specifies requirements in the following areas:

— development and full description of protocols for technology and scenario evaluations;

— execution and reporting of biometric evaluations reflective of the parameters associated with biometric evaluation types.

2 Conformance

A test shall claim conformance to either the technology evaluation or scenario evaluation clauses of this part of ISO/IEC 19795.

The set of clauses to which a scenario test shall conform differs from the set of clauses to which a technology test shall conform. In addition, the set of clauses to which an identification-system test shall conform differs from the set of clauses to which a verification-system test shall conform. To conform to this part of ISO/IEC 19795, an evaluation shall conform to clauses of this part of ISO/IEC 19795 as shown in Table 1.

Table 1 — Conformance for evaluation methodologies and comparison types

Evaluation methodology	Comparison type	Required clauses
Technology or scenario	Identification or verification	Clauses 5 and 8
Technology	Identification	All of Clause 6, except 6.3.3
Technology	Verification	All of Clause 6, except 6.3.4
Scenario	Identification	All of Clause 7, except 7.3.4
Scenario	Verification	All of Clause 7, except 7.3.5

3 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 19795-1, *Information technology — Biometric performance testing and reporting — Part 1: Principles and framework*

4 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 19795-1:2006 and the following apply.

4.1 Biometric data

4.1.1
biometric reference
⟨template, model⟩ user's stored reference measure based on features extracted from enrolment samples

4.2 Components of a biometric system

4.2.1
feature extractor
apparatus that extracts features from a sample

4.2.2
biometric reference generator
apparatus that transforms a sample into a biometric reference

4.3 User interaction with a biometric system

4.3.1
acclimatization
reduction, over the course of an evaluation, in a temporal condition of a biometric characteristic that may impact the ability of a sensor to process a sample

4.3.2
effort level
number of presentations, attempts or transactions needed to successfully enrol or match in a biometric system

4.3.3
enrolment attempt
submission of one or more biometric samples for a Test Subject for the purpose of enrolment in a biometric system

NOTE 1 One or more enrolment attempts may be permitted or required to constitute an enrolment transaction. An enrolment attempt may comprise one or more enrolment presentations.

NOTE 2 See Annex B for illustration of the relationship between presentation, attempt and transaction.

4.3.4
enrolment attempt limit
maximum number of attempts, or the maximum duration, a Test Subject is permitted before an enrolment transaction is terminated

4.3.5
enrolment presentation
submission of an instance of a biometric characteristic for a Test Subject for the purpose of enrolment

NOTE One or more enrolment presentations may be permitted or required to constitute an enrolment attempt. An enrolment presentation may or may not result in an enrolment attempt.

4.3.6
enrolment presentation limit
maximum number of presentations, or the maximum duration, a Test Subject is permitted before an enrolment attempt is terminated

4.3.7
guidance
direction provided by an Administrator to a Test Subject in the course of enrolment or recognition

NOTE Guidance is separate from feedback provided by a biometric system or device in the course of enrolment or recognition, such as audible or visual presentation queues.

4.3.8
habituation
degree of familiarity a Test Subject has with a device

NOTE A Test Subject having substantial familiarity with a biometric device, such as that gained in the course of employment, is referred to as a habituated Test Subject.

4.3.9
comparison attempt
submission of one or more biometric samples for a Test Subject for the purpose of comparison in a biometric system

4.3.10
comparison attempt limit
maximum number of attempts, or the maximum duration, a Test Subject is permitted before a comparison transaction is terminated

4.3.11
comparison presentation
submission of an instance of a single biometric characteristic for a Test Subject for the purpose of comparison

NOTE One or more comparison presentations may be permitted or required to constitute a comparison attempt. A comparison presentation may or may not result in a comparison attempt.

4.3.12
comparison presentation limit
maximum number of presentations, or the maximum duration, a Test Subject is permitted before a comparison attempt is terminated

4.4 Performance measures

4.4.1
failure at source rate
proportion of samples discarded from the corpus either manually or by use of an automated biometric system prior to use in a technology evaluation

EXAMPLE A proportion of images collected in a face data collection effort may be discarded due to lack of a face in the image.

5 Overview of technology evaluations and scenario evaluations

This standard addresses two types of evaluation methodologies: technology evaluations and scenario evaluations. A test report shall state whether it presents results from a technology evaluation, a scenario evaluation, or an evaluation that combines aspects of both technology and scenario evaluations.

Technology evaluation is the offline evaluation of one or more algorithms for the same biometric modality using a pre-existing or specially-collected corpus of samples. The utility of technology testing stems from its separation of the human-sensor acquisition interaction and the recognition process, whose benefits include the following:

— Ability to conduct full cross-comparison tests. Technology evaluation affords the possibility to use the entire testing population as claimants to the identities of all other members (i.e. impostors) and this allows estimates of false match rates to be made to on the order of one in N^2, rather than one in N.

— Ability to conduct exploratory testing. Technology evaluation can be run with no real-time output demands, and is thus well-suited to research and development. For example, the effects of algorithmic improvements, changes in run time parameters such as effort levels and configurations, or different image databases, can be measured in, essentially, a closed-loop improvement cycle.

— Ability to conduct multi-instance and multi-algorithmic testing. By using common test procedures, interfaces, and metrics, technology evaluation affords the possibility to conduct repeatable evaluations of multi-instance systems (e.g. three views of a face) and multi-algorithmic (e.g. supplier A and supplier B) performance, or any combination thereof.

— Provided the corpus contains appropriate sample data, technology testing is potentially capable of testing all modules subsequent to the human-sensor interface, including: a quality control and feedback module(s), signal processing module(s), image fusion module(s) (for multi-modal or multi-instance biometrics), feature extraction and normalization module(s), feature-level fusion module(s), comparison score computation and fusion module(s), and score normalization module(s).

— The nondeterministic aspects of the human-sensor interaction preclude true repeatability and this complicates comparative product testing. Elimination of this interaction as a factor in performance measurement allows for repeatable testing. This offline process can be repeated *ad infinitum* with little marginal cost.

— If sample data is available, performance can be measured over very large target populations, utilizing samples acquired over a period of years.

NOTE 1 Collecting a database of samples for offline enrolment and calculation of comparison scores allows greater control over which samples and attempts are to be used in any transaction.

NOTE 2 Technology evaluation will always involve data storage for later, offline processing. However, with scenario evaluations, online transactions might be simpler for the tester — the system is operating in its usual manner and storage of samples, although recommended, is not absolutely necessary.

Scenario evaluation is the online evaluation of end-to-end system performance in a prototype or simulated application. The utility of scenario testing stems from the inclusion of human-sensor acquisition interaction in conjunction with the enrolment and recognition processes, whose benefits include the following:

— Ability to gauge impact of additional attempts and transactions on system's ability to enrol and recognize Test Subjects.

— Ability to collect throughput results for enrolment and recognition trials inclusive of presentation and sample capture duration.

NOTE 3 In online evaluations, the Experimenter may decide not to retain biometric samples, reducing storage requirements and in certain cases ensuring fidelity to real-world system operations. However, retention of samples in online tests is recommended for auditing and to enable subsequent offline analysis.

NOTE 4 Testing a biometric system will involve the collection of input images or signals, which are used for biometric reference generation at enrolment and for calculation of comparison scores at later attempts. The images/signals collected can either be used immediately for an online enrolment, verification or identification attempt, or may be stored and used later for offline enrolment, verification or identification.

Information on differences between technology and scenario evaluations is presented in Table 2.

Table 2 — Distinctions between technology and scenario evaluations

	Technology Evaluations	Scenario Evaluations
What is tested	Biometric component (comparison or extraction algorithm).	Biometric system.
Objective of test	Measure performance of algorithm(s) on a standardized corpus.	Measure performance of end-to-end system in simulated application.
Ground truth	Known associations between data samples and source of samples, subject to data collection errors and intersections in merged data sets.	Known associations between system decisions and independently recorded sources of presented samples, subject to data collection errors and tester failure to note unwanted Test Subject behaviour.
Test Subject behaviour controlled by Experimenter	Not applicable during testing. May be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled.	Controlled (unless Test Subject behaviour is an independent variable).
Test Subject has real-time feedback of the result of attempt	No.	Yes.
Repeatability of results	Repeatable.	Quasi-repeatable (if test environment conditions and human factors variables are controlled).
Control of physical environment	May be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled.	Controlled and/or recorded.
Test Subject interaction recorded	Not applicable during testing. May be recorded when biometric data recorded.	Recorded.
Typical results reported	Relative robustness of biometric components or versions of components (e.g., comparison or extraction algorithms). Determine critical performance factors.	Relative robustness of biometric systems. Determine critical performance factors. Measure simulated performance.
Typical metrics	Most error rates. Not end-to-end throughput. Good for large-scale identification system performance where difficult to assemble large test crew.	Predicted end-to-end throughput. False match rate, false non-match rate. Failure to acquire, failure to enrol. GFAR, GFRR.
Constraints	Appropriate test database, e.g., gathered with one or more sensors, the identity of which may or may not be known.	Operational, instrumented system.
Human test population	Recorded.	Real time participation.

