INTERNATIONAL STANDARD

ISO/IEC 19795-4

First edition
2008-06-01

Information technology — Biometric performance testing and reporting —

Part 4: Interoperability performance testing

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

Partie 4: Essais de performances d'interopérabilité

Reference number
ISO/IEC 19795-4:2008(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 vi

Introduction vii

1 Scope 1

2 Conformance 1

3 Normative references 2

4 Terms and definitions 2

5 Abbreviated terms 4

6 Goals 5
6.1 Coverage 5
6.2 Target application 8
6.2.1 Biometric application 8
6.2.2 Interoperable application 9
6.3 Purpose 10
6.3.1 Interoperability testing 10
6.3.2 Sufficiency testing 11

7 Metrics 12
7.1 General 12
7.2 Figures of merit 12
7.2.1 Recognition performance figure of merit 12
7.2.2 Measuring component failure 13
7.3 Interoperability matrices 14
7.3.1 General 14
7.3.2 Interoperability with sBDB generators 14
7.3.3 Interoperability with sBDB generators 15
7.3.4 Fixed operating point interoperability 16
7.3.5 Reporting failure of sBDB generators 16
7.4 Proprietary performance 16

8 Conducting a test 17
8.1 Structure of test 17
8.2 Sample data 17
8.2.1 Acquisition 17
8.2.2 Representative data 18
8.2.3 Collection of ancillary data 18
8.2.4 Corpus size 18
8.2.5 Removal of subject-specific metadata 18
8.2.6 Removal of unrepresentative metadata 18
8.2.7 Origin of samples 19
8.2.8 Untainted samples 19
8.2.9 Sequestered data 19
8.3 Conformance testing 19
8.3.1 Conformance 19
8.3.2 Executing conformance tests 19
8.3.3 Reporting 20
8.4 Constraints on the sBDBs 20
8.4.1 Optional encodings 20
8.4.2 Optional encodings from profile standards 20
8.4.3 Deviation from the base standard 20
8.4.4 Data encapsulation 20

8.5 Components 21
8.5.1 Components for sufficiency testing 21
8.5.2 Establishing modularity requirements 21
8.5.3 Components for interoperability testing 21
8.5.4 Underlying algorithms 21
8.5.5 Capture device user interfaces 21
8.5.6 Multimodal components 22
8.5.7 Component variability 22
8.5.8 Component reporting requirements 22
8.6 Planning decisions 22
8.6.1 Computational intensity 22
8.6.2 Supplier recruitment 23
8.6.3 Provision of samples to suppliers 23
8.6.4 Equivalency of generator resources 23
8.6.5 Handling violations of test requirements 24
8.6.6 Comparison subsystem output data encapsulation 24
8.6.7 Fundamental generator requirement 24
8.6.8 Fundamental comparison subsystem requirement 25
8.6.9 General requirements on software implementations 25
8.7 Prevention and detection of gaming 26
8.7.1 General aspects 26
8.7.2 Modes of gaming 26
8.7.3 Prevention and detection of gaming 28
8.8 Test procedure 29
8.8.1 Primary test 29
8.8.2 Uncertainty measurement 30
8.8.3 Variance estimation 30
8.8.4 Remedial testing 30
8.8.5 Survey of configurable parameters 30

9 Interpretation of the interoperability matrix 30
9.1 Determination of interoperable subsystems 30
9.1.1 General 30
9.1.2 Identifying interoperable combinations of subsystems 31
9.1.3 Acceptable numbers of interoperable subsystems 33
9.1.4 Combinatorial search for maximum interoperability-classes 33
9.1.5 Multiple interoperable subgroups 34
9.1.6 Statistical stability of the test result 34
9.2 Interoperability with previously certified products 35
9.2.1 Decertification considerations 35
9.2.2 Continuity of testing 35
9.2.3 Interoperability with previously certified generators 35
9.2.4 Interoperability with previously certified comparison subsystems 36
9.2.5 Treatment of systematic effects 36
9.2.6 Retroactive exclusion from analysis 37
9.3 Overall sufficiency 37

Annex A (informative) Procedures for conducting a test of sufficiency and/or interoperability 38

Annex B (informative) Example Interoperability Test 42

Bibliography 45

Figure 1 — General biometric interoperability 6

Figure 2 — Specific interoperability: enrolment BDB is standardized 6

Figure 3 — Specific interoperability: enrolment BDB is proprietary 7

Figure 4 — Offline interoperability testing 7

Figure 5 — Biometric capture device interoperability 8

Figure 6 — Cells of an example interoperability space 10
Figure 7 — Sufficiency testing: proprietary vs. standard interchange formats 12
Figure 8 — Cross-generator performance matrix 15
Figure 9 — Example performance matrix 15
Figure 10 — Proprietary performance matrix 16

Table 1 – Conformity with ISO/IEC 19795-2 1
Table 2 – Sample size adjustment of error rate requirement 31
Table 3 – Confidence levels of the standard Normal distribution 32
Table A.1 – Interoperability test procedure, phase 1: planning 38
Table A.2 – Interoperability test procedure, phase 2: setup 39
Table A.3 – Interoperability test procedure, phase 3: sBDB and pBDB generation 39
Table A.4 – Interoperability test procedure, phase 4: verification 40
Table A.5 – Interoperability test procedure, phase 5: identification 40
Table A.6 – Interoperability test procedure, phase 6: reporting 41
Table A.7 – Interoperability test procedure, phase 7: variance estimation 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 19795-4 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*

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

— *Part 4: Interoperability performance testing*

Part 6: Testing methodologies for operational evaluation is under preparation.

Introduction

The multi-part biometric data interchange format standard, ISO/IEC 19794, has been developed to foster interoperable exchange of biometric data. By defining open containers for image, signal and feature data, and constraining some of the properties of the samples, the standards enhance interoperability by requiring implementers to be able to handle a restricted set of all possible biometric samples. Examples of this are the template standards of ISO/IEC 19794-2 and ISO/IEC 19794-8 which embed compact processed data from fingerprint images. Only samples of the same format type (several of which can be defined in the same part of ISO/IEC 19794) are intended to be interchangeable.

One common assertion prior to SC 37's formulation of data interchange standards was that proprietary templates offer greater recognition performance than any likely standard on the grounds that the proprietary instances are the product of processes that embed considerable, private, intellectual property. The question of whether the emerging standards are sufficient then arises: that is, do they code data (feature, image, etc.) representations that allow matching with accuracy comparable to that available from the proprietary solutions?

A second issue, interoperability, arises in those applications where standardized data are generated and matched by different institutions and systems. If a company's feature extraction subsystem processes acquired samples to produce ISO/IEC 19794-x compliant instances, then can other companies' comparison subsystems attain performance comparable with that obtained from the originator's own comparison subsystem? A further question is then whether a third company can successfully recognize enrolment and user samples from two different sources.

This part of ISO/IEC 19795 defines tests to specifically address absolute performance, sufficiency, and interoperability available from biometric data formatted to comply with established standards, particularly those developed in the various parts of ISO/IEC 19794. However, because this part of ISO/IEC 19795 references interchange formats generically, by referencing only their black box generation and use, it also applies to other open standards. One consequence of this approach is that the success of a test is predicated on the correctness and appropriateness of lower-level data elements and values, i.e. conformance to the respective standards. Therefore, the approach here is to require conformance testing as an integral part of the test. This is achieved by referencing formal published conformance tests or profiles of standards. For instance, an interoperability test of the ISO/IEC 19794-5 face format might reference an application profile of its Token image, which in turn might rely on ISO/IEC 15444-1 (JPEG 2000 core coding system).

This part of ISO/IEC 19795 conceives of the following three kinds of tests:

- **online:** a scenario test in which a volunteer population enrols on suppliers' products and subsequently uses suppliers' verification or identification implementations to make genuine and impostor attempts;
- **offline:** a technology test in which an archived corpus of captured samples, not necessarily collected with any intent to simulate the operational conditions of a particular application, is used as input to suppliers' enrolment, verification or identification products to make genuine and impostor attempts;
- **hybrid:** a test in which the sample corpus is collected online under conditions which attempt to simulate the operational conditions of a particular application, and is then processed offline.

In each case, an interoperability test needs to embed multi-supplier generation, exchange, and comparison of samples of the standard interchange format. Online collection from a live population is appropriate when the biometric capture device, and/or the subject interaction with the biometric capture device, is considered to have a material effect on the interoperable performance of the intended application. An offline test is appropriate when a representative corpus of samples is already available (for example passport photographs to be converted into Token instances of ISO/IEC 19794-5). An offline test may be appropriate when the collection of representative data is neither practical nor necessary to determine the interoperable performance of specific subsystems, such as feature extraction and/or comparison.

In all cases, an interoperability test must enrol subjects on one or more products and verify or identify on one or more others. This should involve subjects making transactions as themselves (genuine trials) and as one or more other people (impostor trials). If a large enough population is available, a disjoint impostor population can be used. Since online tests can become onerous on the test population when many products and impostor attempts are needed, hybrid and offline testing allow execution of many zero-effort impostor attempts.

In an interoperability performance test, J generators of standardized biometric data blocks (BDBs) are applied to the samples assembled as part of a hybrid or offline test. By applying K comparison subsystems to the standard BDBs, up to KJ^2 verification or identification trials are conducted, each following ISO/IEC 19795-2. The BDB may be an image or signal, or a standardized template. Optional encodings allowed by the standard interchange format should be fully specified. This might be achieved by normatively referencing one of the ISO/IEC 24713-x profiles. If the format in question is an image, a subsequent internal (usually proprietary) template would be used, but its existence here is subsumed by the notion of a black-box comparison of two instances of the given format.

The test advanced by this part of ISO/IEC 19795 demarcates the generic aspects of interoperability from the meaning associated with each particular biometric format of ISO/IEC 19794-x.

Information technology — Biometric performance testing and reporting —

Part 4: Interoperability performance testing

1 Scope

This part of ISO/IEC 19795 prescribes methods for technology and scenario evaluations of multi-supplier biometric systems that use biometric data conforming to biometric data interchange format standards.

It specifies requirements needed to assess

— performance available from samples formatted according to a standard interchange format (SIF),

— performance available when samples formatted according to a SIF are exchanged,

— performance available from samples formatted according to a SIF, relative to proprietary data formats,

— SIF interoperability, by quantifying cross-product performance relative to single-product performance,

— performance available from multi-sample and multimodal data formatted according to one or more SIFs, and

— performance interoperability of biometric capture devices.

In addition, this part of ISO/IEC 19795

— includes procedures for establishing an interoperable set of implementations,

— defines procedures for testing interoperability with previously established sets of implementations, and

— gives testing procedures for the measurement of interoperable performance.

It does not

— establish a conformance test for biometric data interchange formats, or

— provide test procedures for online data collection.

2 Conformance

An interoperability performance test conforms to this part of ISO/IEC 19795 if it satisfies the requirements specified in Clauses 6, 7, 8 and 9 of this part of ISO/IEC 19795 and the requirements specified in the clauses of ISO/IEC 19795-2 referenced in Table 1.

Table 1 — Conformity with ISO/IEC 19795-2

Structure of ISO/IEC 19795-4 test	ISO/IEC 19795-2 conformance
Online (8.2.1.3)	Clause 7 (Scenario evaluation)
Hybrid (8.2.1.4)	Clause 6 and Clause 7
Offline (8.2.1.2)	Clause 6 (Technology evaluation)

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*

ISO/IEC 19795-2, *Information technology — Biometric performance testing and reporting — Part 2: Testing methodologies for technology and scenario evaluation*

4 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 19795-1, ISO/IEC 19795-2 and the following apply.

4.1
basic interoperability
ability of a supplier's generator to create BDBs that can be processed by other suppliers' comparison subsystems, and the ability of a supplier's comparison subsystem to process BDBs from other suppliers' generators

4.2
biometric capture device
BCD
device that collects a signal from a biometric characteristic and converts it to a biometric sample

NOTE A device can be any piece of hardware, and supporting software and firmware.

4.3
biometric data block
BDB
block of data with a defined format that contains one or more biometric samples or biometric templates

4.4
captured biometric data block
cBDB
block of universally understood, possibly standardized, image or signal data produced by a biometric capture device

NOTE A cBDB is, by definition, an sBDB. It is used in Figures 1 to 5 to indicate the minimal unprocessed output of a biometric capture device.

EXAMPLE Greyscale raster image from a fingerprint scanner.

4.5
comparison subsystem
subsystem capable of comparing standardized or proprietary biometric data blocks

NOTE 1 When a test of an image-level SIF is conducted, a verification comparison subsystem will compare two images (usually by internally producing, then comparing, two proprietary and non-interoperable templates). Because each of the input samples will be used again, it will be more computationally efficient for the proprietary templates to persist within the comparison function. This part of ISO/IEC 19795 ignores the internal operation of each supplier's test software, but a throughput computation may need to break out rates for "first" comparisons and "second" (i.e. already stored template) comparisons.

NOTE 2 The definition should not be construed to exclude systems that legitimately perform more than a single one-to-one comparison in order to verify. Certain cohort normalization techniques, for example, perform additional internal

comparisons intended to improve performance. But such a comparison subsystem remains a black box that accepts two instances to produce a comparison score.

4.6
failure to acquire rate
FTA
proportion of recognition samples for which a generator fails to produce an instance suitable for comparison

NOTE In cases where a single sample is used for each subject, the sample-failure rate is the same as the attempt-failure rate and this definition agrees with, but is a special case of, the definitions given in ISO/IEC 19795-1 and ISO/IEC 19795-2.

