INTERNATIONAL STANDARD

ISO/IEC 19795-6

First edition
2012-02-01

Information technology — Biometric performance testing and reporting —

Part 6: Testing methodologies for operational evaluation

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

Partie 6: Méthodologies d'essai pour l'évaluation opérationnelle

Reference number
ISO/IEC 19795-6:2012(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

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 19795-6 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 5: Access control scenario and grading scheme*

— *Part 6: Testing methodologies for operational evaluation*

— *Part 7: Testing of on-card biometric comparison algorithms*

Introduction

Operational tests evaluate complete biometric systems in the targeted operational environment with the target population. Tests may encompass performance monitoring of operational systems or assessment of performance in operational trials.

Operational performance assessment may be based on:

— data collected by the operational system in the course of normal operation;

— additional data collected during normal system use, but with the system running in an “evaluation mode” allowing extra data to be collected;

— data collected with a set of test subjects considered separately from the subject base of the operational system.

Operational evaluation differs from technology or scenario evaluation in that the subject base, environment, and system design are no longer controlled for the purpose of repeatable testing, but vary in accordance with operational use. Examples of uncontrolled variables include the legitimacy of the subject’s identity claim, environmental effects from weather or lighting, or the variability of system use by different individuals.

The overarching goals of operational testing are to measure or monitor operational biometric system performance over a period of time.

Subgoals of operational testing may include:

— to determine if performance meets the requirements specified for a particular application or the claims asserted by the supplier;

— to determine the need to adjust or configure the system to improve performance;

— to predict performance as the numbers of subjects, locations, or devices increase;

— to obtain information on the target population and environmental parameters found to affect system performance;

— to obtain performance data from a pilot implementation;

— to obtain performance data to benchmark future systems.

This part of ISO/IEC 19795 provides the test planning, test conduct, performance measurement, test reporting, and record keeping requirements to be followed during a biometric system’s operational evaluation.

Information technology — Biometric performance testing and reporting —

Part 6: Testing methodologies for operational evaluation

1 Scope

This part of ISO/IEC 19795:

— provides guidance on the operational testing of biometric systems;

— specifies performance metrics for operational systems;

— details data that may be retained by operational systems to enable performance monitoring; and

— specifies requirements on test methods, recording of data, and reporting of results of operational evaluations.

NOTE Some operational biometric systems perform a single biometric function. For example, in the initial stages of rollout of biometric passports, the operational system might be performing biometric enrolment only. Operational evaluation of such systems is within the scope of this part of ISO/IEC 19795.

This part of ISO/IEC 19795 does not:

— cover testing of operational systems in the laboratory or

— address vulnerability testing.

2 Conformance

An operational evaluation is in conformance with this part of ISO/IEC 19795 if it is planned, executed and reported in accordance with the requirements of Clause 6.

3 Normative references

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

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

4 Terms and definitions

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

4.1
acclimatization
change, over the course of one or more transactions, of a biometric characteristic that might impact the ability of a system to process a sample

NOTE Acclimatization is primarily associated with a subject's temporal adjustment to environmental effects, such as skin temperature.

4.2
attendant
agent of the biometric system operator who directly interacts with the biometric capture subject

4.3
biometric capture subject
individual who is the subject of a biometric capture process

4.4
biometric data subject
individual whose individualized biometric data is within the biometric system

4.5
biometric probe
biometric data input to an algorithm for comparison to a biometric reference(s)

4.6
biometric operational personnel
individuals, other than the biometric capture subjects, who take an active role in the operation of the biometric system

NOTE Biometric operational personnel includes biometric system administrators, attendants, and examiners.

4.7
biometric system administrator
person who executes policies and procedures in the administration of a biometric system

4.8
biometric system operator
organization responsible for defining policies and procedures in the operation of a biometric system

4.9
biometric reference
one or more stored biometric samples, biometric templates or biometric models attributed to a biometric data subject and used for comparison

4.10
comparison attempt limit
maximum allowed number or duration of attempts in a comparison transaction

4.11
enrolment attempt limit
maximum allowed number or duration of attempts in an enrolment transaction

4.12
habituation
familiarity a subject has with the biometric device, system and application

NOTE The level of habituation can affect biometric sample presentation and acquisition device.