NOTE 5 Although in some cases there may be exceptions to the entries in this table, these are the main distinctions.

6 Technology evaluation

6.1 Test design

6.1.1 Goals

An evaluation shall be designed to evaluate a system's enrolment, acquisition and matching functions on the target application.

6.1.2 Application realism

If the test intends to evaluate performance within an application or concept of operations, the test shall be designed and executed so that it mimics the functional (input to output) and procedural (e.g. enrolment or verification processes) aspects of such an application or concept of operations.

EXAMPLE If several images are typically gathered to constitute an enrolment transaction in a real-world enrolment attempt, technology test design should follow a similar process.

For testing purposes, the implementations under test should, if possible, return the comparison score of each comparison attempt.

6.1.3 Determination of appropriate performance measures

Experimenters shall determine which performance measures are applicable to their evaluation, in addition to those listed at clause 6.3.

Test design shall ensure that all required metrics can be generated.

Experimenters shall determine and report on the type(s) of comparison functionality to be incorporated within the technology test. One or more of the following types of comparison shall be specified:

a) verification

b) open-set identification

c) closed-set identification

The rationale for selection of one or more types of comparison functionality within a technology test shall be reported. The comparison functionality evaluated should be applicable to the algorithm in question, such that systems designed to conduct a specific type of comparison such as watchlist identification are tested in a fashion that generates the appropriate type of result.

NOTE Formulae for error rate calculation are provided in ISO/IEC 19795-1:2006, Clause 7.

6.1.4 Implementation primacy

The test plan shall not dictate the method(s) by which the biometric recognition system implements its functions. It is the responsibility of the biometric recognition implementation to perform its functions in its own way.

NOTE The separation of what a tested biometric system does from how it does it is the fundamental construct for allowing offline testing to be done. It is primarily useful in establishing the responsibilities of tester versus supplier. The system under test should be regarded wherever possible as a black box: Its essential function is to render decisions on input samples. The internal details of how this occurs may be proprietary, but in any case, are of no concern to the tester. This construction facilitates the testing of arbitrary biometric samples.

EXAMPLE 1 If a fingerprint is sampled at 1000 dpi, and a test device is known to process only half that, then the tester should a) not execute the down-sampling on the basis that the method for doing so is non-trivial, and b) apprise the supplier of the need to handle the down-sampling internally.

EXAMPLE 2 A set of simultaneously acquired non-frontal face images could be processed by a biometric system and device in at least three ways: selection of the best image; fusion of all images; or stereoscopic synthesis of a 3D model. In any case the biometric system or device decides.

EXAMPLE 3 Most automated fingerprint identification system (AFIS) machines (i.e. those identifying multi-fingerprint records) implement some binning mechanism to partition the database according to some criterion (most simply, the Henry class) and to search only that part of the database of the same category as a user or impostor sample, thereby obtaining throughput benefits, but possibly incurring accuracy losses. Such a tradeoff is achieved by the supplier setting internal binning parameters, and measured by conducting full-scale repeats of the test for each configuration.

EXAMPLE 4 In a study seeking to demonstrate the utility of multiple fingers' prints in a recognition system, the tester should not pass separate samples through the device and perform subsequent score-level fusion, but should instead compose all the imagery as a sample (e.g. in an American National Standards Institute [ANSI]-National Institute of Standards and Technology [NIST] record, or a Common Biometric Exchange Formats Framework [CBEFF] wrapped instance of 19794-2) so as to let the biometric device perform the fusion internally. See ISO/IEC 19794-2, *Information technology — Biometric data interchange formats — Part 2: Finger minutiae data* for further information on CBEFF wrapped instances. See ANSI/NIST-ITL 1-2000 NIST Special Publication 500-245 for information on ANSI-NIST records.

6.1.5 Policies on disclosure of information to suppliers

The tester shall formulate policies before testing begins that govern what information will be disclosed to the suppliers a) before test equipment is configured, shipped or installed, and b) at execution time.

6.1.6 Non-interchangeability of identification and verification attempts

Comparison scores that result from a one-to-many identification search shall not be presented as the results of verification attempts without justification.

NOTE 1 The principle of operational realism indicates that performance shall be estimated from outcomes of attempts (i.e. rejects and acceptances). A verification system shall be evaluated on the results of a sequence of user claims of identity. Likewise an identification system shall be tested over one-to-many searches. Even in the case that a one-to-many search produces a full candidate list, the candidate list is atomic, meaning that it should not be regarded as the result of N verification attempts (to be used in the computation of verification performance).

NOTE 2 The non-equivalency of a single identification attempt and N one-to-one verifications arises because verification can be improved by comparing the user sample with additional hidden samples in a process known as cohort-normalization. The method adjusts the raw single comparison score in order to drive down false accept rates by effectively setting user-specific thresholds. The method trades off performance for throughput because the additional comparisons mean 1:1 verification incurs the expense of 1:M, where M is the size of the hidden biometric reference set.

NOTE 3 The use of cohort normalization is properly conducted internally to the device, making private use of an internally selected enrolled population.

6.1.7 Acknowledgement of models

If a model, approximation or prediction of identification performance is reported in place of, or in addition to, an empirical trial, the model shall be verified to the extent possible with the available data and fully documented.

6.1.8 Sequential use

The test plan shall define the order of use of the test data. This order shall be appropriate to the application. The implementation should process the test data in this sequence.

NOTE 1 Transactions are ordinarily executed separately. Therefore the implementation would need to complete one transaction before commencing the next transaction.

NOTE 2 The majority of biometric applications involve the sequential and separate use of the biometric systems or device by individuals, subsequent, in the case of genuine users, to prior enrolment.

NOTE 3 Certain identification tasks may not be sequential. For example batch identification of all persons in a closed room is easier because it reduces to the linear assignment problem.

6.1.9 Pre-test procedures

6.1.9.1 Installation and validation of correct operation

The test organization shall take steps to ensure that the hardware/software is installed and configured appropriately and shall verify that the system is operating correctly.

NOTE Installation, configuration, and verification of system operations may involve supplier(s).

6.1.9.2 Data preparation

Data preparation shall ensure that Test Subject-identifying information and any associated metadata that would not ordinarily be available to the application (e.g. sex, age) is expunged from the samples. Otherwise a supplier might deduce the true identities to game the test.

6.1.10 Generic test execution sequence

The following is a generic description of the sequence of technology test execution:

— Enrolment samples are converted to biometric references and may be stored in a linear collection.

— Identification and verification samples are converted to sample features.

— Verification attempts are a direct comparison of sample features to a biometric reference.

— Closed-set identification attempts are a search of the enrolled population intended to return the user's identifier.

— Open-set identification attempts search the enrolled database and

  — return one or more identities;

  — return a null identity, indicating Test Subject is not found in the enrolled database.

NOTE 1 The above functionality may be implemented at the API level, or by scripting around executables.

NOTE 2 Annex A describes test execution sequences for specific types of technology tests.

6.2 Assembling an appropriate test corpus

6.2.1 General

Technology evaluation is designed to evaluate one or more biometric algorithms for enrolment and comparison performance. Technology test planning is contingent on the type of data an Experimenter wishes to generate.

6.2.2 Unique enrolment

All corpus samples should correspond to real people. An evaluation design should not intentionally enrol different samples from the same individual as if they were from different individuals. For tests in which each

identity corresponds to a different individual, the testing organization shall report processes implemented to ensure this.

If it is possible that an individual has multiple identities in the corpus, the corpus may be "cleaned" to reconcile such instances if practical. Otherwise the test should proceed under the assumption that each identity corresponds to a different individual.

NOTE 1 Biometric systems are intended to uniquely identify single individuals. If more than one image or signal is available for an individual it should be encapsulated as a single sample and used for enrolment or comparison.

NOTE 2 Populating an identification system with more than one sample per individual (from within one or more modalities) and then regarding the enrolments as nominally separate is a deprecated practice for the following reasons:

Identification entails a search through enrolled samples and generation of a candidate list. When multiple samples are separately enrolled a score-level fusion of each user's samples using the max criterion is implied because the largest scored entry wins. Even if the number of samples per person is equal for all persons, the practice is deprecated because it is the supplier's responsibility to combine each individual's samples in what it deems the best way.