4.7
failure to enrol rate
FTE
proportion of enrolment samples for which a generator fails to produce an instance suitable for comparison

NOTE In cases where a single sample is used for each subject, the sample-failure rate is the same as the person-failure rate and this definition agrees with, but is a special case of, the definitions given in ISO/IEC 19795-1 and ISO/IEC 19795-2.

4.8
generator
subsystem capable of producing a standardized or proprietary biometric data block

NOTE 1 Under this definition, a biometric capture device might constitute a generator.

NOTE 2 The subsystem may be implemented in software and/or hardware.

NOTE 3 Referring to ISO/IEC 19785-1 (CBEFF data element specification), a generator would transform a source BDB to a target BDB.

4.9
interoperable performance
performance associated with the use of generator and comparison subsystems from different suppliers

4.10
native performance
performance associated with the use of generator and comparison subsystems from a single supplier

4.11
performance interoperability
measure of the adequacy of interoperable performance

NOTE Performance interoperability expresses the ability of biometric subsystems from different suppliers to generate and compare samples, and to either meet an absolute level of performance or constrain error rates within some relative (i.e. non-absolute) bound.

4.12
proprietary format
PF
format defined in a privately controlled biometric data format specification

4.13
proprietary biometric data block
pBDB
biometric data block conforming to a proprietary format

4.14
proprietary performance
performance associated with the use of proprietary generator/comparison subsystems

4.15
standardized biometric data block
sBDB
block of data with a standard interchange format that contains one or more biometric samples or biometric templates

NOTE This part of ISO/IEC 19795 conceives of a biometric sample as a set of one or more instances of acquired biometric data. This definition therefore includes multi-sample and multimodal data. While none of the parts of ISO/IEC 19794 defines multimodal containers, many of them allow multiple instances. The inclusion of multi-sample and multimodal data is supported by the view of generators and comparison subsystems as black boxes in this part of ISO/IEC 19795.

EXAMPLE 1 An sBDB could be a fingerprint minutiae template conforming to ISO/IEC 19794-2.

EXAMPLE 2 Three ISO/IEC 19794-5 Token face images produced from a person on three separate occasions.

EXAMPLE 3 An ISO/IEC 19794-6 iris image and an ISO/IEC 19794-11 hand geometry image wrapped together in a complex ISO/IEC 19785-1 CBEFF structure.

4.16
standard interchange format
SIF
format defined in a part of ISO/IEC 19794 or in any other publicly available biometric data format specification

4.17
sufficiency
measure of the adequacy of native performance using a standard interchange format

NOTE 1 Sufficiency may be assessed relative to proprietary performance, or against a specified performance level, e.g. "the standard interchange format is sufficient to achieve an EER below 2%" or "the standard interchange format is sufficient to achieve an EER at most 1,5 times that of proprietary performance".

NOTE 2 Sufficiency aims to quantify whether the interchange standard unambiguously embeds sufficient information to attain performance comparable with that available from existing proprietary formats.

NOTE 3 Sufficiency of a standard interchange format is dependent on the intended application. A data interchange format that is sufficient for high quality images, or for a 1% equal error rate, may be insufficient for low quality images, or for a more stringent accuracy requirement. Nevertheless, any finding of a lack of sufficiency does however indicate the SIF was either incapable of marking up the same data as the proprietary instance or, at least, was not exploited to maximum effect.

4.18
supplier
researcher, commercial entity, organization or institution providing a biometric capture device, generator or comparison subsystem

5 Abbreviated terms

For the purposes of this document, the following abbreviations apply.

— API application programming interface
— BCD biometric capture device
— BDB biometric data block
— CBEFF Common Biometric Exchange Formats Framework (i.e. ISO/IEC 19785)
— SIF standardized interchange format
— cBDB captured biometric data block
— sBDB standardized biometric data block

— PF proprietary format
— pBDB proprietary biometric data block
— FAR false accept rate
— FRR false reject rate
— FMR false match rate
— FNMR false non-match rate
— FTA failure to acquire rate
— FTE failure to enrol rate
— FNIR false negative identification rate
— FPIR false positive identification rate
— GFAR generalized false accept rate
— GFRR generalized false reject rate

NOTE 1 In a fingerprint template interoperability test, the reader may find benefit in mentally replacing the sBDB acronym with the term "standard template instance". The term is used here to allow this part of ISO/IEC 19795 to refer generically to standardized signals, images and templates.

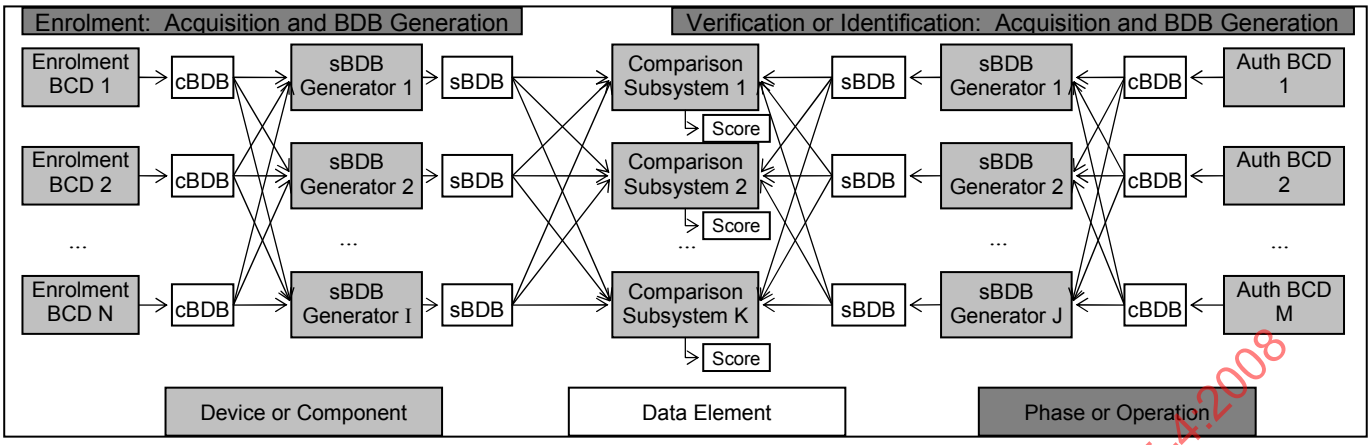
NOTE 2 The quantities FAR, FRR, FMR, FNMR, FTA, FTE, FNIR and FPIR are defined in Clause 4 of ISO/IEC 19795-1:2006. The quantities GFAR and GFRR are also addressed there, in 8.3.4.

6 Goals

6.1 Coverage

The test plan and test report shall document the specific aspects of interoperability that are being investigated. The test report shall include the numbers of suppliers who provided the various components essential to the target interoperability application. A test shall assess sufficiency or interoperability or both. The test plan and test report shall relate its goals to the following overview.

EXAMPLE Six suppliers provided ISO/IEC 19794-5 Token image generators. Each supplier teamed with a supplier of the ISO/IEC 15444 (JPEG 2000) compression format. Four generators employed supplier A's compressor while the other two used supplier B's. In all cases, captured face images were acquired using the biometric capture device from supplier X. These were stored without any compression. Products from the six suppliers were used generate Token instances represented enrolment samples. Comparison subsystems from the same six suppliers were used to compare Token images from each generator with captured images representing authentication samples.



NOTE 1 The sBDB references in this Figure may be replaced by pBDB with the exception that a data format interoperability test will not involve pBDBs in both the enrolment and verification/identification phases. The crossed arrows, which depict interchange, would not be appropriate when pBDBs are generated.

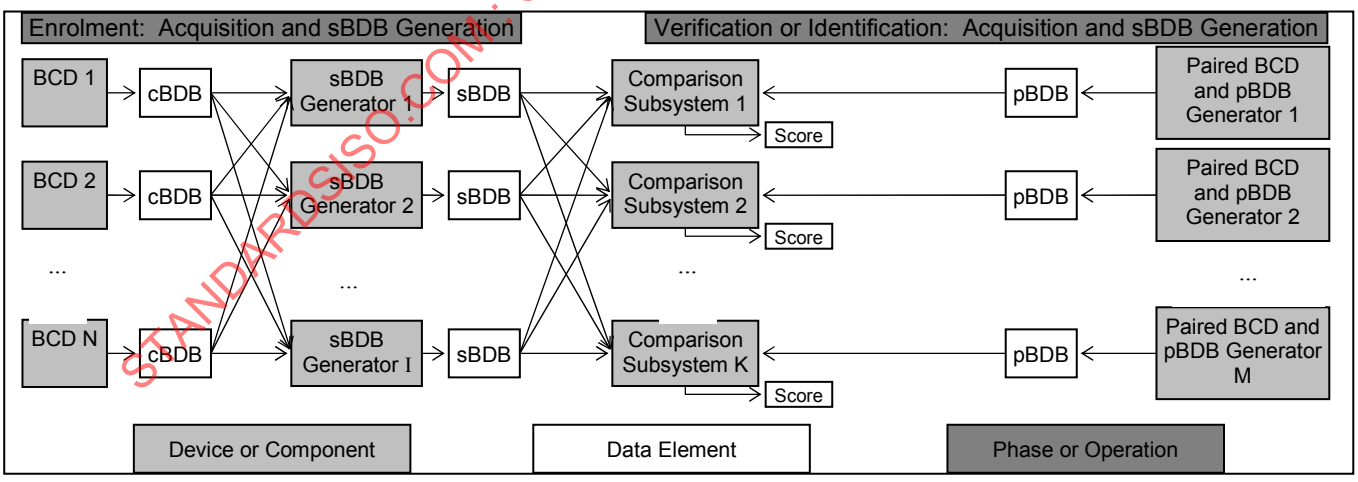
NOTE 2 For identification systems the term "Score" here would more appropriately be replaced by "Candidate List".

NOTE 3 As discussed in clause 7.2.2, each device or component (shown in the mid-grey boxes above, and in the remaining figures) will, in general, have an associated failure-to-process rate.

NOTE 4 This and subsequent figures depict biometric capture devices as generating captured sBDBs (typically unprocessed images) which can be interoperably accepted by all generators. Figure 2 and it's note depict variations on this configuration in which the BCD and the generator are combined with only internal non-standardized data flow.

Figure 1 — General biometric interoperability

Figure 1 depicts the general biometric interoperability problem: different biometric capture devices are used to acquire data that is enrolled in sBDB format by each of I generators for later use in K comparison subsystems. This data is compared with verification or identification data gathered on M biometric capture devices and converted to sBDB form by J generators.



NOTE In some applications, biometric capture devices and generators will be paired. This may arise because there is no need to retain captured samples. A biometric capture device supplier might team with more than one generator supplier, or vice versa. There may be a performance benefit inherent in the BDB generator being tailored to the biometric capture device (rather than having to deal with all possible biometric capture devices).

Figure 2 — Specific interoperability: enrolment BDB is standardized

Some special cases of Figure 1 are described in the following list.

— A common commercial case is depicted in Figure 2: The verification or identification product produces a pBDB which is compared to an enrolled sBDB. Such is the case with an identity credential storing sBDBs for off-card verification (see [1] as an example of such a test).

— The reverse of this situation (a pBDB is enrolled and later compared with a sBDB) is also possible in, for example, a match-on-card application. This is depicted in Figure 3.

— When an offline test is conducted (see, for example, [2]), or when data collection has been done separately, Figure 4 may be appropriate. Note that one but not both of the enrolment and verification BDBs may be pBDBs.

— When the effect of the biometric capture device on performance is of interest (see, for example, [3]) a single BDB generator and comparison subsystem may be appropriate, as shown in Figure 5. Although a biometric capture device interoperability evaluation of this kind does not necessarily involve exchange of sBDBs it is consistent with the definition of performance interoperability in clause 4.12, and is notable because it quantifies biometric capture device performance in terms of recognition error rates rather than its imaging properties.