4.13
subject base
set of individuals whose biometric data is intended to be enrolled or compared in operational use of a biometric system

4.14
system acceptance rate
proportion of recognition transactions in an operational system in which the subject is recognized

NOTE 1 Though the acceptance of an impostor is an incorrect recognition, it can still count as a system acceptance.

NOTE 2 System acceptance rate = 1 – system rejection rate.

4.15
system identification rate
proportion of identification transactions in an operational system in which one or more subjects are identified

4.16
system rejection rate
proportion of recognition transactions in an operational system in which the subject is not recognized

NOTE The system rejection rate differs from the false reject rate in that, in addition to false rejections, it also includes any rejected impostor transaction and any improper genuine transaction.

4.17
test crew member
selected biometric data subject whose use of the operational system is controlled or monitored as part of the evaluation

NOTE In an operational evaluation, test subjects can be subjects of the operational system or they can be members of a test crew using the system specifically for evaluation purposes

5 Operational evaluation overview

5.1 Operational evaluation goals

The overarching goals of operational testing are to measure or monitor operational biometric system performance.

Subgoals of operational testing may include:

— to determine if performance meets the requirements specified for a particular application or the claims asserted by the supplier;

— to determine the need to adjust or configure the system to improve performance;

— to predict performance as the numbers of subjects, locations, or devices increase;

— to obtain information on the target population and environmental parameters found to affect system performance;

— to obtain performance data from a pilot implementation;

— to obtain performance data to benchmark future systems.

Operational evaluation considers the performance of people as well as the equipment, algorithms, and environment. Consequently, operational testing includes aspects of social science in addition to physical science, whereas technology testing does not. In general, operational performance will vary over time due to uncontrolled conditions in people, equipment and environment. For example, if the majority of subjects are enrolled at the start of operations, with few new enrolments, performance of the system might improve as

subjects habituate or degrade as subjects' biometric characteristics age over time away from their enrolled references.

The performance observed in testing can depend on the operational personnel, such as attendants or biometric examiners, as well as the biometric subjects. Operational personnel based factors should be taken into consideration in all aspects of the test from scope definition to reporting (see e.g., references [1] and [2]).

5.2 Operational performance metrics

Recognition metrics for operational testing differ from those used in technology and scenario testing. In technology and scenario tests, false accept and false reject rates can be measured because the underlying ground truth is known to the experimenters. Ground truth will generally be unknown in an operational setting such that an operational test will measure system acceptance and system rejection rates.

Determining the false accept rate and false reject rate from the number of system rejections and acceptances will require additional observations or controls to determine the legitimacy of identity claims and device interactions. Similarly, in the case of identification systems, determining identification error rates from the number of system identifications also requires additional observations or controls.

Performance calculations in technology and scenario tests often exclude rejections in which the subject did not provide an ideal presentation or did not correctly follow instructions. Operational testing will include such rejections in measuring the system rejection rate.

5.3 Operational evaluation methods

Operational performance assessment is based on data collected through an operational system. A system may be configured in "evaluation mode" to collect additional data during normal system use.

Operational performance assessment may be based upon different groups of test subjects:

— data collected with a non-controlled set of test subjects (i.e., a set of test subjects reflective of the subject base of the operational system);

— data collected with a controlled set of test subjects defined as the "test crew members" (i.e., a set of test subjects considered and controlled separately from the subject base of the operational system).

If the test crew is specially instructed in their use of the system, evaluation results can be expected to differ from those encountered operationally. Operational performance can be highly time variant because of uncontrollable variations in the population, equipment and environment. Variation in performance across these conditions cannot be predicted.