Error metrics that depend on the size of the enrolled population, N, will be incorrect if the size of the enrolled population is not the number of separate individuals.

NOTE 3 An evaluation that seeks to investigate the effect of multiple (separate) enrolled biometric references per Test Subject is exempted from this clause, provided that this is documented in both the test plan and the test report.

6.2.3 Recurrence of data acquisition

Depending on the Experimenter's level of access to the test population, each Test Subject may be able to provide data multiple times over the course of multiple visits. The number of transactions and visits can be maximized in order to enable granular measurement of biometric reference aging effects though this will also be informed by habituation effects.

6.2.4 Test Subject identification

The Experimenter shall report information related to Test Subject identification, including at a minimum the following:

a) Types of identifiers used to identify Test Subjects

b) Amount and type of personal data collected

6.2.5 Provision of non-biometric information

If available in the corpus, metadata normally available to a deployed system shall be provided to the system(s) under test. The test report shall state the names and types of any metadata variables that were made available to the systems under test.

EXAMPLE Such data may be sensor-specific (e.g. sensor settings), environmental (e.g. temperature, humidity), Test Subject-specific (e.g. gender, age), or any other germane information.

NOTE Technology testing is unable to incorporate multiple aspects of real-world biometric operations, but evaluation design should not exclude aspects of real-world biometric operations unnecessarily.

6.2.6 Representativeness of corpus

Evaluation design shall consider, and a test report shall document, whether the data in the test corpus is appropriate for the goals of the test or the applications of interest.

If data is acquired under the supervision or control of the test organization, information pertaining to Experimenter-Test Subject interaction shall be recorded in the areas of acclimatization, training, habituation, and guidance.

NOTE 1 The utility of technology evaluation in producing predictive estimates of deployed performance is predicated upon the assumption that it is possible to consistently acquire samples from users in the same format and with the same quality as the data used in the test.

NOTE 2 Ideally, data collected for different modalities has equivalent levels of habituation, acclimatization, guidance, etc.

6.2.7 Untainted corpus

A corpus may be considered "tainted" to a greater or lesser extent if

a) any implementation supplier has had possession of the corpus;

b) any implementation supplier has provided equipment used in collecting or processing the corpus, particularly if this activity influenced the nature or quality of the corpus such as by excluding samples;

c) a system being tested has previously been tested and tuned using the corpus.

When use of a tainted corpus is unavoidable, this fact shall be documented in the test report.

Sample data should not be used in an evaluation if one or more of the participating suppliers has had possession of it. Previous testing / tuning of the system using the test corpus (in whole or in part) shall be documented in the test report.

NOTE 1 This clause is necessary because performance may be improved via gaming.

NOTE 2 Note that it is generally insufficient to trivially alter the sample to skirt the reuse prohibition. Gaming may still be possible if any identifiable trait of the previously seen samples remains.

6.2.8 Retirement of corpus

Samples should not be reused in an evaluation if one or more systems under test have been tuned on the basis of performance measured in a previous test with that data.

NOTE 1 This is most readily achievable by using sequestered data.

NOTE 2 This may be expensive in that it implies additional collection activity.

6.2.9 Corpus validation

Validation is the process whereby Test Subject data is screened for the purpose of removal of data not suitable for the purpose of the evaluation.

Validation may include checking to ensure that Test Subject data is present, that the data is in the correct format, that the correct instance has been collected, and that ground truth errors are identified.

Experimenters shall report whether Test Subject data has been validated. If data has been validated, the Experimenter shall detail the method(s) applied in validating the data. The proportion and criteria for data removal shall be reported.

EXAMPLE 1 Database quality control might be used to avoid images with bad contrast in the Test Subject data.

EXAMPLE 2 Data samples might be excluded that do not show a face in face recognition technology (e.g. no face at all or full body) or that do not show a fingerprint in a fingerprint recognition technology test (e.g. a palm print).

NOTE 1 Since some types of biometric data may be more easily validated than others, use of data validation could introduce a bias in performance results.

NOTE 2 Data removed by "corpus validation" is distinct from that discarded as "Failure at source." Sometimes a judgement call will be needed as to whether excluded data should be considered invalid or failure at source.

6.2.10 Corpus collection environment

Environmental conditions present during data collection may be known or specified. Such collection would typically be intended to measure performance under specific environmental conditions relative to baseline environmental conditions. Such controls may be established for temperature, lighting, humidity, and other factors known or suspected to impact biometric performance.

Available information pertaining to environmental conditions during corpus acquisition relevant to the modalities under evaluation should be reported, such as the following:

— temperature;

— exposure to elements;

— lighting, including type, direction, intensity;

— ambient noise;

— vibration.

If applicable, Experimenters shall report that such information was not available.

NOTE See ISO/IEC 19795-1:2006, C.2.6 for information on environmental factors that can impact performance.

6.2.11 Failure at source

Offline tests use stored biometric samples, which may have been gathered with or without a biometric system in the acquisition process. The test report shall disclose any known information about how the data was processed at any stage before use in the test. Particularly if samples were discarded, either manually or by use of an automated biometric system, then a *failure at source* rate (FAS) shall be reported.

NOTE 1 FAS may relate to a different biometric sensor or image quality assessment algorithm than the system under test.

NOTE 2 A judgment call may be needed: For example if a few legacy image samples are found to be entirely blank then these could be legitimately not counted in FAS, unless of course such samples would routinely occur in the application that the test is intended to mimic.

6.3 Performance measurement

6.3.1 Enrolment

Offline tests shall record, as the failure to enrol rate (FTE), the proportion of Test Subjects for whom an implementation elects to reject enrolment of their designated enrolment samples in the corpus. Criteria by which failure to enrol is declared shall be defined.

NOTE 1 Failure to enrol measured in a technology test is only partially representative of the failure modes possible in a live acquisition.

NOTE 2 A failure to enrol can be declared by a system for any reason. A frequent reason is that the system (as configured with its native image or signal detection and processing capability and with some quality acceptance criteria) fails to detect the needed signal on the basis of low quality.

NOTE 3 By declaring a failure to enrol, a system can attain better comparison performance. This trade-off must be accounted for by combining failure to enrol and false non-match rate to produce generalized false reject rate (GFRR).

The Experimenter shall specify the minimum number of samples required, and the maximum number of samples permitted, for successful enrolment.

For each biometric system tested, the Experimenter should calculate the following:

a) distribution of enrolment quality scores, if available;

b) failure to enrol for different demographic groups, or associated with different environmental conditions, or for other logical segments of the corpus.

6.3.2 Failure to acquire

Offline tests shall record the proportion of verification or identification attempts for which the system fails to capture or locate an image or signal of sufficient quality. This is the *failure to acquire* rate (FTA).

NOTE 1 The failure to acquire rate is the comparison-phase analogue of the enrolment-phase failure to enrol, and thus Notes 1 and 2 in Clause 6.3.1 apply analogously.

NOTE 2 The failure to acquire must be used along with false match rate to compute false accept rate.

NOTE 3 In a technology test, FTA is typically declared by an encoding or comparison component and is attributable to failure to process an attempt.

The Experimenter shall specify the minimum number of samples required, and the maximum number of samples permitted, to create sample features.

The formula for calculating FTA can be found in ISO/IEC 19795-1.

6.3.3 Verification metrics

For each verification system tested, the Experimenter shall compute the following:

a) false match rate (FMR) and false non-match rates (FNMR);

b) false reject rate (FRR) and false accept rates (FAR), unless test design is such that false accept rate and false reject rate are identical to false match rate and false non-match rate;

c) number of genuine and impostor comparisons executed;

d) for genuine Test Subjects, distribution of time lapsed between enrollment and acquisition of sample features, if available;

e) uncertainty of test results, as well as basis and formulae for estimating uncertainty.

False match rates and false non-match rates, as well as false accept and false reject rates, may be rendered in the form of a receiver operating characteristic (ROC) or detection error tradeoff (DET) curve. The number of Test Subjects and transactions used to arrive at these rates shall be computed.

NOTE For systems that return match/non-match decisions as opposed to comparison scores, performance may be reported at a single operating point on the ROC or DET.

For verification systems, the Experimenter should calculate the following:

a) distribution of comparison scores for genuine Test Subject and impostors;

b) verification results for different demographic groups, or associated with different environmental conditions, or for other logical segments of the corpus.

6.3.4 Identification metrics

For all identification systems, the Experimenter shall calculate uncertainty of test results, as well as basis and formulae for estimating uncertainty.

For closed-set identification systems, the Experimenter shall calculate the following:

a) cumulative match characteristics (CMC);

b) number of searches executed.