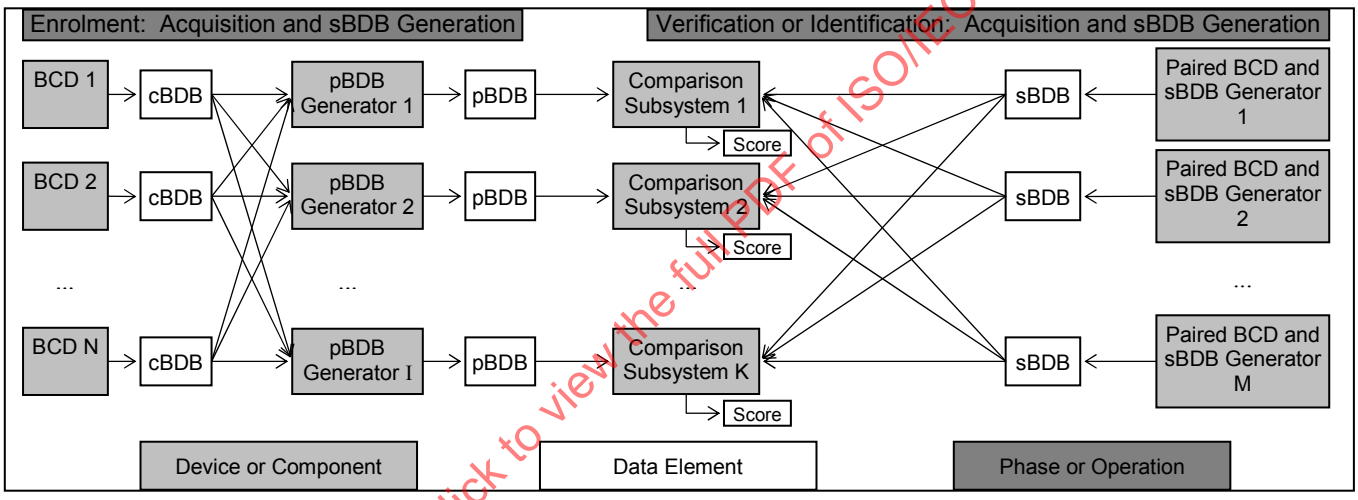
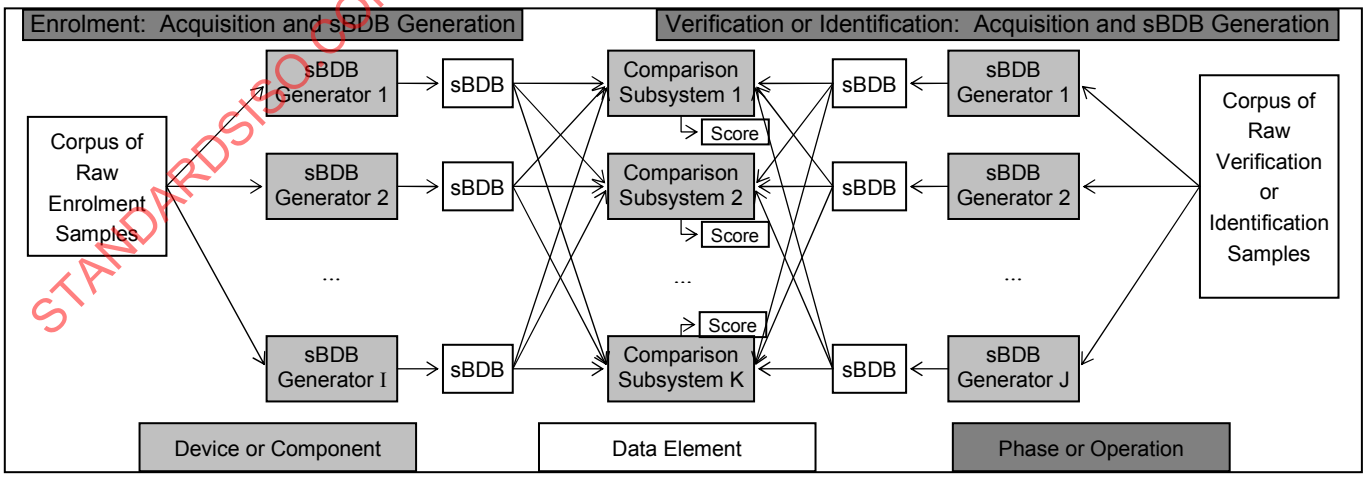


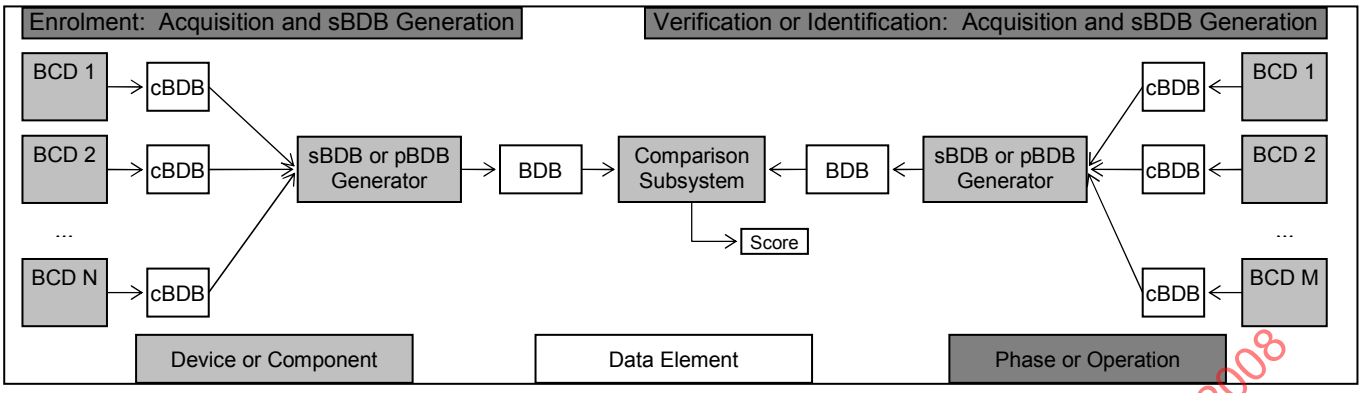
Figure 3 — Specific interoperability: enrolment BDB is proprietary



NOTE 1 The interoperability of capture devices can be tested if the samples from the corpus of captured verification or identification samples stem from a different capture device than the samples in the corpus of captured enrolment samples.

NOTE 2 Either, but not both, of the phases could be mediated by pBDBs here, instead of sBDBs (as in Figure 3). If both phases used pBDBs then this would depict a traditional technology test of the kind standardized in ISO/IEC 19795-2, Clause 6.

Figure 4 — Offline interoperability testing



NOTE This figure is idealized in that it depicts a single comparison subsystem. This is the minimum requirement for comparing biometric capture devices. Practically a test might embed more comparison subsystems and the figure would resemble Figure 1.

Figure 5 — Biometric capture device interoperability

6.2 Target application

6.2.1 Biometric application

6.2.1.1 Defining a transaction

The test plan and test report shall describe the verification or identification trials that the evaluation uses to represent one or more target applications.

The test plan and test report shall define what constitutes a transaction. For an online transaction, this documentation will include how users interact with the biometric capture device, how many attempts that may make, what feedback, if any, is provided to the user, and whether the user is provided with decision from any attached comparison subsystem.

For an offline transaction, this documentation shall specify the numbers of samples that are input to the components under test. It should also specify the order they are provided, and any contingencies associated with their provision.

NOTE 1 ISO/IEC 19795-1 formally defines the term transaction in Clause 4.

NOTE 2 An offline test might "replay" the sequence of events that constituted an online transaction in a prior collection phase. This would support offline estimation of fielded performance.

NOTE 3 As with most biometric performance tests, interoperability tests that allow multiple attempts are likely to report fewer false rejections than those using a single-attempt policy because multiple samples are involved. In an interoperability test, multiple attempts may mask an underlying interoperability problem.

EXAMPLE 1 An online test might define a transaction to mean that live users may make up to three attempts, with a yes/no access decision provided to the user after each one.

EXAMPLE 2 An offline test might provide a face recognition engine with up to three images of a user. Each of these could be used in an identification attempt. The provision of the second and third images might occur only after a request from the implementation under test.

EXAMPLE 3 See Annex B for a documented example of an interoperability test and target application.

6.2.1.2 Reporting for identification systems

Operational systems might enrol BDBs prepared by more than one supplier. This could occur, for example, if the sBDB generation process is distinct from the enrolment database function. In such cases, an identification search will proceed over BDBs from different sources. This may complicate analysis. Therefore the test plan

and test report shall describe the interoperability application in terms of whether the enrolled instances are all prepared by one supplier's product or several. The test report shall state the proportions of enrolment BDBs from each generator, and the total.

NOTE Some identification applications may indeed enrol heterogeneously-sourced sBDBs. Such applications require analyses and test methods beyond the current content of this standard (particularly the procedures in clause 8.8.1.3).

6.2.2 Interoperable application

6.2.2.1 Statement of coverage

In an interoperability test, the test plan and the test report shall include a statement of coverage that clearly identifies the scope of the interoperability that it is seeking to assess.

EXAMPLE 1 A border crossing application involves the comparison of an enrolment sBDB from an issuing country's passport with a verification sample captured at a host country's port of entry. The country will deploy a single biometric capture device and comparison subsystem. It elects to conduct a test of several suppliers' products, in order to procure the best performer. During the test the interoperability space has dimension of 2; once the system is deployed the interoperability space has dimension of 1. The statement of coverage might be:

> "This test will measure verification performance using data from the following sources: 1. a set of one or more combined camera and sBDB generators configured to execute an attended enrolment (i.e. passport application); 2. a set or one or more combined camera, pBDB generators and comparison subsystems configured for a live acquisition in an immigration booth. The second system compares the enrolment sBDB to the live sample data and renders a decision."

EXAMPLE 2 Two financial services firms merge, each retaining their installed base of fingerprint biometric capture devices and a comparison subsystem used for logical access. All equipment will be retained post merger, but with firmware revised to write sBDBs instead of pBDBs. The new company conducts a test to assess any performance penalty. The appropriate interoperability matrix is 2 x 2 x 2. The statement of coverage might be:

> "This test will measure verification performance using data from the following sources: 1. a set of two combined biometric capture devices and minutiae extractors producing enrolment templates from those images; 2. the same set of two biometric capture device and minutiae extractors producing verification templates; 3: a set of two comparison subsystems comparing enrolment and verification templates."

6.2.2.2 Dimension of the interoperability space

As discussed in the Figures of clause 6.1, an interoperable application will involve exchange of data between combinations of products from multiple suppliers. The interoperability problem depicted in Figure 1 can be viewed as having five dimensions. The results of a performance test can be viewed as occupying an interoperability space with as many dimensions as there are device or component classes that are neither sole-sourced, nor proprietary, nor already known to be interoperable, in the target application.

Practically a test might subtract, or add, various interoperable components to properly reflect its target application. The dimension of the interoperability space shall be reported.

EXAMPLE 1 In Figure 1 the dimensionality of the interoperability space is 5. Thus each supplier from a group A builds a ten-finger enrolment biometric capture device, each supplier from a group B's product yields ISO/IEC 19794-8 enrolment templates, each supplier from a group C compares those with like verification templates from each supplier from a group D's product, as generated from images acquired by each supplier from a group E's single-finger biometric capture device.

EXAMPLE 2 In Figure 2 the dimensionality of the interoperability space is 3. Thus if each supplier from a group A builds a face camera whose output is enrolled as a ISO/IEC 19794-5 token image by each supplier from a group B's product, then this may be identified by each supplier from a group C's comparison subsystem with a pBDB produced by each supplier from a group C's combined camera and feature extraction algorithm.

EXAMPLE 3 In Figure 3 the dimensionality of the interoperability space is 3. Thus if each supplier from a group A's fingerprint biometric capture device is used with each supplier from a group B's proprietary template generator to populate a smart card then the bearer may be verified by submitting a ISO/IEC 19794-2 minutiae template from each supplier from a group C's generator to each supplier from a group B's match-on-card implementation.

EXAMPLE 4 In Figure 4, with its Note 2 in effect, the dimensionality of the interoperability space would be 2. Thus if each supplier from a group A's converts captured iris images from a database into ISO/IEC 19794-6 polar irides then each supplier from a group B can compare those with proprietary templates it produced from other archived images.

EXAMPLE 5 In Figure 5 the dimensionality of the interoperability space would be 2. Thus if each supplier from a group A's converts captured iris images from a database into ISO/IEC 19794-6 polar irides then each supplier from a group B can compare those with proprietary templates produced from further archived images.

EXAMPLE 6 The most common commercial operational scenario arises when manufacturers of BDB generators team with biometric capture device manufacturers to supply a finished product. Thus if each supplier from a groups A and B make the enrolment and verification products, respectively, then the interoperability space has dimension 2.

6.2.2.3 Number of products

For each dimension of the interoperable application, various numbers of suppliers will elect to participate in the test. The numbers of products shall be reported. The number of suppliers shall be reported.

EXAMPLE 1 Two fingerprint enrolment biometric capture devices are submitted for testing with five ISO/IEC 19794-8 skeletal template generators, three comparison subsystems, and six single-finger verification biometric capture devices. Referring to Figure 1 the values of the component counts are: N = 2, I = J = 5, K = 3 and M = 6.

EXAMPLE 2 An offline test of the ISO/IEC 19794-2 fingerprint minutiae template is conducted. A single universal biometric capture device was used to produce a sample corpus. The test is intended to measure core interoperable capability of comparison subsystems using just sBDBs. Two suppliers provide enrolment template generators, four provide verification template generators and three submit comparison subsystems. Thus I = 2, J = 4 and K = 3 with N = M = 1 because biometric capture device interoperability is assured by conformance to an optical imaging specification. This is depicted in Figure 6.

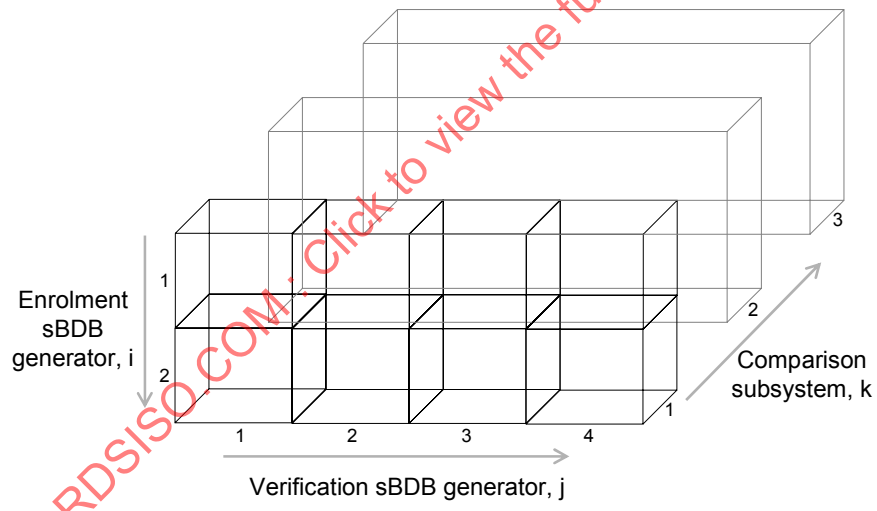


Figure 6 — Cells of an example interoperability space

6.3 Purpose

6.3.1 Interoperability testing

An interoperability test is appropriate to quantify or compare performance levels when standardized data is exchanged, or when biometric capture devices are used interchangeably, or when pBDBs are compared with sBDBs. Possible objectives include:

a) Produce an estimate of performance interoperability;

b) Be a part of an iterative development process in which a standard is developed, subsystems are produced and tested, a consensus on needed modifications is brokered, and the standard is updated. Each test phase will embed a type 1 test;

c) Use an estimate of interoperable performance from a test of type (a) to certify a core group of products as being interoperable;

d) Measure performances against one or more subsystems previously found to be interoperable, such as that produced in a type (c) test. This test is appropriate when one or more products are being evaluated for inclusion on a certified interoperable products list.

e) Make a prediction of operational performance.

f) Assess the feasibility of replacing one supplier's component of a biometric system with another supplier's.