5.4 Determining operational performance

Performance estimates may be determined from operational data in at least three ways:

— through direct observation of throughput rates, acceptance rates and/or rejection rates;

— through offline computation of throughput and acceptance and rejection rates based on comparison scores and timing metrics recorded during operations; and

— through offline computation of comparison scores, and acceptance and rejection rates based on samples acquired during the test and on stored reference data.

Each approach is likely to yield different measures. Further, each approach carries different risks for miscalculation. For example, recording only direct observations of an access control system does not reveal whether a rejection was due to a biometric error or an error in the operation of the gate; recording only comparison scores will not show cases where a reference or probe has been stored or transmitted incorrectly.

5.5 Use of technology and scenario evaluation methodologies in evaluating operational systems

In addition to tests based on real operational use of the system, scenario and technology evaluation (ISO/IEC 19795-2 [3]) can also have a role in determining some aspects of operational performance.

Testing solely in live operation might not be capable of measuring all aspects of operational performance. Depending on the purpose of the evaluation, certain performance measures might only be determined by testing for them specifically. Testing in live operation is not meant to guarantee that the system will be operating under the specific conditions, or with sufficient frequency, in order to draw statistically valid conclusions. Furthermore, when testing operationally, it might be infeasible to isolate these effects from other operational factors that also affect performance.

EXAMPLE Environmental factors such as sunlight or humidity can affect sensor performance. Unless the system is tested to monitor performance in the specific environments, the effects that such factors have on the operational performance cannot be quantified.

6 Operational evaluation

6.1 Purpose and scope

6.1.1 General

The purpose and scope of the test need to be determined before the test design can be drawn up. The following elements shall be addressed:

— criteria for system inclusion,

— system specification,

— enrolment and comparison functionality to be evaluated, and

— performance measures of interest.

6.1.2 Criteria for system inclusion

The experimenter shall address the criteria by which biometric system(s) are included in an operational test. Biometric systems may be included in an operational test due to their having been previously deployed, due to selection on the part of the biometric system operator, or due to selection on the part of the evaluating entity.

NOTE Testing multiple independent systems could compromise the operational realism of the tests. However, some elements of the test must be controlled if meaningful comparisons are to be made. Some elements can be controlled without jeopardizing the operational value of the tests. (See reference [4].)

6.1.3 System specification

Details of the system under test shall be specified as fully as possible. The following elements should be reported:

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

— for biometric algorithms: provider, version, and revisions, and the values of all field-variable parameters or settings — biometric algorithms include quality assessment, feature extraction, binning, comparison and fusion algorithms, and any or all of these might be supplied by different vendors;

— if the operational system incorporates a biometric application, such as a logical access interface: provider, title, version, and build of the application;

— for systems tested on or through personal computers, personal digital assistants or other computing devices: platform, operating systems, processing power, memory, manufacturer, and model of computing device;

— details of system architecture and data flow between biometric data acquisition, processing, and storage components;

— data flow between system components;

— system configuration (e.g., in the case reference updating, does the system use a single biometric reference for all subsequent comparison attempts or is the biometric reference updated following each successful attempt).

6.1.4 Biometric functionality

In operational tests of previously deployed systems, the biometric functionality (i.e., enrolment, verification or identification) under evaluation should be that of the deployed system. In operational tests of systems deployed for the purpose of operational testing, the experimenter may determine which comparison functionality(ies) to implement and evaluate. The rationale for selecting the comparison functionality components to be evaluated by the operational test shall be reported.

NOTE An operational test can incorporate both identification and verification functions if, for example, data is used to execute searches against watch lists and also for verification against an existing enrolment.

6.1.5 Performance measures

Performance measures relevant for operational evaluation include:

— throughput for enrolment and recognition transactions,

— failure-to-enrol rate,

— system rejection rate (in verification systems),

— system identification rate (in identification systems)

— false accept rate and false reject rate (in verification systems when the evaluation can establish ground truth),

— false-positive identification error rate and false-negative identification error rate (in identification systems when the evaluation can establish ground truth).