For open-set identification systems, the Experimenter shall calculate the following:

c) false positive identification rates (FPIR) and corresponding false negative identification rates (FNIR) (preferably over a range of thresholds);

d) binning error rate and penetration rate if binning is used.

For identification systems, the Experimenter shall calculate the following:

e) identification results for different demographic groups, or associated with different environmental conditions, or for other logical segments of the corpus.

6.3.5 Generalized error rates including failure to enrol and failure to acquire

6.3.5.1 General

The immediate output of an offline test — a set of paired (false match rate, false non-match rate) values — shall be combined with the measured values of failure to acquire and failure to enrol.

NOTE 1 Because a system can improve its false acceptance and rejection performance by abstaining from processing low quality samples, it is necessary to combine failure to acquire and failure to enrol rates with the measured false match rate and false non-match rate to produce the final statement of performance.

If FTE and FTA are known to be zero this fact should be noted. In this case, GFAR and GFRR for single-attempt transactions do not differ from FMR and FNMR. If FTE or FTA are known to be non-zero, false accept rate and false reject rate should be computed, and thereby differentiated from, false match rate and false non-match rate.

NOTE 2 In certain tests, samples that resulted in failure to acquire or failure to enrol may be released to suppliers for further study.

NOTE 3 For systems that return match/non-match decisions as opposed to comparison scores, performance may be reported at a single operating point on the ROC or DET.

NOTE 4 By declaring very many failures to enrol or to acquire, an implementation under test may achieve low GFAR values, but will thereby increase GFRR.

6.3.5.2 Single-attempt transactions

For each implementation under test, the experimenter shall determine the generalised FAR (GFAR) for single-attempt transactions and the generalised FRR (GFRR) for single-attempt transactions.

In cases where transactions consist of single attempts, generalised false accept rate may be computed as the proportion of impostors who are acquired and matched at some operating point threshold, t:

GFAR(t) = (1-FTA) FMR(t) (1-FTE)

Similarly the generalised false reject rate is that proportion of genuine users who are either unable to be acquired during use, or who can acquire but can't enrol, or who can enrol, be acquired and are falsely rejected, at some operating point threshold, t:

GFRR(t) = FTA + (1-FTA) FTE + (1-FTA) (1-FTE) FNMR(t)

The given formulas for GFAR and GFRR hold only in the special case of n=1, where n is the number of attempts allowed in a transaction.

NOTE 1 Different formulae may be needed if it is acceptable to not enrol certain individuals.

NOTE 2 Explicit failure to enrol and failure to acquire measurements can be avoided by specifying that all verification comparisons will result in a comparison score. A supplier can satisfy this requirement by internally recording a failure-to-enrol or failure to acquire condition and report suitably low values when such a biometric reference is used in a one-to-one comparison. This method correctly includes failure to enrol and acquire in the DET characteristic.

NOTE 3 Alternatively, GFAR and GFRR can also be determined by:

— including failed (neither accept nor reject) impostor transactions and impostor transactions for individuals whose enrolment failed in the total number of impostor transactions,

— including failed (neither accept nor reject) genuine transactions and genuine transactions for individuals whose enrolment failed in the total number of genuine transactions, and

— counting failed (neither accept nor reject) genuine transactions and genuine transactions for individuals whose enrolment failed as false rejects.

6.3.5.3 Multi-attempt transactions

In cases where transactions consist of multiple attempts, GFAR and GFRR calculations are more complex. Formulae for such tests should be generated on a test-specific basis.

6.3.6 Throughput performance

6.3.6.1 General

The testing organization may measure throughput suitable to the implementation under test.

If transaction time is an aspect of performance that a test intends to measure, then the experimenter shall specify a method of measuring transaction time suited to the implementation under test.

NOTE 1 Ideally, all enrolment and comparison operations would be timed.

NOTE 2 Offline testing will address only the computational components of throughput. For example in enrolment an offline test will only capture the image analysis and biometric reference creation phase of the full enrolment process while excluding ergonomic and transactional aspects of human related activities (e.g. placements on the sensor, removing eye glasses). Thus the enrolment throughput rates as measured in a technology test represent a lower bound on operational throughput rates.

6.3.6.2 Reporting throughput performance

If throughput summary statistics are calculated, the mean shall be reported. Other summary statistics such as the following may also be reported:

a) minimum;

b) maximum;

c) median;

d) standard deviation.

NOTE If populations of different sizes are enrolled (particularly in identification trials) the tester should report sufficient timing information to allow an assessment of the functional dependency on population size to be made; for example O(N) or $O(N^2)$.

6.3.6.3 Reporting comparison and throughput performance

Throughput performance is vital in technology evaluation because, in general, recognition errors can be reduced if throughput is decreased. In such cases a full statement of performance would be a DET characteristic with an additional third axis, throughput rate, allowing a deployer to select a suitable operating point. Moreover, with many biometric systems, altering the decision threshold will require that more, or fewer, presentation samples to be needed for successful recognition of a genuine attempt, thereby affecting throughput.

6.3.6.4 Measuring biometric reference generation and sample feature extraction timing

Some systems are asymmetric in that generation of biometric references uses different kinds of samples and different algorithmic processes than those used in feature extraction for verification and identification. Timings for biometric reference generation and for feature extraction may therefore be reported separately.

6.3.6.5 Simultaneous measurement of throughput and recognition error rates

Throughput measurement may be made during the same trial as the recognition error rates, and statistics may be bound to them.

6.3.6.6 Throughput in impostor and genuine user attempts

By the principle of operational realism user throughput may be measured for both impostor and genuine user attempts, and statistics may be reported separately.

6.3.6.7 Post-enrolment "fixing" overhead

In identification trials the tester should be cognizant that after a population is enrolled a post enrolment "fixing" overhead may be incurred. Fixing here means any process that a system invokes at the end of enrolment such is typically used to separate feature vectors for better performance.

6.3.6.8 Uniqueness searches on enrolment

Real-world enrolment of a population often embeds a uniqueness search for each new candidate enrolee. This implies that the enrolment of a population of size N, has $O(N^2)$ expense. In a technology test the uniqueness of the samples is usually assured by design, such that enrolment has O(N) cost.

The experimenter should determine whether 1:N uniqueness determination is a component of the enrollment process. This might be ascertained through measurement of enrollment durations as the size of the enrolled database increases. If such behavior is observed or known to be implemented, then the experimenter should report results accordingly. To separate the time required to determine uniqueness from the fixed enrollment time, a test design may require that implementations disable duplicate detection functionality.

NOTE A dedicated enrolment test, building a database from scratch, can be conducted.

6.3.6.9 Hardware

If an evaluation is to be implemented only in software, and several implementations are being compared, then throughput measurements may be made on fixed hardware and a fixed runtime environment, with the system being re-started between testing throughput times on each implementation.

NOTE The runtime environment here includes a fixed operating system, a fixed compilation and linking setup, and a fixed roster of background processes that should not be consuming significant resources (e.g. I/O, CPU).

6.4 Reporting

6.4.1 General

The results from the evaluation shall be presented in a test report. A test report shall document the entire test process. All reporting requirements listed in Clause 6.1 through 6.3 shall be documented in the test report. If a requirement was out of the scope or not applicable, the report shall state that the requirement was out of scope or not applicable.

EXAMPLE 1 Results cannot be reported at multiple attempt and transaction levels for biometric systems that do not permit multiple attempts or transactions to match. The report would state that the requirement to report performance at multiple effort levels was not applicable.

If a requirement was not addressed due to information being unavailable, the report shall state that the applicable data is unknown. The report shall explain why the data is unknown.

EXAMPLE 2 An organization may not be permitted to record demographic information during testing for privacy reasons. The test report would state that demographic information was not collected for privacy reasons.

The report may be released to different audiences in separate sections on different timelines.

6.4.2 System information

6.4.2.1 Specifications

For the biometric system(s) tested, the Experimenter shall report the following product information:

a) For acquisition devices: manufacturer, model, version, and firmware as applicable. If the acquisition device's core acquisition components are integrated within a third-party device, such as in the case of a fingerprint sensor incorporated into a peripheral, then manufacturer, model, version, and firmware of the core acquisition components shall be reported.

b) For comparison algorithms: provider, version, revision.

c) Specification of the platform through which for systems were tested, including but not limited to platform, OS, processing power, memory, manufacturer, database type, database size, and model.

6.4.3 Data collection processes

The Experimenter shall report the following information related to data collection:

a) methods of recording data for each performance element, including those not logged by the system(s);

b) processes for auditing and validating performance data collection, including those not logged by the system(s).

The Experimenter shall provide examples of data collection elements such as spreadsheets and logs, whether as screenshots or reproduced forms.