The test type shall be reported. The Clauses of this standard include material specifically for certain of these test types.

NOTE 1 Scenario or operational tests are likely to be most suitable for estimating fielded interoperable performance. A test of type 5 should therefore be conducted using a human population making authentication or identification attempts in the scenario-style codified in ISO/IEC 19795-2. However the size of the population in scenario-like tests is often constrained by the availability of funds and a commensurate increase in the uncertainty of the measured performance should be anticipated. It may be possible to conduct a purely offline type (e) test by using an archived corpus of samples. The accuracy of performance predictions from such tests will depend on the extent to which the data is representative of the intended application.

NOTE 2 If a SIF is amended significantly then test of type (a), (b) or (c) would usually be appropriate. The amendment of a standard is usually undertaken with the knowledge that existing interoperability results, sufficiency results, products, qualification lists, and circulating sBDBs, will become obsolete. In such circumstances a type (d) test is inappropriate.

NOTE 3 The need to conduct such a type (f) test may arise operationally for a number of reasons for example, if the component supplier goes out of business, or if the fielded performance is not acceptable, or if the maintenance fees become prohibitive. The aim of the test would be to measure performance of the system before and after the component replacement. This aim may imply an asymmetry in the goal of the test: supplier B's performance on supplier A's data is of interest, whereas A's performance on B's data is not.

6.3.2 Sufficiency testing

Sufficiency is a measure of the performance of implementations of the standard versus purely proprietary implementations. A sufficiency test may be appropriate when an interchange standard has been newly developed or significantly revised (for fingerprint templates, see [2]). A test of sufficiency requires at least one pBDB generator and one pBDB comparison subsystem. A comparison of proprietary and standardized formats requires that the pBDB and sBDB data shall be generated from common cBDBs. The test shall therefore embed an offline comparison phase.

NOTE 1 It is possible to conduct a test in which only one supplier participates. Such a test would serve only to demonstrate sufficiency of the SIF by expressing SIF-based performance relative to proprietary instance performance. While a conclusion of sufficiency from a single-supplier test is technically consistent with this standard, a multi-supplier test would inevitably allow more robust conclusions to be drawn.

NOTE 2 Figure 7 does not depict exchange of sBDBs between generators and comparison subsystems because only native comparisons are needed in sufficiency testing. In this case I = J = K.

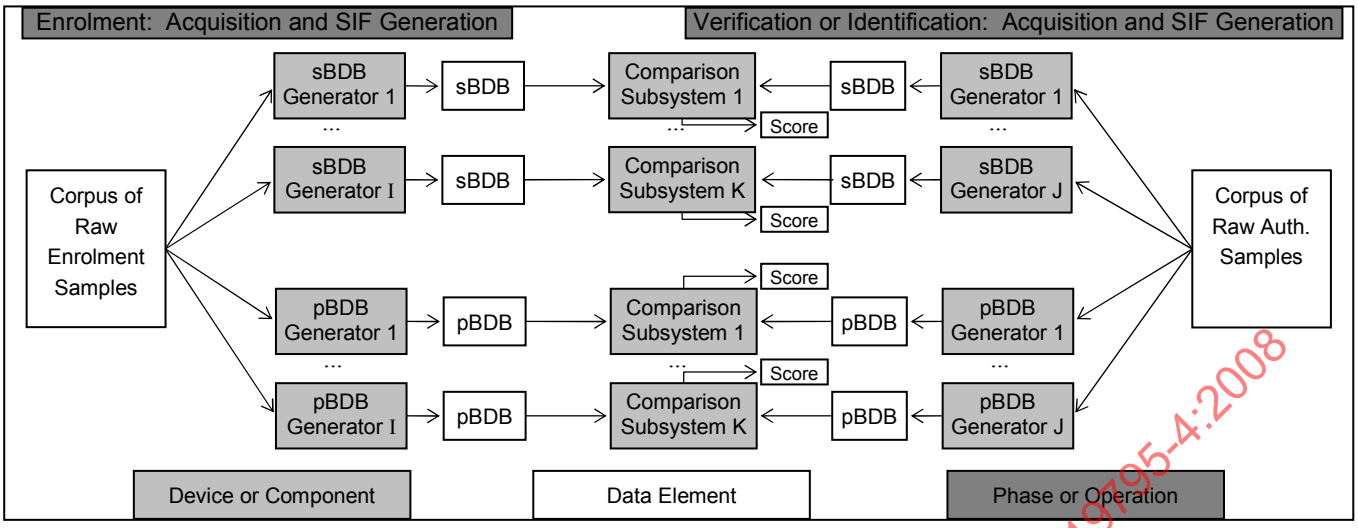


Figure 7 — Sufficiency testing: proprietary vs. standard interchange formats

7 Metrics

7.1 General

The test plan and test report shall state whether the test includes testing of interoperability, sufficiency or both. In an interoperability test, the test shall report values of standardized performance metrics for all cells of the interoperability space covered by the statement of coverage.

NOTE See Clause 2 of this standard for conformance requirements referencing ISO/IEC 19795-2:2006.

7.2 Figures of merit

7.2.1 Recognition performance figure of merit

Measures of performance for determining interoperability and sufficiency shall be defined in terms of one or more underlying figures of merit. These shall be selected to be operationally meaningful and to adequately represent performance. Examples of figures of merit are:

— for verification, generalized false reject rate (GFRR) at a specified generalized false accept rate (GFAR) – mandatory, see below

— for verification, false reject rate (FRR) at a specified false accept rate (FAR);

— for verification, FAR at a specified FRR;

— for verification, GFAR at a specified GFRR;

— for verification, equal error rate;

— for verification, the FAR and FRR values at an operating threshold fixed for the comparison subsystem;

— for identification, false negative identification rate (FNIR) at a specified false positive identification rate (FPIR);

— for identification, FPIR at a specified FNIR.

For interoperability performance testing of verification systems, GFRR at a fixed GFAR is mandatory as one of the figures of merit. This is because different suppliers' products can fail at different stages of processing and, for comparison purposes, it is necessary to included all sources of failure to process in the performance metric.

EXAMPLE 1 A suitable figure of merit for a face recognition system could be FRR at FAR = 0.01.

NOTE 1 It is possible for the set of subsystems meeting specified interoperability criteria to change depending on the threshold setting. This may be obviated by basing the test on an operational performance requirement, such as FAR = 0.01, which would have the effect of a fixing the threshold. In the absence of such a specification a test report shall report interoperability at several operating points.

EXAMPLE 2 Report FRR at FAR = 0.0001, 0.001, and 0.01.

NOTE 2 Results of interoperability tests (whether of absolute or relative performance) do not necessarily indicate the same levels of interoperability for other applications, or in other environments.

NOTE 3 The metrics above correspond to single points on a DET characteristic, and can be operationally targeted if the subsystem can be pre-configured with the corresponding threshold. More general summary statistics such as area under the ROC (one minus the area under the DET curve) and the d-prime statistic are not as operationally relevant because they are computed independently of an operating threshold. But they may be of use for two cases: first as an indication of the biometric "power" associated with a biometric trait, the samples thereof, and an algorithm used to process them; Second in the case where the degree to which systems are interoperable (see clause 7.3) is found to be threshold-dependent and a summary value can be used as a generic (i.e. application non-specific) performance statistic.

NOTE 4 Interoperability testing of commercial-off-the-shelf (COTS) products that do not provide comparison scores, and whose operating threshold is not configurable, may require a figure of merit that takes account of both the GFRR and whether the GFAR is within the acceptable range; e.g., figure of merit, F

$$F = \begin{cases} \text{GFRR} & \text{GFAR} \le 0.01 \\ 1.0 & \text{GFAR} > 0.01 \end{cases}$$

NOTE 5 For identification, the test may benefit from a study of the tradeoff between processing time and BDB size.

NOTE 6 Identification metrics, such as fraction of genuine matches at rank less than or equal to r (as given on a cumulative match characteristic), are also suitable and may be included in addition to the figures of merit listed above.

NOTE 7 Complements of the error rates (for example, TAR = 1 – FRR) are not included here because those Clauses that use inequality terms (such as "smaller than") would have to be reversed (to "larger than", say), and because the operation is trivial.

Methods for interpretation and analysis of such figures of merit (e.g. to quantify performance interoperability) is addressed in Clause 9.

7.2.2 Measuring component failure

In general any component of a biometric system may fail to execute its function. Such component-level failure rates shall be measured and reported in addition to the generalized transactional error rates.

Component-level failure may occur at many stages of processing, for example during

- acquisition, or
- image or signal processing, or
- quality control, or
- data conformance, or
- template encoding, or
- comparison.

These phases are not necessarily distinct. For example quality control may be integral to an image processing step, or may be a by-product of the template encoding step. The examples below correspond to the phases identified above.

EXAMPLE 1 A user wearing a glove attempts to use a vascular imaging system. The user-interface provides no guidance and the system does not detect the problem and initiate feedback.

EXAMPLE 2 A face detection algorithm fails to find a face in the scene. The result is a failure to produce a verification template.

EXAMPLE 3 An iris capture subsystem images a user, segments the left iris, computes its area, and issues a failure because the area of the iris at 0.18 cm^2 is below an internally configured minimum.

EXAMPLE 4 An ISO/IEC 19794-8 fingerprint skeletal generator, designed to process ISO/IEC 19794-4 fingerprint images fails when it is given an ANSI-NIST Type 4 record as input. This is failure by syntactic non-conformance.

EXAMPLE 5 A minutiae extractor may return an error code and not produce an ISO/IEC 19794-2 template if it is unable to find more than three minutiae in the input image.

EXAMPLE 6 An ISO/IEC 19794-5 token face comparator may fail if the eyes are not actually in the locations specified by the standard for a Token image (i.e. failure by semantic non-conformance).

A component failure may be due to user error (as in example 1), inadequate performance (as in example 2), or it may be elective (as in examples 3) or due to incorrect operation of a previous component (as in example 4). Errors may depend on the properties of the input image (as in example 5). Some errors may go undetected (as in example 6, where the eye misplacement may not impede computation of a comparison score but will give rise to a low comparison score or rejection decision). The reason for the failure will not generally become known if the test treats the components as black boxes. This would not be true if the box signals the failure with an appropriate and documented code, message or alert.

A test should measure component-level failure rates. These should be reported in addition to the generalized overall transactional error rates.

NOTE 1 The requirement to test conformance of all sBDBs in clause 8.3.2 is needed because a comparison subsystem may be justified in rejecting non-conformant sBDBs.

NOTE 2 Many biometric tests measure only failure to enrol and failure to acquire. In an interoperability test, which often involves one supplier's output data as another's input, failure measurement is required if a correct understanding of the error rates is needed.

NOTE 3 The generalized transactional error rates embed all the effects of component failures. Generalized error rates allow systems that fail at different stages of operation to be fairly compared.

7.3 Interoperability matrices

7.3.1 General

For each relevant cell of the interoperability application space of clause 6.2.2.2 the interoperable performance shall be measured in accordance with Clause 8 and figures of merit derived. This yields one or more interoperability matrices which shall be included in the test results. Methods for summarising and interpreting the interoperability matrix are provided in Clause 9.

7.3.2 Three-way interoperability with sBDB generators

When an interoperability test involves comparison of BDBs from different generators, the test report shall include, for each comparison subsystem, a matrix of the form in Figure 8. Element F_{ijk} is the figure of merit for comparison subsystem k operating on enrolled sBDBs prepared by supplier i and user sBDBs prepared by supplier j.

Comparison subsystem k	User sBDB generator 1	User sBDB generator 2	User sBDB generator 3
Enrol sBDB generator 1	F_{11k}	F_{12k}	F_{13k}
Enrol sBDB generator 2	F_{21k}	F_{22k}	F_{23k}
Enrol sBDB generator 3	F_{31k}	F_{32k}	F_{33k}

NOTE 1 The subscripts i and j index sBDB generators, and k indexes the comparison subsystem.

NOTE 2 The indices 1, 2 and 3 label distinct products without any connotation of their supplier. The set of suppliers of generators and the set of suppliers of comparisons subsystems might have no intersection. The matrix will generally be rectangular and non-symmetric.

NOTE 3 In the special case where the supplier of a generator and a comparison subsystem is the same, the element gives the performance for sBDBs from the same supplier. It may aid interpretation if such items are assigned row and column indices so that they appear as the diagonal elements of the performance matrix.

NOTE 4 Figure 8 depicts a rank two "vertical" slice of the rank three absolute interoperability space of Figure 6. It contains the figure of merit for the cross-generator comparison performance of the k-th comparison subsystem.

NOTE 5 This kind of reporting is given in [2] for fingerprint templates for each of fourteen fingerprint comparison subsystem suppliers.

NOTE 6 If a reference comparison subsystem is available, then a single table comparing performance of different combinations of sBDB generators may be appropriate.