Considerations for determining and reporting these performance measures are specified in Clause 6.4. In addition, experimenters shall determine any specific performance measures to be generated through the operational test. Annex A provides a list of potential performance metrics.

6.2 Application characteristics

6.2.1 General

Application characteristics shall be considered in order to plan for test data collection that will be representative of operational use. The following elements shall be addressed:

— concept of operations,

— guidance and instruction,

— levels of effort and decision policies,

— use of multiple instances,

— environment,

— deployment factors,

— acclimatization, and

— habituation.

6.2.2 Concept of operations

The concept of operations of the operational application being tested shall be determined and reported. Description of the concept of operations shall include, but not be limited to, discussion of:

— integration of biometric systems with external systems such as logical or physical access systems;

— authentication methods and systems that the biometric system is replacing or complementing, if applicable; and

— the category of fielded application under evaluation, e.g.:

  — enrolment,

  — physical access control,

  — logical access control,

  — surveillance or screening,

  — identification, or

  — examiner assisted identification.

6.2.3 Guidance and instruction

While operational test subjects should not be accompanied by test support personnel in their interaction with the biometric system, as would be the case in a scenario evaluation, they may still be given instructions on how to utilize the technology. Any instructions on how to utilize the biometric system shall be reported. Information given to the test subject regarding the operational test shall also be reported.

The experimenter shall define the following elements related to guidance provided during enrolment and recognition for each system tested:

— attended/unattended operation;

— information provided by attendant to test subject prior to interaction;

— type and extent of feedback provided by biometric system to test subject during interaction;

— type and extent of feedback provided by attendant to test subject during interaction;

— use of scripts, instructions, guidance tools, or other mechanisms to inform test subjects as to the optimal method(s) of interacting with the system;

— correction and recording of improper interaction with device;

— constraints on test subject appearance and apparel; and

— differences in guidance and feedback from that provided for operational use.

NOTE For operational tests in which the system is not already in operational use prior to testing (i.e., a pilot deployment), the amount of attendant-test subject interaction should approximate that which would be provided by the biometric system vendor or administrative personnel for a typical deployment.

6.2.4 Levels of effort and decision policies

The level of effort involved in enrolment and recognition, as well as the test system's decision policies, are dictated by the operational environment. In most cases the system will have a specific configuration based on the concept of operations and will execute a match/no match/enrol or other decision in real time, such that the subject's subsequent activities in an operational environment are contingent on the outcome.

For tests in which enrolment is in scope, the experimenter shall determine and report enrolment attempt limits and decision policies for the system to the extent feasible. Such limits and policies may include the following:

— minimum and maximum number of attempts required or permitted to enrol (see Annex B). A system may allow enrolment after one attempt, or may require multiple attempts.

— minimum and maximum duration permitted or required to enrol within a given enrolment attempt or transaction. A system may terminate an enrolment transaction e.g., after a fixed duration. This may be due to (1) the inability to capture or acquire a biometric sample of sufficient quality or distinction, or (2) the inability to sense any biometric characteristic whatsoever. Incident (1) means that a biometric system has acquired and processed data, but found it lacking; incident (2) means that the data was not acquired and/or processed.

— whether the enrolment process includes an immediate attempt at verification using the enrolment to confirm satisfactory enrolment and to help subject familiarization, and

— quality criteria or threshold applied during enrolment.

For tests in which recognition is in scope, the experimenter shall determine and report comparison attempt limits and decision policies for system to the extent feasible. Such limits and policies may include the following:

— minimum and maximum number of attempts required or permitted per comparison transaction. A system may make a decision after one attempt, or may require multiple attempts.

— minimum and maximum duration permitted or required to match within a given comparison attempt or transaction. A system may terminate a comparison transaction after a fixed duration. For systems that do not time out, a time limit may be established for the purposes of testing.

— quality criteria and thresholds applied during comparison.