6.4.3.1 Architecture

For the biometric system(s) tested, the Experimenter shall report the following elements:

a) biometric data acquisition, processing, and storage architecture;

b) data flow between system components.

6.4.3.2 Outputs

For biometric system(s) tested, the Experimenter shall report each of the following:

a) types of outputs the system reports, including but not limited to comparison scores, accept/reject decisions, candidate lists, enrolment quality scores, sample quality scores;

b) range of comparison scores system is capable of reporting as well as associated supplier-specified thresholds;

c) range of enrolment quality scores system is capable of reporting as well as supplier-specified thresholds;

d) range of sample quality scores system is capable of reporting as well as supplier-specified thresholds;

e) method(s) through which outputs are provided by the system.

6.4.3.3 Method of implementation

For each biometric system tested, the Experimenter shall report system implementation information corresponding to each of the following:

a) method of biometric and platform system acquisition;

b) level of supplier involvement in system implementation.

6.4.4 Disclosure

6.4.4.1 External reporting

A test plan shall disclose what input, intermediate, and output material is to be made available to non-suppliers on what schedule and to whom.

NOTE 1 Some tests will be done in secret, in private, or with full disclosure.

NOTE 2 A full disclosure test would conceivably publish: the names of the participating suppliers, contact points, the protocol, the raw samples, the biometric references, the raw comparison scores, records of transaction times, records or anomalous system behaviour, error rates, and final conclusions.

NOTE 3 Comparative tests are commercially sensitive, so complete formal declarations of the kind and type of results to be released are vital.

6.4.4.2 Sample properties disclosure

The test plan shall specify what sample related information will be provided to the suppliers, and on what schedule. This may be modified in response to formal comments from suppliers.

NOTE Generally any information that a supplier requests that would not be known operationally should not be provided. Suppliers will legitimately inquire about the target application because their implementations may usually be configured to balance some tester-specified requirements on, for instance, numbers of enrolees, users, impostors, size of images, compression ratios, lengths of video sequences, etc.

6.4.5 Report structure

The following sections shall be incorporated in the test report:

— Executive Summary;

— Characteristics of Corpus Data;

— Specific Test Processes;

— Data Collection;

— Data Analysis;

— Record Keeping;

— Performance Results;

— Full Test Plan.

7 Scenario evaluation

7.1 Test design

7.1.1 Characteristics of simulated application

7.1.1.1 Concept of operations

The application being modeled in a scenario test shall be specified.

NOTE The application modeled in a scenario test may range from generic to specific. A generic application is one in which a limited number of parameters are specified, such as test of 1:1 authentication systems in an indoor office environment. A specific application is one in which many parameters are specified, such as test of 1:1 token-based access control systems in an indoor office environment with a non-habituated crew.

7.1.1.2 Comparison functionality

Experimenters shall determine whether verification, open-set identification, and/or closed-set identification shall be incorporated in the scenario test.

The comparison functionality evaluated shall be applicable to the prototype or simulated application.

The rationale for selection of one or more types of comparison functionality within a scenario test shall be provided.

NOTE 1 Scenario evaluations typically evaluate 1:1 systems in which a transaction is executed based on a claimed identity. Identification systems can be evaluated in scenario evaluations so long as transactions are executed in real time and results are available to the observer within sufficient time to direct further interaction with the system as required.

NOTE 2 A scenario evaluation may compare performance of verification and identification systems. Such tests require a careful approach to test design and results reporting to ensure equitable presentation of findings. For example, the order of enrolment, genuine, and impostor trial may change depending on whether systems perform identification or verification.

7.1.1.3 Evaluation environment

The environment in which a scenario evaluation is executed shall be reported inclusive of the following:

— indoor or outdoor;

— if indoor, type of facility;

— if outdoor, degree of exposure to elements.

Environmental conditions relevant to the systems and application under test shall be measured and reported.

EXAMPLE 1 Because temperature and humidity of the test environment are understood to be capable of impacting the performance of certain fingerprint sensors, in a scenario evaluation of fingerprint technology temperature and humidity would be measured and reported.

Measurement of environmental conditions shall be taken at sufficient intervals such that temporal environmental conditions can be characterized.

NOTE Environmental conditions may be introduced or controlled specifically for the purposes of the evaluation, or may be unconstrained.

EXAMPLE 2 Air conditioning may generate background noise sufficient to impact the performance of voice recognition systems.

EXAMPLE 3 Lighting from windows may impact the performance of face recognition systems.

7.1.1.4 Test platform

Systems' processing power and specifications shall be commensurate with the scenario being evaluated.

7.1.2 Test execution

7.1.2.1 Test information and general test instructions

Test information and general test instructions provided to Test Subjects prior to a scenario evaluation shall be reported.

NOTE 1 Test information encompasses the general purpose of the evaluation, characterization of the devices or technologies to be evaluated, and characterization of the target application. General test instructions encompass overall test flow and processes not limited to usage of any specific systems, such as moving from system to system.

NOTE 2 Providing certain types of test information and general test instructions can inform the way that Test Subjects interact with devices. For example, if a Test Subjects knows that certain presentations correspond to impostor trials, he may present a characteristic in a different fashion.

7.1.2.2 Training

The extent and method of training provided to Test Subjects prior to a scenario evaluation shall be reported.

NOTE 1 Training encompasses Test Subject interaction with systems under test, including presentation of characteristics to each device as well as feedback and prompts from each system.

NOTE 2 It may be appropriate to provide no training to Test Subjects if untrained usage is consistent with the target application.

NOTE 3 Training may be provided in the form of written and/or verbal instructions.

NOTE 4 Separate training may be necessary for enrolment and recognition if characteristic presentation or system feedback differ for enrolment and recognition.

Use of supplier-provided scripts, instructions, or other training tools shall be reported.

For comparative scenario tests in which training is provided to Test Subjects, training shall be implemented in a consistent fashion across all systems. Examiners shall ensure that the average duration between Test Subject training and device usage is roughly consistent across all devices. If a Test Subject is trained on several systems directly prior to testing, the initial devices with which he interacts may have an advantage over devices used later in the evaluation.

7.1.2.3 Attended / unattended testing

The presence of an Administrator and/or Operator in the test environment at all times when a Test Subject is engaged in a test activity is recommended.

NOTE The presence of an Administrator and/or Operator provides an opportunity for the test organization to identify and correct cases of incorrect Test Subject interaction with biometric systems.

7.1.2.4 Guidance

Guidance provided to Test Subjects in the course of enrolment and recognition shall be consistent with that of the test's target application.

NOTE 1 The degree of guidance in a scenario test may have a substantial impact on error rates, particularly failure to acquire and failure to enrol, and throughput rates. Increasing the amount of guidance provided during an evaluation will tend to reduce false non-match rate, failure to acquire, and failure to enrol. Providing excessive or insufficient guidance may result in non-representative failure to acquire and failure to enrol.

EXAMPLE 1 A scenario test that evaluates biometric systems in an application wherein no guidance is provided would not incorporate guidance, as performance may not be reflective of that of the target application.

EXAMPLE 2 If consistent with usage in the target application, an Administrator might provide corrective instructions to a Test Subject using a device incorrectly during enrolment but not during recognition.

For scenario tests in which guidance may be provided to Test Subjects, guidance policies shall be documented that address the following:

— point(s) in an enrolment or recognition attempt at which guidance is permitted or required;

— specific guidance an Administrator is to provide to Test Subject;

— aspects of guidance at the discretion of the Administrator, if any.

NOTE 2 A guidance policy might dictate that guidance only be provided for exception cases, such as a Test Subject's incorrect usage of a capture device or in which a system is unable to acquire samples despite proper presentation of a characteristic. In this case, the Administrator would need to observe the presentation and/or the system response to determine whether to provide guidance. Conversely, a guidance policy might dictate that an Administrator provide identical guidance to each Test Subject at the same juncture, regardless of the success of a presentation.

For scenario tests in which guidance may be provided to Test Subjects, guidance shall be implemented in a consistent fashion across all systems.

NOTE 3 Despite best efforts to implement guidance policies consistently across all systems under test, systems may be inadvertently rewarded or penalized depending on the extent of guidance provided to Test Subjects. A system whose differentiator is ease of use or ability to provide automated corrective instructions during usage may not benefit from guidance as strongly as a difficult-to-use system or one that provides no corrective instructions during usage. This may apply to systems of different modalities (e.g. facial recognition and fingerprint) as well as to different systems within a given modality.

Where operator guidance goes beyond a (pre-determined) level consistent with the target application, this shall be recorded and the proportion of such cases reported.