Figure 8 — Cross-generator performance matrix

7.3.3 Two-way interoperability with sBDB generators

The test shall compute and report all figures of merit for all pairs of generators and comparison subsystems. For any given figure of merit, the symbol F_{ijk} shall be used to denote performance of supplier k's comparison subsystem on supplier i's sBDBs, and these values shall be reported as the performance matrix in the format depicted in Figure 9. If suppliers are required to provide paired generators and comparison subsystems the performance matrix will be square; otherwise the performance matrix will be rectangular and the elements will be subscripted by different indices.

	BDB generator and comparison subsystem 1	BDB generator and comparison subsystem 2	BDB generator and comparison subsystem 3	BDB generator and comparison subsystem 4
Enrol BDB generator 1	F_{111}	F_{122}	F_{133}	F_{144}
Enrol BDB generator 2	F_{211}	F_{222}	F_{233}	F_{244}
Enrol BDB generator 3	F_{311}	F_{322}	F_{333}	F_{344}

NOTE 1 The ikk-th element gives the performance figure of merit of comparison subsystem k on BDBs from generator i. The double k subscript indicates that product k implements both BDB generation and comparison.

NOTE 2 As explained in the text of clause 6.1 (concerning Figure 2 and Figure 3), the generator here may produce sBDBs or pBDBs. Likewise the comparison subsystem may compare those with sBDBs or pBDBs under the constraint that pBDBs cannot be interoperably compared with pBDBs.

NOTE 3 The indices 1, 2, 3 and 4 label distinct products without any connotation of their supplier. The set of suppliers of generators and the set of suppliers of comparisons subsystems might have no intersection. The matrix will generally be rectangular.

NOTE 4 In the special case where the supplier of the enrolment generator and a comparison subsystem is the same, the element gives the within-supplier or native performance. It may aid interpretation if such items are assigned row and column indices so that they appear as the diagonal elements of the performance matrix.

NOTE 5 Native performance of the SIF is compared with fully proprietary performance to quantify sufficiency, see clause 6.3.2.

Figure 9 — Example performance matrix

7.3.4 Fixed operating point interoperability

The performance matrices of Figure 9 state a figure of merit for comparison subsystem A on sBDB instances from suppliers A, B, C etc. In a verification application the figure of merit might reasonably be the FRR for a fixed FAR. However, if comparison subsystem A is configured to use a fixed threshold then both the FAR and FRR will vary depending on the source of the sBDB instances, A, B, C etc. In practice, an operational system will either use a fixed threshold for all input sBDBs, or tailor the threshold to the source. This latter approach requires the enrolment BDB to be associated with its generator, and for some calibration to be available indicating how the threshold should be set. To address this issue an interoperability test plan and test report shall document the threshold setting policy of the target application. If the application includes source-specific thresholds then the performance matrix of Figure 9 will be a sufficient statement of performance. If, however, a fixed threshold is considered then the test shall report performance related variables at that threshold that gives a specified value of the figure of merit for one particular system.

EXAMPLE If supplier X can match its own sBDBs with a FNMR of 0.02 at a fixed FMR of 0.01, then the test report would include the value of both FNMR and FMR for supplier Y's sBDBs also. Typically both of these values will depart from the native X values, so FNMR may be 0.022 and FMR may be 0.008.

7.3.5 Reporting failure of sBDB generators

When a generator fails to produce a sBDB from an input sample the result is a failure to enrol or a failure to acquire. These shall be handled according to the requirements of clause 7.2.3. Note that in an interoperability test a failure to enrol will have the consequence that error rates such as FRR and FNIR will be altered for all comparison subsystems. Failure to enrol rates should be reported for each generator.

EXAMPLE Suppose a supplier generator fails or otherwise elects not to convert 4% of captured image sequences into conformant ISO/IEC 19794-6 polar image instances. Even if comparison subsystems are able to correctly reject impostors and accept genuine users, the false reject rate will still be 4%.

7.4 Proprietary performance

In a test in which suppliers produce and match their own (i.e. potentially non-standard) images, signals or templates, all figures of merit shall be computed. For any given figure of merit, the symbol P_{kk} shall be used to denote performance of supplier k's comparison subsystem on its own pBDBs. These on-diagonal elements are necessary if a test seeks to quantify sufficiency; they may be assumed to reflect the maximum-effort performance available from that supplier on the given corpus. Together these values shall be reported as the proprietary performance matrix in the format shown in Figure 10.

The off-diagonal elements of the proprietary performance matrix are usually unavailable because the pBDBs are generally non-interoperable. A test may seek to assess and document the extent of interoperability by, for example, examining the proprietary instances, or by running the various comparison subsystems on the proprietary formats.

NOTE Any differences between pBDB and sBDB performance may be due to differences in the computational resources used in preparation of those instances or in the matching process. See clause 8.6.4.

	Supplier 1 proprietary comparison subsystem	**Supplier 2 proprietary comparison subsystem**	**Supplier 3 proprietary comparison subsystem**
Supplier 1 enrolment and user pBDB generator	P_{11}	NA	NA
Supplier 2 enrolment and user pBDB generator	NA	P_{22}	NA
Supplier 3 enrolment and user pBDB generator	NA	NA	P_{33}

Figure 10 — Proprietary performance matrix

8 Conducting a test

8.1 Structure of test

A test shall execute genuine and impostor transactions and these may be conducted online or offline. A sufficiency test shall be based on execution of offline transactions. Thus the test shall be conducted

— entirely online (with acquisition and verification or identification transactions as an integral part of the test), or

— in a hybrid manner in which samples are acquired before a separate phase embedding the verification or identification transactions, or

— entirely offline (using archived samples).

The test plan and test report shall document which of these approaches is used.

8.2 Sample data

8.2.1 Acquisition

8.2.1.1 General

Sample data for interoperability testing may be collected online or may be available offline. Dedicated online data collection offers the best opportunity to collect data in an environment representative of the application and in a manner representative of the protocol (for example three attempts). Online acquisition of data is usually the best means of constructing a corpus for evaluations that are intended to give performance estimates as predictive as possible of the fielded performance of systems. Offline data consists of archived samples of data collected previously perhaps in an operational setting. It may be available in very large quantities.

NOTE Offline testing is appropriate if the goals of the test include an examination of the causes of poor interoperability, an analysis of the limits of a SIF, or a comparison of algorithmic functionality. Offline trials are repeatable, can be scaled to very large populations, and can be conducted with calibrated or deliberately processed laboratory data.

8.2.1.2 Offline acquisition

Offline data sets used in interoperability testing may originate in an unrelated collection phase, or may be collected specifically for the test. Offline data for which the biometric capture device is unknown shall not be used in biometric capture device comparison or biometric capture device interoperability trials.

If samples have been excluded after initial data collection effort and before delivery to the testing organization then the test shall only proceed if the fraction of samples so rejected is known, documented and included in the computations required by clause 7.2.2.

EXAMPLE Some biometric capture devices embed a supplier's quality assessment algorithms and reject (fail to acquire) samples, on the grounds that they're not suitable for matching.

8.2.1.3 Online acquisition

If a test includes the acquisition of samples from a live population then the collection should conform to the provisions of ISO/IEC 19795-2 that apply to online acquisition. However, this collection need not include biometric verification or identification trials if, instead, those transactions will be later conducted offline. Such an activity is termed hybrid testing.

NOTE Clause 7 of ISO/IEC 19795-1 gives requirements and guidance on data collection.

8.2.1.4 Hybrid acquisition

It will be particularly beneficial if the exact circumstances surrounding the online capture of samples can be logged. This information will support offline "replay" of the sequence of operations used during enrolment,

verification or identification attempts. This is useful in interoperability testing to ensure that each supplier's sBDB generator and comparison subsystems are tested on all samples in a verifiably equitable and repeatable manner. The test plan shall therefore define procedures and data formats to support transactions. These shall include formats and mark-ups for recording the temporal sequence of user actions (e.g. finger placements), biometric capture device responses (e.g. authentication feedback results), and storing of samples or features.

8.2.1.5 Biometric capture device performance testing

Online acquisition shall be conducted if the test scope includes a biometric capture device interoperability component. This requirement can be waived if an existing corpus of offline data is available, for which the biometric capture device is known.

8.2.2 Representative data

An offline test shall use a corpus of acquired samples, usually containing at least two from each person, which is representative of the intended applications. It may enhance the usefulness of the test to include data from sources beyond any immediate target application. This will be particularly true for newly-standardized formats for which limited public testing has been conducted. The test organization might repeat the tests defined in this standard with dedicated application-specific data sets, and report each repetition (see also clause 8.6.4).

8.2.3 Collection of ancillary data

If a test organization can identify covariates, such as subject-age or acquisition environment, for which interoperability is either known, or expected, or found to be particularly sensitive to, then the test should be expanded to include further trials and analyses.

It is imperative that a test design should establish which covariates to collect before collection begins because it will be either impossible or very difficult to recover this information after collection. For example, environmental factors (e.g. humidity) may not be available at all, and population-specific variables (e.g. eye colour) will only be available from that portion of the population who can be contacted after the fact.

NOTE The database could be purpose-built (see example 1 below) or could be some archival imagery gathered over many years, possibly without any intention to use automated biometric recognition methods (example 2).

EXAMPLE 1 In a fingerprint-based logical or physical access control application with a particular biometric capture device, the most representative test data would, by definition, be those images captured on site with that biometric capture device. Otherwise images from a different site with the same biometric capture device may be suitable.

EXAMPLE 2 Passport face images from the 1960s.

8.2.4 Corpus size

In a test that seeks to quantify both interoperability and sufficiency the question of whether the SIF is significantly inferior to proprietary formats will arise. Sufficiency aside, the same issue will arise for the various interoperability measurements. The test designer should ensure availability of data sets large enough to resolve small differences in the chosen figures of merit. This test may use all of this material or not, depending on the measured performance and on formal estimates of the confidence intervals.

8.2.5 Removal of subject-specific metadata

The test shall remove subject identifying information from the acquired samples. This should apply to any biographical information, for example date of birth, which may heuristically indicate whether two samples are from the same or different persons.

8.2.6 Removal of unrepresentative metadata

The test shall remove any information from the acquired samples that would not be available to a system in the context of the intended application.

EXAMPLE The eye coordinates in a face image might be present in the header of a ISO/IEC 19794-5 instance, but would often not be available to an application.

8.2.7 Origin of samples

The test report shall document the origin of the samples used in the test. Such documentation should at least include the number of samples and individuals, and when available, the biometric capture devices used in the acquisition and any relevant physical characteristics of the samples (compression ratio, sampling frequency, resolution, colour space, etc.).

NOTE The samples should preferably have been acquired using the same biometric capture device, in the same environment, such that the corpus is homogeneous.

8.2.8 Untainted samples

The acquired samples used in the test shall not have been pre-processed, filtered, restored or enhanced by any supplier participating in the test. The tester shall verify that none of the original samples have been discarded by a participating supplier, prior to delivery to the tester.

8.2.9 Sequestered data

In an offline or hybrid test it is often productive to provide some sample data to participants for development purposes. However, the test itself shall use a disjoint part of the corpus which shall not be provided to the test participants at any time prior to the conclusion of the test.

8.3 Conformance testing

8.3.1 Conformance

An interchange test report shall include a description of the SIF, noting its title, documentary source, nature, origin, age, maturity, and availability of implementations. The test report shall reference any documentation of relevant prior conformance tests, any known conformant implementations, and cite any evidence regarding the ability of the SIF to be implemented. This requirement may be waived for tests that do not involve sBDBs (e.g. biometric capture device interoperability tests using only pBDBs).

NOTE 1 Conformance to a biometric interchange standard does not guarantee interoperability. This departs from some other fields in that matching performance is also heavily dependent on non-standardized factors such as algorithms, biometric capture devices, environment etc.

NOTE 2 A performance test might not meet its objectives if any of an underlying stack of standards is either poor, or poorly implemented. Low-level data interoperability is clearly required for a performance test to produce meaningful results.

8.3.2 Executing conformance tests

A SIF performance interoperability or sufficiency test is likely to fail and give erroneous results if the sBDB instances are not conformant to the underlying SIF. This is because a set of uniformly understood and implemented definitions is a necessary condition for sBDBs to be exchanged. Therefore an interoperability or sufficiency test shall assess the conformance of all sBDBs generated during the test. This requirement may be waived for tests that do not involve sBDBs (e.g. biometric capture device interoperability tests using only pBDBs). This requirement may be waived for those sBDBs that are not used in the computation of figures of merit.

NOTE 1 The requirement for conformance testing of all sBDBs (instead of just an initial sample) is needed because some conformance problems will be data dependent and will only occur when certain perhaps unusual input samples are input to a generator.

NOTE 2 It will usually be beneficial for the test to include an initial conformance testing phase. Supplier's implementations, possibly submitted in preliminary form, may be used to generate sBDBs from trial datasets geared toward debugging. These might include empty or poor images, or in an online situation a deliberately defective presentation. In addition a test might assess conformance by acquiring and inspecting sample sBDBs from prospective participants. This method, while ad hoc, offers the possibility of rapidly detecting obvious impediments to a direct progression to performance testing.

8.3.3 Reporting

The interoperability test report should state whether the conformance of generators to the standard interchange format was tested.

8.4 Constraints on the sBDBs

8.4.1 Optional encodings

If an interchange format has standardized but optional (either-or) formats or parameterizations, the test plan shall fully specify the allowed, disallowed, undefined, required or optional values for each conditional.

EXAMPLE 1 A test of the face image interchange format test may require the sBDBs to conform to just full frontal or token images.

EXAMPLE 2 A fingerprint minutiae may be in the normal or compact format. This also applies for fingerprint spectral and skeletal pattern formats.

EXAMPLE 3 A fingerprint image test might require images to be compressed with ISO/IEC 15444 Parts 1-10 (i.e. JPEG 2000).

8.4.2 Optional encodings from profile standards

A test plan shall consider whether existing relevant application profile standards contain appropriate or useful specifications of the optional content of the interchange format.

NOTE It may be worthwhile to consider the Biometric Profiles for Interoperability and Data Interchange standards, ISO/IEC 24713-x ($x \geq 2$), which give complete specified values for the optional content of each ISO/IEC 19794-x part.

EXAMPLE Minutiae records with valley-skeleton bifurcation points should not be compared with minutiae records with ridge-skeleton end points, even though all these options conform to ISO/IEC 19794-2.

8.4.3 Deviation from the base standard

A test plan shall describe any allowed deviations from the SIF. Deviations should usually be confined only to the header fields and not to the material functional data. The tester may judge such determinations by referencing of the goals of the interoperability test.

EXAMPLE 1 It may be necessary to remove the Creator and PID fields as defined by the ISO/IEC 19785-1, Common Biometric Exchange Formats Framework (CBEFF) standard.

EXAMPLE 2 The use of uncompressed image data encoding is not permitted for ISO/IEC 19794-5 face image types. An interchange test that uses this though non-compliant to the base image standard, may be necessary if, for example, compression had been shown to cause interoperability problems.

8.4.4 Data encapsulation

The test plan shall specify data representations for acquired samples and sBDBs. The clause may require the tester to produce, publish, and seek comment on, extensive documentation of the various formats.

EXAMPLE 1 The acquired sample formats might be simply files in a standardized sample format such as JPEG (ISO/IEC 10918 Parts 1-4). The sBDBs might also be files.

EXAMPLE 2 Given that the ISO/IEC 19794-x standards define sBDBs at CBEFF's biometric data block (BDB) level, they can be used as is. They could equally be wrapped with header and signature blocks to produce a simple CBEFF biometric information record structure.

NOTE The BDB may contain one or more biometric samples or templates. Beyond that, the complex CBEFF structure (ISO/IEC 19785-1) allows for a record that can contain multiple BDBs, each having its own standard biometric header (SBH), plus additional SBHs that express the relationships among the BDBs. As such, a complex biometric information record could theoretically contain ten fingerprints, the minutiae sBDBs thereof, multiple iris codes or some arbitrary combination of multimodal multi-sample biometric data. This standard can be used to evaluate any such complex instance.

8.5 Components

8.5.1 Components for sufficiency testing

A test organization conducting an interoperability test shall prescribe whether a test participant is required to produce both a generator whose output can be matched by others and a comparison subsystem that can perform on others' inputs or whether either function can be sufficient. This question is essentially a commercial one. If the market for sBDB generators is separate from the market for sBDB comparison subsystems, then a supplier may seek to enter one but not both roles. In such cases analyses of the interoperability matrices shall allow for $K \neq I$. Comparison subsystem in this context may refer to apparatus for verification and/or identification.

8.5.2 Establishing modularity requirements

The test plan should establish at what level a black box is decomposed into distinct internal black boxes. This may depend on the interoperability goals of the test.

EXAMPLE A black box component might consist of a combined hand geometry capture subsystem and a ISO/IEC 19794-9 conformant sBDB generator. Alternatively the two functions may be separated into two black boxes with a raw image output from the reader being input into the sBDB generator. These are functionally equivalent except that the failure rates may be measured separately. In an interoperable situation, the acquisition device output may be fed into several generators (offline say) and this decomposition is vital in stating performance of the components.

8.5.3 Components for interoperability testing

If a test seeks to assess interoperability, the test plan shall state which of the following must be provided by suppliers:

— an enrolment sBDB generator;

— a user sBDB generator;

— a sBDB comparison subsystem.

NOTE The enrolment and user sBDB generators might be functionally identical and share the same invocation.

8.5.4 Underlying algorithms

A test may seek to evaluate the performance loss attributable to just the SIF versus the PF by using the same underlying comparison algorithms. In this case the test plan and test report shall document steps taken to ensure that each supplier did embed the same core comparison algorithms. Such steps might include written requests or instructions, inspection of source or compiled code, and analyses of timing and results. Such a constraint would be useful in comparing just the two data formats. However it might not be immediately possible for a supplier to comply. Another consideration is that because of the black box nature of most implementations the test organization may not be able to detect departures from a same-algorithm requirement. If the SIF and PF comparison subsystems do not use the same underlying algorithm, then it may not be possible to attribute any insufficiency of one of the comparison subsystems to the data interchange format in use.

NOTE This standard also covers the use of multiple algorithms to the extent that together they comprise a single black box comparison subsystem that internally fuses data together. An assessment of the efficacy of a single component of such a comparison subsystem will require cooperation of the supplier.

8.5.5 Capture device user interfaces

The user interface is an important component of acquisition. A good interface may give improved performance. In an interoperability test, if two acquisition subsystems, consisting of identical hardware and different user interfaces, give different performance then this presents an interoperability problem. A test report shall document interface differences and any recorded interoperability effects.

8.5.6 Multimodal components

The outcome of a test of multimodal sBDBs might be that interoperability can be achieved. This conclusion may be specious because a lack of interoperability in one or more of the modes may be hidden by the black box nature of the comparison subsystems. Therefore a test of multimodal sBDBs shall document in a test plan and test report the steps taken to determine whether only a single mode is internally interoperable. Such measures may include requiring generators and comparison subsystems to be capable of producing and accepting single mode samples, in addition to the primary multimodal sample. These may be assessed separately according to the provisions of this standard.

8.5.7 Component variability

The performance a biometric system deployed at a number of sites may vary even if the configuration of each is thought to be identical. This will be the result of multiple sources of variation. These might include environmental effects (e.g. indoor vs. outdoor illumination), population effects (habituation, demographics, etc), manufacturing variability (e.g. of a biometric capture device embedding a charged coupled device camera), calibration (linearity of response with temperature in an infra-red device), and configuration changes (e.g. different driver software) or errors (e.g. in the operating threshold or the width of bandpass filter). Measurement of performance may itself introduce variation, if the testing protocols are themselves not uniform. A test will give more robust results if multiple copies of a component are tested. The test design shall consider the feasibility of increasing the number of copies. Factors such as labour and cost will need to be considered. The test report shall document known sources of variability, any attempts to measure them and any attempts to mitigate them.

EXAMPLE A test may measure the interoperability of one component, with an identical copy of itself. So considering a special case of Figure 5 Biometric capture device interoperability, two copies of a single fingerprint biometric capture device model could be used to capture images from a population. These could be matched offline and error rates compared. Any significant difference between inter- and intra-device performance would indicate inconsistencies between the biometric capture devices.

8.5.8 Component reporting requirements

An interoperability test might involve several copies of components provided by different suppliers, each with a particular configuration. The future applicability of the test results may depend on an adequate description of what has been tested. A test report shall include statements of the following.

- The components in use. The term component includes acquisition device, processing software, encoder or generator, comparison subsystem. This definition should be expanded to include any distinct component that has, or is thought to have, a material effect on the performance or is involved in interoperable exchange.
- Full identification of each component. This will generally include the name of the manufacturer, the model number, the version, edition or series number, firmware version or build numbers, and any driver version and configuration information associated with a host computer.
- The number of instances of each component.

If a test is conducted with products whose supplier must remain anonymous, then any published test report shall record at least the number of instances of the product and should report any specifications that do not uniquely identify it.

8.6 Planning decisions

8.6.1 Computational intensity

For a N person test population, the test design shall include an estimate of the processing time required to execute the test. The total number of sBDB or pBDB generation operations and the number of verification or identification transactions shall be set to satisfy any constraints on total time, expense, and available population size and resource availability. Estimates may be needed for the times required to

— generate enrolment sBDBs (for each person the first is notionally the enrolment sample) using each enrolment sBDB generator,

— generate user sBDBs (the remaining samples represent the verification samples) using each user sBDB generators,

— perform verification comparisons, for all possible supplier-generator pairings using each comparison subsystem,

— perform identification searches, with each comparison subsystem, against the population enrolled by each enrolment generator, using user instances from each user sBDB generator,

— generate enrolment pBDBs from each enrolment pBDB generator,

— generate user pBDBs from each user pBDB generator,

— perform verification comparisons for each pBDB supplier,

— perform identification searches for each pBDB suppliers.

NOTE The test organization may solicit throughput rates from likely participants. Participants may need information on the sample data to make these estimates. More sophisticated throughput models with needed coefficient information may also be sought.

8.6.2 Supplier recruitment

If the number of enrolment sBDB generator is I, the number of user sBDB generators is J, and the number of comparison subsystems is K, then:

— a sBDB comparison subsystem interoperability test requires each comparison subsystem to process sBDBs from more than one source, so $I \geq 2$, $J \geq 2$ and $K \geq 1$, and

— a sBDB generator interoperability test requires a generator's output to be handled successfully by more than one comparison subsystem, so $I \geq 1$ or $J \geq 1$, and $K \geq 2$

8.6.3 Provision of samples to suppliers

The test plan should specify whether or not suppliers will be given samples for development. The suppliers shall not be given the actual test data.

It may be appropriate for a test organization to provide reference samples (sBDBs, pBDB, or input samples, for example) to test participants. This would support development of interoperable implementations. The test organization should consult with the owners of such data as to how or whether the data may be shared.

8.6.4 Equivalency of generator resources

The test plan shall establish upper and lower bounds for at least

— the amount of storage for a sBDB (in memory and/or on disk),

— the time needed to generate a sBDB,

— the time needed to match sBDBs.

These bounds may be either enforced in the worst case (for example, size is always less than 257 bytes) or on average (for example, the median size will be less than 200 bytes). Maximum limits are easier to implement, adhere to and measure for data sizes. Limits on average performance are often more appropriate for processing times. But in all cases the quantities and their bounds shall each be accompanied by a statement of whether the limits apply to the means, medians, minima, or maxima or some other statistic.

NOTE 1 A test may put requirements on timing so loose so as to be effectively absent yet compliant with this Clause; i.e. the upper and lower bounds may be zero or infinity.

NOTE 2 If a supplier is allowed to devote considerably more (or fewer) resources to the generation of instances of its proprietary format than to the sBDBs, then relative performance may be tailored.

NOTE 3 In the simplest case, proprietary and sBDB generators will differ only in their output / formatting codes, and will accordingly differ very little in their resource requirements.

8.6.5 Handling violations of test requirements

The test plan shall establish policies and appropriate penalties for dealing with generators or comparison subsystems that violate the resource constraints of the clause 8.6.4.

8.6.6 Comparison subsystem output data encapsulation

The test plan shall specify data representations for the outputs of comparison subsystems. For verification this specification shall cover, at least, comparison scores. For identification this specification shall cover, at least, candidate lists.

8.6.7 Fundamental generator requirement

8.6.7.1 Functional properties

A test shall regard

— a sBDB generator as a black box that converts acquired biometric data into sBDB instances, and

— a pBDB generator as a black box that converts acquired biometric data into pBDB instances.

8.6.7.2 Generator implementation

An interoperability test shall be implemented at either the

— executable level - a compiled and linked application capable of accepting an acquired biometric sample and writing a sBDB or pBDB to a file. A large scale test shall use a scripting language to invoke the executable.

— API level - a library that provides a suitable class (function) instances of which can be constructed (called) from (with) an arbitrary sample of acquired biometric data and accessed to provide a standalone sBDB.

NOTE Various parts of clause 8.7.2 on gaming are relevant in consideration of the above choice.

8.6.7.3 Failure to process

The test plan shall establish appropriate mechanisms for each component to declare a failure. This may necessitate documented interaction with suppliers before testing begins. Once a test is underway certain other errors may occur. The test organization should measure and report the occurrence of such events, and include documentation of their nature.

NOTE An outright crash or failure of a component will usually be unacceptable and the supplier would normally be required to resubmit the affected component.

EXAMPLE The test plan might define some non-zero error codes that a SIF generator should return. When this occurs matching might proceed using an empty sBDB (if such is validly part of the SIF) as a default.

8.6.7.4 Generator error logging

The failure or refusal of a generator to produce an output shall be counted and used in the computation of a failure to enrol and or acquire. However, a generator may produce an output but also indicate some problems in doing so. The test plan shall therefore establish a mechanism by which a generator may report problems encountered when processing an acquired sample.

NOTE The test might allow an integer error code to be returned; the meaning of the various values would be defined in the planning and comment phases. The incidence of the various warnings may have value in identifying implementation problems or ambiguities in the SIF.

8.6.8 Fundamental comparison subsystem requirement

8.6.8.1 Functional requirement

An interoperability test shall regard a verification comparison subsystem as a black box capable of comparing a user sBDB with an enrolment sBDB to output a comparison score. Similarly a comparison subsystem in a biometric identification system shall be regarded as a black box capable of comparing a user sBDB with a set of enrolled sBDBs to produce a candidate list.

8.6.8.2 Comparison subsystem implementation

An interoperability test shall be implemented using one or more of the interfaces enumerated below.

a) Executable level - compiled and linked application capable of accepting two arbitrary sBDBs that are stored as standalone files.

b) API level - a library that provides a class (function) instances of which can be constructed (called) from (with) two arbitrary sBDBs and accessed to provide (that returns) a comparison score.

NOTE Various parts of clause 8.7.2 on gaming are relevant in consideration of the above choice.

8.6.8.3 Comparison subsystem errors

The test plan shall establish a mechanism by which a comparison subsystem can declare a refusal to process the inputs.

NOTE 1 An outright crash or failure of a component will usually be unacceptable and the supplier would normally be required to resubmit the affected component.

NOTE 2 The test might allow an integer error code to be returned; the meaning of the various values would be defined in the planning and comment phases.