7.1.2.5 Test order and acclimatization

In a multi-system test, the order in which Test Subjects interact with systems shall be arranged to balance acclimatization, habituation and order effects. Each system should be tested a roughly equivalent number of times in the first position, second position, etc., through to the last position; also each system should be preceded by each other system a roughly equivalent number of times. Any observed effects of the ordering on performance shall be reported.

NOTE When testing multiple systems in a scenario evaluation, order is a major consideration. As the degree of habituation to biometric systems can presumably improve within the span of a handful of interactions with biometric systems, the first systems tested within a given modality would likely be at a disadvantage relative to later systems, as Test Subjects learn within the course of a test session the most effective method of interacting with a given modality. Similarly, an individual may "tire" over the course of a test session, particularly if required to sign or recite pass phrases across a number of systems.

Test design shall minimize the presence of temporal conditions of a biometric characteristic known to impact the ability of sensors to process samples.

EXAMPLE A Test Subject might enter a test facility from a cold outdoor environment and immediately begin interacting with fingerprint devices. As many fingerprint devices are less capable of acquiring samples from cold, dry fingerprints than from room-temperature fingerprints, the first fingerprint device(s) that the Test Subject interacts with could be more subject to errors than subsequent devices utilized as fingerprints return to their normal indoor temperature and moisture. Proper test design would ensure that the Test Subject is given enough time to adjust to the test environment; if this is not viable, the test order would be set such that the effects of acclimatization are distributed evenly across devices in question.

7.1.2.6 Test subject identifiers

Use of Test Subject identifiers shall be specified as follows:

— identifiers used to identify Test Subjects;

— method by which identity is claimed in verification systems;

— method by which the subject's true identity is ascertained in identification systems.

7.1.3 Levels of effort and decision policies

7.1.3.1 Enrolment level of effort and decision policies

For each system tested, levels of effort and decision policies for enrolment shall be specified, for example:

— minimum and maximum number of presentations, attempts, and transactions required and permitted to enrol;

— maximum duration permitted and required within each enrolment presentation, attempt or transaction.

NOTE 1 A system may terminate an enrolment attempt or transaction after a fixed duration. This may be due to (1) rejection of an acquired sample due to insufficiently distinctive data or (2) inability to acquire a sample.

NOTE 2 A system may be capable of enrolling a Test Subject after one attempt or may require multiple attempts to enrol.

NOTE 3 The maximum number or duration of presentations, attempts, and transactions during enrolment are referred to as enrolment presentation limits, enrolment attempt limits, and enrolment transaction limits, respectively.

7.1.3.2 Comparison level of effort and decision policies

For each system tested, levels of effort and decision policies for comparison shall be specified, for example:

— minimum and maximum number of presentations, attempts, and transactions permitted or required for comparison;

— minimum duration required and maximum duration permitted for each comparison presentation, attempt or transaction.

NOTE 1 A system may terminate a comparison attempt or transaction after a fixed duration. This may be due to (1) rejection of an acquired sample due to insufficiently distinctive data or (2) inability to acquire a sample or (3) inability to generate a template from the acquired sample.

NOTE 2 A system may be capable of matching a Test Subject after one attempt or may require multiple attempts to match.

NOTE 3 The maximum number or duration of presentations and attempts during comparison are referred to as comparison presentation limits and comparison attempt limits, respectively.

7.1.3.3 Reference adaptation

The Experimenter should address whether systems under test utilize biometric reference adaptation in recognition transactions. If systems utilize biometric reference adaptation, the manner in which reference adaptation was accommodated should be reported. If the proportions of genuine and impostor recognition transactions in which biometric reference adaptation occurred is known, these proportions should be reported.

7.1.3.4 Appropriateness of levels of effort and decision policies

Level of effort and decision policies for enrolment and comparison shall be appropriate to the systems and scenario under test.

NOTE While attempt and transaction limits should be equivalent for all systems under test, systems' acquisition, enrolment, and comparison processes may vary substantially.

EXAMPLE A facial recognition system might declare a failure after a predetermined period of time, whereas a fingerprint system may declare a failure after a certain number of attempts to enrol.

7.1.3.5 Implementation of native and customized levels of effort and decision policies

For each system, implementation of native and customized levels of effort and decision policies shall be specified.

NOTE A system may utilize native, non-adjustable enrolment and recognition functions in which a fixed number of attempts or a fixed amount of time is permitted to enrol or compare. Alternatively, a system may utilize adjustable enrolment and recognition functions in which an Experimenter can modify the number of attempts or amount of time permitted to enrol or compare.

7.1.4 Multiple visits and transactions

Multiple transactions and visits may be used to maximise the amount of data for estimation of performance. When this is done, the repeat transactions should, as far as is possible, be faithful to the scenario under test. This will normally mean that multiple visits are preferred to multiple transactions at the same visit.

NOTE Depending on the Experimenter's level of access to the test population, each Test Subject may be able to execute multiple transactions per visit over the course of multiple visits.

Distribution of time lapsed between enrollment and acquisition of sample features shall be calculated.

7.1.5 Executing genuine and impostor trials

Methods of executing genuine and impostor transactions shall be specified.

Test Subject-discernible test processes and system behaviors should not differ for accepted and rejected attempts and transactions.

NOTE A fundamental issue to be addressed in executing a scenario test is whether results from genuine and impostor comparison attempts are recorded on a transactional basis, as match/no match decisions subsequent to N presentations, attempts, and transactions, or in terms of comparison scores that result from each biometric reference-to-biometric reference comparison. A scenario test protocol may dictate recording of comparison scores subsequent to

genuine and impostor attempts, such that error rates can be determined after the fact through score analysis as opposed to in real time. A protocol may also dictate that a decision be rendered in real-time, which necessitates the use of fixed thresholds on which to make decisions.

7.1.6 Data collection

Methods of data collection shall be specified as follows:

— methods of recording data for each performance element, including those not logged by the system(s);

— processes for auditing and validating performance data collection, including those not logged by the system(s).

The Experimenter shall provide in the Test Report examples of data collection elements such as spreadsheets and logs, whether as screenshots or reproduced forms.

7.2 Test crew

7.2.1 General

A crew shall be recruited for the purpose of enrolment and recognition in test systems.

7.2.2 Habituation

The degree to which the crew is familiarized with each device under test shall be reported.

If Test Subjects' levels of habituation are such that the test population can be clearly categorized according to the level of habituation, error rates should be reported for each such categorization.

NOTE 1 The degree to which a test crew is habituated to device(s) under test can have a substantial impact on error rates and throughput rates. Testing with a crew habituated to devices under test will tend to generate lower false non-match rate, failure to acquire, and failure to enrol than testing with a non-habituated crew.

NOTE 2 The degree of habituation in a crew may range from zero (no experience for any crew members) to total (extensive experience for all crew members). To avoid Test Subjective reporting of the degree of habituation of a crew, quantitative data on habituation, such as historical frequency of use, should be reported.

In evaluations that utilize a habituated test crew, the method by which the crew became habituated to each device under test shall be reported.

EXAMPLE A crew may have become habituated to device(s) under test in the course of employment or through pre-evaluation usage or training in the test environment.

In a multi-device evaluation, the crew's degree of habituation shall be equivalent for each device under test.

NOTE 3 Though habituation is measured in terms of a Test Subject's familiarity with a device, experience with a *type* of device may be sufficient to habituate a Test Subject or crew to similar devices. For example, habituation to a fingerprint device that utilizes a sweeping motion for presentation may be extensible to other devices that utilize a similar presentation method.

NOTE 4 Habituation effects may not impact performance equally across different devices under test. Habituation is less likely to be a factor when evaluating devices whose characteristic presentation process is passive than when evaluating devices whose characteristic presentation process requires careful positioning or incorporates feedback loops. Similarly, habituation effects may not impact performance equally across different modalities.

A scenario evaluation measuring performance in an application whose users are typically habituated should utilize a crew habituated to the device(s) under test. A scenario evaluation measuring performance in an application whose users are typically non-habituated should utilize a crew not habituated to the device(s) under test.

NOTE 5 Recruiting a habituated crew can be difficult. Test Subjects may be trained prior to an evaluation to emulate habituation.

7.2.3 Crew composition

Crew composition shall be reported to include distribution of age and gender.

NOTE 1 Where practical, educational level attained, occupation, and ethnicity should be reported.

NOTE 2 In some circumstances, experimenter may need to permit Test Subjects to opt out of reporting certain elements.

NOTE 3 Crew composition is typically controlled in scenario evaluations through recruitment. This differentiates scenario evaluations from certain technology evaluations in which samples have been previously collected, as well as from certain operational evaluations in which the test population is externally dictated.

NOTE 4 While scenario evaluations can attempt to differentiate performance by age, gender, ethnicity, education level attained, occupation, or other factors of interest, recruitment of a sufficient-sized crew may be difficult or costly.