8.6.9 General requirements on software implementations

8.6.9.1 Invocation

The components should be separated to express the interoperable paradigm and mimic the logical separation of the three functions (enrol, user template, matching). Practically the separation allows an offline test to be run in modular stages, with flexibility in the scheduling of the operations, and in assessment of the conformance and storage of the sBDBs.

In an interoperability test the sBDB generators and comparison subsystems shall be distinct, separately invoked, entirely independent of one another.

8.6.9.2 Side effects

The generator and comparison subsystem shall not change their operating environment, other than in ways explicitly allowed.

8.6.9.3 Memory access

Implementations shall not access memory locations other than those pointed to by the calling implementation. Such activity may be useful to a gaming strategy. System stability will also, clearly, be degraded if out-of-memory access occurs.

Implementations shall access only that system memory that it allocates or that corresponds to the provided inputs.

8.6.9.4 Communication

Unless explicitly permitted in the test plan, implementations shall not communicate with external processes, devices, or computers. Neither receipt nor transmission of information from or to another source is needed for correct function. It is disallowed here because a variety of unrealistic performance improvements could be realized.

8.7 Prevention and detection of gaming

8.7.1 General aspects

8.7.1.1 General

An interoperability performance test design and administration shall embed appropriate means to prevent, detect, and obviate any mechanism by which one or more suppliers may seek to advantage themselves, disadvantage others, or misrepresent available performance.

8.7.1.2 Assessment of gaming risk

The amount of effort a tester should expend on prevention or detection of gaming may well be determined by a consideration of the risk reward trade-off for a supplier and the possible modus operandi. The testing laboratory should assess gaming risk and document it.

The test should be designed with consideration of the advantages to suppliers of successfully implementing any of the gaming strategies.

EXAMPLE The interchange format standard might itself be undermined if sufficiency statistics are manipulated (downwards).

8.7.2 Modes of gaming

8.7.2.1 General

The test plan should establish appropriate steps to address the gaming risks described in the remaining subclauses of 8.7.2.

Note that gaming techniques can be used probabilistically (i.e. on a fraction of the samples or transactions) and yet still be effective. Therefore, if any means of detecting gaming are applied, they should be applied across all samples, trials, suppliers and instances.

NOTE This agrees with the clause 8.3.2 requirement to test conformance of all sBDBs.

8.7.2.2 Cartels

It is possible that more than one supplier will unite for the purposes of disadvantaging one or more others. Such a collaboration constitutes a cartel. A test should take appropriate steps to identify cartel behaviour.

8.7.2.3 Exploitation of test environment to alter performance

Tests implemented at either the API or executable level shall be implemented with considerable attention paid to information hiding and prevention of heuristic attempts at identifying match or non-match information. The API level implementation presents some hazards: For instance, if matching pairs of sBDBs were stored contiguously in memory then a library could implement a gaming strategy in which proximity of memory addresses of the two BDBs is used as a (perhaps collateral) factor in reporting a high or low comparison score. Such a strategy can be defeated by careful randomization of the BDB memory locations and the sequence of calls.

8.7.2.4 Acquired sample pass through

If a test uses a data structure (a class, in the object oriented sense) to contain a sBDB, the test should institute appropriate checks to detect whether the acquired biometric data sample is being stored as-is inside the sBDB. Such a gaming strategy would allow matching to be done entirely bypassing the sBDB. The test specification should be written such that the sBDB class can be written, examined offline, and read in before use.

EXAMPLE A supplier implementation appends the original image to the end of a fingerprint minutiae sBDB that is otherwise perfectly conformant to ISO/IEC 19795-2.

8.7.2.5 Proprietary data pass through

The test should include any checks to detect if a generator is appending, or otherwise hiding, its supplier's own proprietary data inside the otherwise standard sBDB. This would allow the comparison subsystem to invoke its own (assumed to be) better pBDB comparison subsystem, thereby exhibiting better performance.

NOTE Some data interchange formats include optional structures for proprietary or undocumented data. Obviously such data can offer performance benefits only to those able to understand it, and may thus skew the interoperability test. A test design should therefore include constraints on the presence and/or contents of these structures (see also clause 8.2.5).

8.7.2.6 Polluted sBDBs

An interoperability test should include checks to detect if a generator is introducing erroneous or spurious data into its output sBDBs.

EXAMPLE A fingerprint minutiae test should include checks to detect if a generator is introducing egregiously false minutiae into its fingerprint minutiae sBDBs. If the location of a minutiae were hardwired, say, that supplier's comparison subsystem could ignore it during matching.

8.7.2.7 Truncated sBDBs

An interoperability test should include checks to detect if a generator is excluding information from its output sBDBs that would ordinarily be included or offer substantial benefit to a comparison subsystem.

EXAMPLE It may benefit a fingerprint minutiae supplier to include fewer minutiae in the sBDB than it finds, or are actually present. This strategy could be effective if that supplier's comparison subsystem is superior for low minutiae count templates.

NOTE It may be difficult to establish criteria against which the activity described in the clause may be judged conclusively.

8.7.2.8 Supplier identifying information

Many BDBs include a field for identification of the product that generated them. For example, some of the ISO/IEC 19794-x data interchange format standards include a capture device type ID field. Operationally this field supports

— BDB generators that tailor their processing to the biometric capture device, or

— comparison subsystems that tailor their processing to the particular generator of the sBDBs, or

— clerical activities.

An interoperability test will be more representative of the target application if the BDBs include whatever information is specified in that application. Therefore, for both biometric capture device and SIF interoperability tests, the default practice shall be to follow the operational specification for such fields. This will usually entail inclusion of conformant and correct product identifiers. However if a test seeks to assess the interoperability of just the core biometric data, then it may be appropriate to require that

— generators not include any identification of themselves in their outputs, or

— the test organization should expunge such information from all samples before they are sent to comparison subsystems.

This option implements a purely blind test of the sBDBs and is appropriate when evaluating the exchange of the core technology. This option is likely to weaken a prediction of operational performance.

In any case, the test report shall state any requirements on the biometric capture devices or generators to omit or otherwise modify any fields that identify the source of the BDBs. Likewise any activities conducted by the test organization in this area shall be reported. Also clause 8.4.3 requires documentation of departures from the SIF.

NOTE 1 Any steganographic technique can be used to allow a comparison subsystem to determine if the sBDB is "one of ours". At its simplest this involves hiding a single bit in the sBDB header or data, and is very difficult to detect. A more classic steganographic technique would be to insert some pattern into the low order bits of a fingerprint sBDBs minutiae locations. Such activity makes it difficult to thwart a supplier who is determined to violate a blind testing requirement.

NOTE 2 To implement blind testing, then supplier-identifying fields in, say, a CBEFF header might need to be zeroed out after sBDB generation and before verification.

NOTE 3 Conformance tests of sBDBs may need to be altered to reflect test-specific requirements in this area.

NOTE 4 Comparison subsystems may need to be changed in order to tolerate, for example, zeroed-out product ID fields.

8.7.3 Prevention and detection of gaming

8.7.3.1 Planning

The test plan shall enumerate any modes of gaming that are specifically prohibited or that will be tested for. The following subclauses describe some known modes of gaming.

8.7.3.2 Consequences of gaming

The test plan shall enumerate any circumstances under which suppliers will be disqualified from the test. The meaning of disqualification shall be formalized.

EXAMPLE If a supplier's implementation were to ignore the immutable nature of an API function parameter (as indicated by the const keyword of C/C++) and alter the contents of a sBDB during a comparison function this may constitute grounds for discontinuation of use of that product.

8.7.3.3 Inspection of anomalous results

A test should include an inspection of each generator's generated sBDBs, the purpose being to detect anomalous values intended to subvert another supplier's comparison subsystem. Particularly departures from default or expected or reasonable values shall be reported.

NOTE In face recognition if two suppliers produce compressed token (eye-positioned) images that achieve a compression ratio (CR) of 10 by compressing the periphery of the image with CR = 5 and the center "face" region at CR = 35. The effect may be to depress the performance of a third supplier whose technology is more sensitive to compression. A test design should include inspection of the sBDBs.

8.7.3.4 Disclosure of participants

Suppliers shall not be informed of the names, nor the numbers, of participants prior to test completion.

NOTE If a supplier is informed that only one other supplier is participating it may aid (or mitigate risk of detection) certain gaming strategies, including a heuristic classification of which sBDB instances belong to them.

8.7.3.5 Removal of non-essential information

The test organization should assess whether each field of each sBDBs header is needed, corrupt, in use for other than its intended purpose, and shall investigate the effect of stripping or zeroing its content.

NOTE If no significant difference were observed between the sanitized and original sBDB performance this may mean that no gaming was involved or that the stripping or zeroing mechanisms were ineffective at preventing it.

8.7.3.6 Perturbation

Test BDBs may be altered or perturbed by a test organization. The test plan should include a policy on whether to additionally invoke comparison subsystems with perturbed sBDB instances.

NOTE 1 One possible means of exposing gaming techniques involves changing the sBDBs from those originally output by a generator. For example a test may add noise, shift, decompress, or rotate a face or fingerprint image. It may reflect all fingerprint minutiae locations about an axis. Or remove a minutia from the list.

NOTE 2 If perturbation is used, it should augment, not replace the generated sBDBs upon which the performance conclusions of the test report should be based.

8.7.3.7 Reporting

The test report shall document the nature and conduct, and optionally the results, of any tests instituted to detect gaming.

8.8 Test procedure

8.8.1 Primary test

8.8.1.1 Overview

A test may be conducted by executing the procedures enumerated in Annex A in Tables A.1 (planning), A.2 (setup) and A.3 (template generation) followed by either Table A.4 (verification) or Table A.5 (identification), and then Table 6 (reporting). A test shall conduct verification or identification trials or both, according to the goals of clause 6.1 and the figures of merit selected in clause 7.2.1. In that case some procedures would be redundantly repeated, and they may be neglected as needed. Steps in the tables applicable to either sufficiency or interoperability measurements may be neglected per the goals of 6.1.

8.8.1.2 Verification

Verification interoperability may be measured by execution of the procedure in Table A.4. This involves the pairing of mated (i.e. same person) sBDBs from suppliers i and j, the pairing of non-mated (i.e. different person) sBDBs from suppliers i and j, the concatenation of all such pairings, and the randomization of that union. Verification proceeds by running a sBDB comparison subsystem sequentially on all such pairs. Concatenation shall be included to prevent a comparison subsystem from predicting which kind of pairing it will encounter next. Randomization shall be included to prevent a comparison subsystem from predicting whether the next transaction is a mate or non-mate.

8.8.1.3 Identification

Identification interoperability may be measured by execution of the procedure in Table A.5. This involves the enrolment of supplier i's sBDBs into supplier k's identification subsystem, then the running of all suppliers' sBDBs against that enrolled population. This yields a row in the cross-generator interoperability matrix of Figure 8. The matrix should only be computed element-by-element (i.e. thereby avoiding the concatenation and randomization steps in Table A.5) if the target application would involve the comparison subsystem being exposed only to user sBDBs from a single source.

8.8.2 Uncertainty measurement

Each figure of merit attained in an interoperability test is accompanied by an uncertainty. Such uncertainty values, and the correlations between them shall be taken into consideration in assessment of which subsystems are considered to be interoperable

NOTE Annex A of ISO/IEC 19795-1 gives guidance on variance, uncertainty, confidence intervals and the issues surrounding their computation.

8.8.3 Variance estimation

One means of estimating the variance in the figures of merit is to repeat the test of clause 8.8.1 on disjoint datasets drawn from the same source. One means of achieving this is to split the whole initial corpus in disjoint subsets, and applying to them separately the procedure of clause 8.8.1. Thereafter the procedure in Table A.7 may be followed.

8.8.4 Remedial testing

If interoperability is found to be uneven in a first round of testing, a test shall examine the possibility of making changes to the stored sBDBs to investigate whether interoperability can be improved. If such changes are deemed worthwhile, then all affected parts of the interoperability test shall be repeated. Such repetition, if undertaken and reported, shall be documented clearly and noted as being initiated by the testing organization and not the supplier. However, if a supplier is contacted with the aim of improving interoperability, then this interaction should be documented.

8.8.5 Survey of configurable parameters

The test shall be repeated for any previously agreed upon changes in configurable parameters.

NOTE Many sBDB generators can be configured to expend greater time in localizing salient features in an image for the purpose of improving recognition accuracy.

9 Interpretation of the interoperability matrix

9.1 Determination of interoperable subsystems

9.1.1 General

A type (a) test (see clause 6.3) produces an estimate of interoperable performance. A type (e) test, intended to predict operational performance interoperability, can usually be treated as a type (a) measurement test. The type (f) test, too, in which one subsystem is to be replaced with another, is a special case of the type (a) test.

A test seeking to qualify a set of interoperable subsystems (i.e. type (c) in clause 6.3) shall establish performance criteria, specify minimum and maximum values for the numbers of subsystems sought, and establish a procedure for how to resolve situations in which interoperability is confined to disjoint sets of suppliers.

EXAMPLE 1 A set of fingerprint minutia extractors may be deemed interoperable if the ISO/IEC 19794-2 instances they produce may be matched by a reference comparison subsystem with FRR less than 2% for a FAR of 2%.

EXAMPLE 2 A set of ISO/IEC 19794-5 token face generators may be deemed interoperable if their output instances may be matched by any three identification comparison subsystems with FPIR of 2% at FNIR of 50% in a population of 1500.

EXAMPLE 3 A set of ISO/IEC 19794-6:2005 polar format iris generators may be deemed interoperable if their outputs can be interoperably verified by a reference comparison subsystem with FNMR less than 1.2 times that achievable by the same comparison subsystem running on the parent ISO/IEC 19794-6 rectilinear images at a fixed FMR of 0.0001.

9.1.2 Identifying interoperable combinations of subsystems

9.1.2.1 General

The test plan shall establish one or more application dependent quantitative measures of interoperability that indicate whether a comparison subsystem is sufficiently interoperable with a set of generators, and whether a generator is sufficiently interoperable with a set of comparison subsystems. Such measures should exclude systems that do not offer adequate performance, and outliers related to conformance failure.

The test report should specify the criteria used with a justification. The test plan, also, should address this issue, if possible. The methods of clauses 9.1.2.2, 9.1.2.3, and 9.1.2.4 are provided for guidance.

9.1.2.2 Interoperability against a performance target

9.1.2.2.1 Method

A set of subsystems under test shall be considered interoperable if each of the corresponding observed (absolute or relative) figures of merit $\overline{p}$ in the performance matrix supports the working hypothesis that the corresponding honest (absolute or relative) figure of merit is below a chosen threshold, p.

An observed error rate supports the working hypothesis if and only if the corresponding null hypothesis, that the corresponding honest error rate is equal or above the threshold, is to be rejected. Whether or not the null hypothesis is to be rejected shall be decided by a one-sided one-sample z test. Using the equation and constraints given in Table 2 the null hypothesis shall be rejected (and the working hypothesis accepted) if $z > z_a$ where the value of z_a, which specifies the confidence level, is discussed in clause 9.1.2.2.3. If $z \leq z_a$ there is not enough evidence to reject the null hypothesis, and the working hypothesis cannot be accepted.

Table 2 — Sample size adjustment of error rate requirement

Formula for significance statistic	Constraints on the application of the formula	
$z = \frac{p - \overline{p}}{\sqrt{\frac{p(1-p)}{n}}}$	np > 10	$n(1\text{-}p)$ > 10
	where	n = number of observations
		p = tolerable error rate
		$\overline{p}$ = measured error rate

NOTE 1 This computation may be applied to one or both of the Type I and Type II error rates (e.g. FNMR and FMR, respectively).

NOTE 2 The key point is that with the max() criterion systems should not be compared directly with a threshold (as a yes/no decision) but rather the normalized distance below the threshold significance (1.6449, for example) of the measurement is used instead. Application of this test is equivalent to lowering the tolerable error rate for any fixed number of trials, n. For example, with n = 60000 the FNMR ≤ 0.01 requirement becomes FNMR ≤ 0.00933 for 95% confidence via:

$$\overline{p} \leq p - z_a \sqrt{\frac{p(1-p)}{n}}$$

9.1.2.2.2 Reporting of data used in significance test computation

The test of clause 9.1.2.2.1 embeds the binomial assumption in which each trial is independent of others and has a fixed probability of error. These assumptions will not be appropriate in the following circumstances.

— The trials are correlated in some way. This will occur, for example, if samples are reused in many impostor trials.

— The error rates associated with samples vary. This could occur if samples are drawn from populations that differ in image quality or demographics.

Thus if the binomial assumption does not apply, then this shall be reported, and the test shall still be performed, but the results shall be interpreted with more caution.

9.1.2.2.3 Setting the significance level

The value of z_a in the significance test of clause 9.1.2.2.1 is derived from a specification of the desired confidence level of the test. It is then uniquely determined by the inverse cumulative distribution function of the Normal distribution, values of which are tabulated in Table 3. The value for *(1-a)* and z_a shall be reported. It is common for the significance tests to be conducted at the 95% level, however in interoperability testing this level may need to increase as follows. If sufficient samples are used such that there is a (*1-a*) confidence that each element is better than some criterion then it is likely that for large matrices some interoperable pairs will actually have a true error rate larger than the requirement, because there is an uncertainty, *a*, in each result. Referring to clause 6.3, certification tests of type (c) and (d) shall therefore increase the numbers of independent samples with the number of cells of the interoperability space. Thus a decrease in *a* to set the single-cell confidence level will maintain a fixed confidence in the overall matrix result.

Table 3 — Confidence levels of the standard Normal distribution

Confidence level, *100 (1-a)*	*a*	$z_a = -\Phi^{-1}(a)$
90%	0.1	1.28155
95%	0.05	1.64485
97%	0.03	1.88079
99%	0.01	2.32635
99.7%	0.003	2.74778
99.9%	0.001	3.09023

NOTE 1 A measurement of performance and a confidence interval should only be construed to apply to the test population. The specification of a confidence interval does not imply that a system or product will always perform within that range. Rather, it means only that a repeated test of the same product on samples acquired from the same population, and in the same manner, is likely to give a performance measurement in that range.

NOTE 2 The size of confidence intervals is dependent on the population. Offline "technology" tests using large populations offer tight confidence intervals. Scenario tests on the other hand are usually conducted on smaller populations (for reasons of expense). Technology tests are appropriate for assessing core algorithmic functionality (in this case interoperable performance) while scenario tests are likely to be more indicative of operational performance because they capture the interactions of the products with live users.

9.1.2.3 Interoperability relative to performance of a reference system

Interoperability may be determined by considering performance figures of merit relative to the performance of a reference system measured on the same data set. The reference system may be a system using a PF or a single-supplier system using a SIF.

For any chosen figure of merit (e.g. FRR at FAR = 0.001), a set of subsystems under test may be considered interoperable if the achieved figure of merit value is less than some specified constant, c, times the corresponding absolute figure of merit for the reference system. The value of c will depend on the tolerance levels and the statistical significance required.

9.1.2.4 Interoperability relative to the group under consideration

A further alternative to an absolute performance target is to certify on the basis of a target computed from the observed performance data. In this case, the most interoperable set might be determined by excluding those

interoperating combinations whose measured performance is significantly inferior than the mean interoperable performance of the whole set.

The average interoperable performance, μ_I, is defined as the mean performance of all interoperating combinations where the components are from different systems, the interoperable deviation, σ_I, is the standard deviation of the performance of those combinations.

Under this method, an interoperating combination shall be excluded from the interoperable set if the performance of that combination, μ_{CR}, is greater than μ_I with a confidence level of *a*. That is, if $\mu_{CR} > \mu_I + \sigma_I z_a$ where z_a is the 100a-th percentile of the standard normal distribution (i.e. the area under the tail of the standard normal distribution from z_a to positive infinity is *a*).

NOTE While absolute figures of merit are highly dependent on the corpus of samples used for measuring, relative figures of merit are fairly consistent across different corpora of samples [4].

9.1.3 Acceptable numbers of interoperable subsystems

In a test of type (c) (see clause 6.3), the test plan and test report shall state any pre-test requirements for the following:

— the minimum number of interoperable subsystems that are acceptable which would not ordinarily be less than two unless the organization commissioning the test is willing to sole-source;

— the maximum number of interoperable subsystems that are acceptable. It will ordinarily be of maximum commercial benefit to qualify as large a number of subsystems as possible, but various commercial considerations may limit this.

When two or more suppliers participate in a test, the activity of clause 8.7.1.3 shall be conducted.

9.1.4 Combinatorial search for maximum interoperability-classes

This establishment of a set of interoperable products against the criteria established by 9.1.2.2 will necessitate extracting the figures of merit to form submatrices of the performance matrix. This strategy is combinatoric: elements of the matrix will be extracted for all combinations of rows crossed with all combinations of columns. The goal is to determine all maximum interoperability classes. A maximum interoperability class is a subset of subsystems under test such that all its elements are interoperable with each other, and there is no subsystem under test that is interoperable with all elements of this subset, but not included in it.

Maximum interoperability classes can be found by systematically grouping the interoperable subsystems under test based on the fields of the interoperability matrix as follows.

1) Initially, put each combination of subsystems that is considered interoperable based on the corresponding cell of the performance matrix into an interoperability class of its own. This yields an initial set of interoperability classes.

2) Repeat

   i) For each interoperability class, check with each of the other interoperability classes whether the two interoperability classes can be united into one interoperability class. If so, unite them and include the union class into the set of interoperability classes. Two interoperability classes can be united if and only if all their elements are considered interoperable with each other according to the chosen criteria.

   ii) Subtract each interoperability class that is a strict subset of another interoperability class from the set of interoperability classes

   until there are no interoperability classes that can be united.

3) The elements of the final set of interoperability classes are all maximum interoperability-classes.

The test report shall describe the method of searching.

NOTE For R components, the test might consider subsets in decreasing order of size, r, i.e. attempt to find the largest subsets first so r = R, R-1, R-2 … 2.

9.1.5 Multiple interoperable subgroups

The outcome of an exhaustive search for interoperable subgroups against a criterion may be that multiple but distinct interoperability classes exist.

EXAMPLE Given four iris capture subsystems A, B, C and D, and two comparison subsystems X and Y, some possible outcomes of a test would be that only

- A, B are interoperable with X (i.e. a single subgroup),
- A, B, are interoperable with X, and C, D are interoperable with Y (i.e. two disjoint subgroups of equal size), and
- A, B, C, D are interoperable with X, and B, D are interoperable with X and Y (i.e. overlapping subgroups of equal size).

Multiple interoperability classes are likely if the size of the largest one is found to be much smaller than the number of subsystems tested. The general case is that some products are members of multiple interoperable classes, while others attain membership only occasionally.

When multiple interoperability classes are found, a straightforward declaration of the interoperable set is not immediate. The test plan should anticipate this circumstance and, if necessary, establish a mechanism for resolving it. This might include the following.

- Application of the number of products criteria of subclause 9.1.3.
- Perturbation of the qualification criterion. For instance, the FRR = 1% at FAR = 1% could be tightened to give FRR = 0.9%. This would inevitably yield a smaller group of interoperable suppliers. In the opposite direction the FRR requirement could be relaxed to FRR = 1.1% say. Such a strategy may produce a single larger interoperable group (e.g. ABCD). This method is ad hoc and would put the test organization into the position of making retroactive policy decisions that may be prejudicial to organizations that targeted any declared interoperability criterion.
- Use the count of the number of times each subsystem is a member of the interoperable subsets (in the third example above, B and C are members twice). This method is somewhat fairer than the previous one, but is less attractive than the discovery of a single interoperable set.

The ultimate choice of an appropriate resolution mechanism might be determined after an investigation of whether the various alternatives or combinations thereof, lead to the same determination. It may be appropriate to disclose in the test plan the possible use of such mechanisms. If such a mechanism is applied then it shall be disclosed in the test report.

9.1.6 Statistical stability of the test result

If a test is conducted and, for example, an interoperating combination is identified, it will be important to examine whether a different outcome would have occurred if the test data had changed. Particularly suppliers of subsystems found to be non-interoperable might reasonably question whether a small test would, if repeated on another sample (even one drawn from the same population) would have produced a different result. As one means of assessing this stability, the testing laboratory should consider repeating the test on partitions of the test corpus. This approach is advanced in Table A.7 of Annex A.

NOTE The computational cost of such an approach can usually be avoided by building in the needed partitioning into the execution plan before testing begins.

9.2 Interoperability with previously certified products

9.2.1 Decertification considerations

Clause 6.3 enumerates several test types. Among them a type (d) test applies to interoperability with products previously evaluated in a prior type (c) certification test. This kind of test is essential when it is desirable to expand the marketplace of products while maintaining interoperable performance. The test is conducted to determine that

— the sBDBs produced by generators can be successfully used by previously certified comparison subsystems, and

— new comparison subsystems can successfully accept the outputs from previously certified sBDB generators.

There is an asymmetry here because, once deployed, the outputs of certified generators (e.g., ISO/IEC 19794-3 fingerprint pattern templates) are persistent while the outputs of comparison subsystems (e.g. scores or decisions) are not, even if they may have downstream consequences (e.g. erroneous duplicate enrolments). This aspect has consequences for testing: Specifically the test design and test plan of a type (d) (see clause 6.3) interoperability test shall state a policy under what conditions sBDB generators and comparison subsystems will be decertified. The consequence of a decertification might be that the product can no longer be installed and used for a particular application.

NOTE 1 If a test is run and an interoperable set of generators and comparison subsystems is identified (against some criteria) then it may be the case that the testing organization will be required or asked to incrementally assess interoperability of a new product. How this is done will depend on whether the new product is a comparison subsystem or a generator. A comparison subsystem may be run against sBDBs produced and archived in the original test. However the output of a newly submitted generator will need to be matched successfully and this will be predicated on the retention, operability and licensing condition of the comparison subsystems submitted in the original test.

NOTE 2 In an incremental test, performance of a new product being evaluated against a previous list of interoperable products would most simply be assessed by requiring interoperability with a reference implementation which could, for example, be an existing commercial installed product.

EXAMPLE Suppose a group of six generators and four comparison subsystems are certified in an initial clause 6.3 type (c) interoperability test. Suppose further that six new sBDB generators are submitted for type (d) evaluation and their outputs are successfully matched by three of the certified comparison subsystems, but not by the fourth. The test plan would state that a comparison subsystem will be decertified if it fails to match all sBDBs with an FNMR < 0.01 at a fixed FMR of 0.004.

9.2.2 Continuity of testing

Although such follow-on or incremental tests may be conducted by a different test organization, there are cost and effectiveness benefits that accrue from continuity of testing. One problem associated with different-organization testing is that different conclusions may be reached. This might occur because

— different data may be used,

— results on data from the same population will still exhibit sample variance, or

— different products may be used.

Additionally if the data changes, a supplier might alter the product to handle it. For these reasons follow-on testing should be conducted by the same organization.

9.2.3 Interoperability with previously certified generators

Referring to clause 6.3, suppose a prior type (a), (b) or (c) evaluation establishes a set of interoperable generators (i.e. capable of producing sBDBs that can be matched with acceptably low error rates by a number of comparison subsystems). Assume further that those products have since been installed and used such that sBDBs from those generators are enrolled in databases or placed on smart cards. A type (d) test conducted to