7.2.4 Test Subject management

Test Subject management processes shall be specified inclusive of the following:

— method of initial Test Subject registration;

— method of ensuring Test Subject uniqueness;

— amount and type of personal data collected;

— use of tokens or badges.

7.3 Performance measurement

7.3.1 General

Experimenters shall determine the types of performance measures to be generated through the scenario test in addition to those listed in clauses 7.3.2 through 7.3.6.

In a scenario test, the following shall be recorded or calculated:

a) for genuine Test Subjects, distribution of time lapsed between enrolment and acquisition of sample features

b) reliability of test results based on number of errors, error rates, test population, and number of transactions executed

c) results as available by demographic group, or associated with different environmental conditions, or for other logical segments of the corpus.

7.3.2 Enrolment

For each biometric system tested, failure enrol rate shall be calculated.

The number of Test Subjects and transactions used to calculate failure to enrol shall be calculated.

For systems in which multiple presentations, attempts, or transactions are permitted or required to enrol, failure to enrol shall be calculated at each effort level from lowest to highest observed.

EXAMPLE For a system in which two to five attempts are permitted to enrol, the percentage of Test Subjects able to enrol in two, three, four, and five attempts would be calculated. If the system also permitted two transactions to enrol, the percentage of Test Subjects able to enrol within one and two transactions would be calculated.

For each biometric system tested, the following should be recorded or calculated:

a) percentage of Test Subjects unable to enrol due to lack of biometric characteristic;

b) average, mean, minimum, maximum, and standard deviation of time to enrol as measured from the point of first presentation of the first characteristic to the sensor to successful enrolment;

c) distribution of enrolment quality scores.

7.3.3 Failure to acquire

Scenario tests shall record the proportion of verification or identification attempts for which the system fails to capture or locate an image or signal of sufficient quality. This is the *failure to acquire* rate (FTA).

The number of presentations used to arrive at this rate, and the point at which a failure to acquire was declared, shall be calculated. The number of Test Subjects and transactions used to arrive at these rates shall be recorded.

NOTE In a scenario test, FTA can be declared by software on a sensor or capture workstation, as well as by a encoding or comparison component. In such tests, FTA is typically attributable to failure to capture or locate an image or signal, but could also be caused by failure to process (e.g. extract features or compare with a reference).

The formula for calculating FTA can be found in ISO/IEC 19795-1.

7.3.4 Verification metrics

For each verification system tested, the following shall be recorded or calculated:

a) False non-match rate and false match rate at the attempt level. Such data may be rendered in the form of a ROC or DET curve. The number of Test Subjects and attempts used to calculate these rates shall be calculated.

b) False accept rate (FAR) and false reject rate (FRR), unless test design is such that false accept rate and false reject rate are identical to false non-match rate and false match rate. Such data may be rendered in the form of a ROC or DET curve. The number of Test Subjects and transactions used to arrive at these rates shall be calculated. For systems in which multiple presentations, attempts, or transactions are permitted or required to match, FRR and FAR shall be calculated at each effort level from lowest to highest observed.

EXAMPLE For a system in which one to three attempts are permitted to match, the percentage of Test Subjects able to match within one, two, and three attempts would be calculated. If this system also permitted two full transactions to match, the percentage of Test Subjects successfully matched within one and two transactions would be calculated.

NOTE 1 For systems that return match / no match decisions as opposed to comparison scores, performance may be calculated at a single operating point on the ROC or DET.

NOTE 2 See Annex C for information on reporting performance results at different levels of effort.

For each biometric system tested, the following should be recorded or calculated:

c) distribution of comparison scores for genuine Test Subject and impostors;

d) average, mean, minimum, maximum, and standard deviation of time to match measured from the point of first presentation of the first characteristic to the sensor to completion of the successful matching transaction.

7.3.5 Identification metrics

For closed-set identification evaluations, cumulative match characteristics shall be calculated.

For open-set identification evaluations, the following shall be recorded or calculated:

a) false match rate and corresponding false non-match rate (preferably over a range of thresholds);

b) false positive identification rates and corresponding false negative identification rates.

The preceding shall be calculated at each effort level for systems in which multiple presentations, attempts, and transactions per permitted to match.

7.3.6 Generalized error rates including failure to enrol and failure to acquire

For each implementation under test, the experimenter shall determine GFAR and GFRR. GFAR and GFRR shall be determined by

— including failed (neither accept nor reject) impostor transactions and impostor transactions for individuals whose enrolment failed in the total number of impostor transactions,

— including failed (neither accept nor reject) genuine transactions and genuine transactions for individuals whose enrolment failed in the total number of genuine transactions, and

— counting failed (neither accept nor reject) genuine transactions and genuine transactions for individuals whose enrolment failed as false rejects.

7.3.7 Interim analyses

Analysis of representative enrolment and matching performance elements shall be conducted at an interim basis prior to conclusion of the testing. Such interim analyses shall be sufficient to validate data collection processes and to ensure that systems are functioning in the manner specific in the test plan. The approach to interim analyses shall be reported. Any anomalous results gathered during these interim analyses that resulted in revision of a test practice or alteration of a system element shall be recorded.

NOTE Interim analysis is necessary in scenario evaluation due to the inability to easily recreate test scenarios linked to reliance on live Test Subjects as opposed to stored data.

7.4 Reporting

7.4.1 General

The results from the evaluation shall be presented in a test report.

All normative elements of test design and performance measurement addressed in Clauses 7.1 through 7.3 shall be documented in the test report. If a requirement in Clause 7.1, 7.2, or 7.3 was out of the scope or not applicable, the report shall state that the requirement was out of scope or not applicable.

EXAMPLE Results cannot be reported at multiple attempt and transaction levels for biometric systems that do not permit multiple attempts or transactions to match. The report would state that the requirement to report performance at multiple effort levels was not applicable.

If a requirement was not addressed due to information being unavailable, the report shall state that the applicable data is unknown. The report shall explain why the data is unknown.

EXAMPLE An organization may not be permitted to record demographic information during testing for privacy reasons. The test report would state that demographic information was not collected for privacy reasons.

7.4.2 System information

7.4.2.1 General

The Experimenter shall collect information regarding the system(s) tested sufficient to execute testing and report test results.

NOTE 1 As opposed to technology evaluation, in which a common hardware platform is used to evaluate multiple components scenario evaluation may entail multiple systems being tested on different platforms, ranging from standalone devices to multi-processor workstations.

NOTE 2 Scenario evaluation may incorporate commercial off-the-shelf (COTS) systems, customized systems, or a mixture of both. There are benefits to enforcing COTS-only or customized-only system requirements. By testing COTS systems, a test organization has an increased certainty that device performance is reflective of a sensor/algorithm combination as available on the market. By allowing for customization, a test organization has an increased certainty that a sensor/algorithm combination can be modified to meet the requirements of a given test scenario. Customization of a biometric system could entail modifying enrolment thresholds to accommodate specific test populations. Customization of systems for an evaluation is not generally considered an ideal process, as the results may be more reflective of a supplier's ability to customize a system to address a specific scenario than of the biometric system's core abilities to enrol and verify users. However it should be noted that there may be substantial interest in how well a supplier can customize a system to meet a certain test protocol.

7.4.2.2 Specifications

For each system tested, the following shall be reported:

a) For acquisition devices: manufacturer, model, version, and firmware as applicable. If the acquisition device's core acquisition components are integrated within a third-party device, such as in the case of a fingerprint sensor incorporated into a peripheral, then manufacturer, model, version, and firmware of the core acquisition components shall be reported.

b) For comparison algorithms: provider, version, revision.

c) If the scenario test incorporates application software, such as a demonstration application or logical access interface: provider, title, version, and build of the application.

d) For systems tested on or through personal computers (PCs), personal data assistants (PDAs), or other computing devices: platform, operating system, processing power, memory, manufacturer, and model of computing device.

7.4.2.3 Architecture

For each system tested, the following shall be reported:

a) biometric data acquisition, processing, and storage architecture;

b) data flow between system components.

7.4.2.4 Outputs

For each system tested, the following shall be reported:

a) available system outputs, such as comparison scores, accept/reject decisions, candidate lists, enrolment quality scores, and sample quality scores;

b) range of values system is capable of outputting for each system output;

c) supplier-provided thresholds and descriptions of values or parameters;

EXAMPLE A system may be capable of providing comparison scores that range from 0-100, with 0 indicating weakest match, 100 indicating strongest match, and 75 indicating a minimum 1:1 match threshold.

d) method(s) through which outputs are provided by the system.

EXAMPLE Comparison scores may be logged by an application or be visually indicated through a graphical user interface.

7.4.3 System acquisition and implementation

For each system tested, the following shall be reported:

a) method of biometric and platform system acquisition;

b) level of supplier involvement in system implementation.

7.4.4 Physical layout of test environment

The physical layout of the test environment shall be reported, including but not limited to the following:

a) dimensional area dedicated to scenario test execution;

b) presence of natural and artificial lighting;

c) positioning of biometric acquisition devices;

d) relative location of each system in the test environment, rendered through a system schematic;

e) photographic images of the test environment sufficient to clearly indicate the relative positioning of devices and Test Subjects during testing.

7.4.5 Report structure

The following sections shall be incorporated in the test report:

— Executive Summary;

— Scenario Description;

— Specific Test Processes;

— Data Collection;

— Data Analysis;

— Record Keeping;

— Performance Results;

— Full Test Plan.

8 Other issues applicable to technology and scenario evaluations

8.1 Parties to a test

An evaluation shall be conducted by a tester. The biometric system under test shall be provided by one or more supplier(s). If the tester and supplier are the same entity, or are affiliated or are otherwise not independent, then this shall be documented in the test report.

Supplier involvement in technology and scenario evaluations is restricted to the supply, installation, and configuration of software and/or hardware. The testing organization executes enrolment and comparison tests without supplier input.

NOTE If it is imperative that an evaluation is construed to be a supplier's own best effort with no possibility of error by the testing organization, an alternative type of test in which tester and supplier roles differ from those enumerated in 8.1 can be conducted. Known as a supplier self-test, this type of test allows the supplier to provide, configure, and operate their own system on tester supplied materials. The tester is absolved of blame if the results are claimed to be deficient. Such an evaluation should use a client-server paradigm. Such tests are problematic in terms of expense, gaming and sample privacy.

8.2 Fairness

A competitive test shall not be designed to favour particular suppliers.

NOTE 1 This clause would not normatively apply to an institution executing technology evaluations for internal research and development purposes.

NOTE 2 After an evaluation is announced, prospective suppliers typically "fish" for information on many issues (sample formats, properties, qualities, the interface, administration procedures, etc.) and the answers to these questions should be made public. A web-based frequently asked question (FAQ) may be an appropriate vehicle. The identity of the questioner shall be suppressed and this practice itself will be noted in the preamble to the FAQ and the test announcement.

NOTE 3 In technology tests, the tester will typically release to all suppliers representative sample data in the format to be used in the test.

Experimenters shall document any involvement on the part of the test organization in the configuration, modification, refinement, or adaptation of the implementation under test.

Experimenters shall document intellectual or physical input on the part of the test organization that materially affects any outcome of the evaluation.

In case multiple components or systems are tested, examiners should report, whether computing systems were tested on equivalent hardware and operating systems or whether images of the operating system were re-installed prior to each test segment in a system by system testing manner.

8.3 Basis for inclusion of test systems

The Experimenter shall report address the basis by which algorithms and systems are included in technology and scenario evaluations. Inclusion of algorithms and systems in the evaluation might be on the basis of

a) an open invitation to participate;

b) selection by a test organization, in which case the selection criteria shall be reported;

c) a contract with a supplier or a 3rd party to test a particular system.

Technology and scenario evaluations can incorporate a single biometric system or multiple biometric components or systems. Technology and scenario evaluations can also incorporate like combinations of multiple biometric components or systems. Testing multiple systems provides the advantage of potentially establishing a range of performance against which different systems can be evaluated. Anomalous

performance can be difficult to gauge from a single-system test. The number of systems tested may be constrained by budgetary constraints, availability of suitable technologies, or time required to acquire samples or process data.

8.4 Use of Frequently Asked Questions

In a competitive technology or scenario evaluation a frequently-asked-questions document may be maintained as a mechanism of communication between a test organization and the suppliers. The author of each question should be suppressed.

8.5 Legal issues

Legal issues in technology and scenario test design, execution, and reporting may need to be addressed. It may be necessary to enact a non-disclosure agreement between the suppliers and the test organization. Certain jurisdictions may require that a Data Privacy Agreement be established between the Test Subject and the test organization.

8.6 Release of test source code

Depending on the type and purpose of the test, it may be appropriate to release test source code to suppliers.

8.7 Supplier comment on test report

Depending on the type and purpose of the test, it may be appropriate to allow suppliers to comment on a pre-release version of the report as prescribed by the testing organization.

Annex A
(informative)

Phases and activities for primary technology test types

A.1 Simple verification test

A simple verification test is the most basic assessment of fundamental biometric power of an algorithm on a database. It can be used repeatedly for component development and also for comparative system evaluation. Can also be used to assess “difficulty” of a data set.

A simple verification test generates false reject rate, false non-match rate, false accept rate, and false match rate.

EXAMPLE May be representative of single-enrolee systems such as a PDA.

Phase	#	Activity
Data Extraction	1	Construct two partitions 1. E, the first sample of each person representing an enrolment sample. 2. 2. U, the second sample of each person in E, representing a user sample.
Execution	2	Make biometric references 1. Run biometric reference generator on all samples from E. 2. Time each operation and store result. 3. Record proportion of samples that were declared unenrolable and compute failure to enrol. 4. Store the (non-failure to enrol) biometric references.
	3	Extract sample features 1. Shuffle the elements of U. Retain the permutation (so as to link matches back to those in E). 2. Run feature extractor on all raw samples from U. 3. Time each operation and store result. 4. Record proportion of samples declared un-usable and compute failure to acquire. 5. Store the (non-failure to acquire) sample features.
	4	Make transactions lists 1. Form the list of biometric references and sample features, A, of N matching pairs from E and U. 2. Form the list of biometric references and sample features, B, of N(N-1) non-matching pairs from E and U. 3. Concatenate A and B and shuffle (randomly permute) the result, C. Retain match and non-match statuses.
	5	Perform full cross-comparisons 1. Run the verifier on each pair from C. 2. Time each operation, record separately for match and non-match pairs. 3. Append each comparison score into separate lists of match and non-match scores.
Reporting	6	DET curve computation 1. Form the set of unique comparison scores, S. 2. For each value, s, from S: a. Compute proportion of genuine comparison scores lower than s, i.e. false non-match rate(s). b. Compute false reject rate(s) from false non-match rate(s) using formula of 5.1.8.4. c. Compute proportion of impostor comparison scores higher than s, i.e. false match rate(s). d. Compute false accept rate(s) from false match rate(s) using formula of 5.1.8.4. 3. Plot DET as (false accept rate(s), false reject rate(s)) for all s.

Phase	#	Activity
	7	Compute throughput statistics 1. Successful and (separately) failed biometric reference generations. 2. Successful and failed genuine and impostor sample feature extractions. 3. Comparisons, separately for matches and non-matches.
	8	Summarize results, and report according to reporting policy

NOTE 1 If the biometric references are asymmetric (i.e. f(enrol,user) ≠ f(user,enrol)) the raw sample sets E and U can be swapped and the test rerun. This may not be representative of the target application, however.

NOTE 2 If there is evidence that the result of an impostor attempt depends on whether another sample from the impostor has previously been used (as an enrolee in E, or user in U) then a true impostor test may be needed: this involves using a third sample partition, I, of true impostors, to be paired with E, such that a sample from I is never used in the enrolee role.

A.2 Verification test with multiple enrolees

A verification test with multiple enrolees is similar to the simple verification test, though modified for evaluation of a multi-user biometric device. This type of test can account for improved verification possible with exploit of other enrolled biometric references such as through cohort normalization. A verification test with multiple enrolees entails exhaustive claims of identity into a linear enrolled population. The algorithm under test may exploit other enrolled data to make dependent biometric references or to do normalization.

A verification test with multiple enrolees generates false reject rate, false non-match rate, false accept rate, and false match rate.

EXAMPLE A verification Physical access control to a building

Phase	#	Activity
Data Extraction	1	Construct two partitions: 1. E, the first sample of each person representing an enrolment sample. 2. U, the second sample of each person in E, representing a user sample.
Execution	2	Enrolment 1. Initialize supplier's enrolee data structure (EDS). 2. For each of N samples from E run biometric reference generator: a. Time the operation, store result. b. if not failure to enrol append biometric reference to EDS. c. Record proportion of samples that were declared un-enrolable and compute failure to enrol. d. Finalize EDS. e. Time this operation and store result.
	3	Sample feature extraction 1. Shuffle the N elements of U. Retain the permutation (so as to link matches back to those in E). 2. Run feature extractor on all raw samples from U. 3. Time each operation, store result. 4. Record proportion of samples declared un-usable and compute failure to acquire. 5. Store the (non-failure to acquire) sample features.