6.2.5 Multiple-instance systems

When it is known that a system utilizes multiple instances from the same test subject for the purposes of enrolment or recognition this shall be reported. In multi-instance, multi-presentation, or multi-attempt systems, fusion techniques may be used to combine information from two or more instances, presentations, and/or attempts. The fusion techniques used, and the method in which they are applied, should be reported when available.

NOTE In an operational system, an attempt in which the subject is rejected might be automatically followed by a fallback attempt in which the system utilizes a different instance of the same modality from the same subject. For example, a system might utilize the left index finger for comparison and prompt for placement of the right index finger if the left index finger fails. This would typically apply to modalities such as fingerprint and iris in which most biometric capture subjects can utilize more than one instance for enrolment and recognition.

6.2.6 Environment

Environmental conditions can play a major part in operational performance levels. All relevant environmental conditions shall be considered. Reference should be made to ISO/IEC TR 19795-3 [5] which lists the physical environmental conditions to be considered for different modalities and operational settings. These include, but are not limited to:

— ambient temperature,

— atmospheric pressure,

— relative humidity,

— exposure to elements (e.g., rain, fog, snow, hail, etc.),

— illumination (including type, direction, intensity), and

— ambient (acoustic) noise.

Environmental conditions that are controlled by the operational system shall be reported to the extent feasible. Any introduction of controls on environmental conditions not normally incorporated in the fielded application should be avoided. If introduced, such controls shall be documented. Conditions that are not controlled may be monitored and reported to the extent that these aspects are of interest.

EXAMPLE 1 If the operational system utilizes dedicated lighting for face recognition, details of the lighting system would be reported.

EXAMPLE 2 If operational performance data is not collected on days where the temperature exceeds 30°C, such occurrences would be documented in the test report. Similarly, if a portable air conditioning unit were used on such days, this would be reported.

For both monitoring and acceptance testing, it may be appropriate to record pertinent environmental data (e.g., ambient temperature, relative humidity, precipitation, cloudiness, illumination levels). The most comprehensive and informative data may be collected at the individual location of the biometric device transaction (e.g., at pedestrian gate N), and at the specific time of the test transaction. However, this approach might be complex, difficult and expensive. An alternative approach may collect the environmental data at a representative location (or set of locations) per test site, including a date-time stamp, and correlate the data off-line with the individual transaction events.

6.2.7 Deployment factors

Deployment factors shall also be considered as these affect the usability of the system. Such factors can include constraints for security, health and safety (e.g., having equipment and attendants on different sides of a glass screen, or fastening equipment so it cannot be moved).

The experimenter should report the physical layout of the operational environment, including but not limited to the following:

— dimensional area dedicated to test execution;

— positions of natural and artificial lighting;

— positioning of biometric acquisition devices.

6.2.8 Acclimatization

Acclimatization should mirror that of operational use. The manner and degree by which biometric characteristics of the test crew are acclimatized to the operational environment shall be reported, and justification shall be provided if this differs from acclimatization in operational use.

6.2.9 Habituation

The degree to which the subjects are habituated to device(s) under test prior to test execution shall be determined and reported, where practical. Habituation may be reported in terms of the period of time over which subjects have interacted with the specific device or class of devices; and the frequency of their use (e.g., several times daily, once daily, once weekly).

NOTE 1 Systems such as time and attendance and network login would in most cases have a habituated subject base. Systems such as voter ID, border entry, and entitlements can require much less regular interaction with a biometric system or device and therefore would be more likely to be categorized as non-habituated.

NOTE 2 Subject habituation to device(s) under test can have a substantial impact on error rates and throughput. Habituated subjects will tend to generate lower false reject rates and failure-to-acquire rates, and shorter throughput times, than non-habituated subjects.

Test crew habituation should be similar to that of the target population. A test crew can become habituated to a device prior to testing, such as in the course of employment, or can become habituated to a device in the course of testing.

Subject habituation is impacted by the frequency of subjects' transactions as well as by operational parameters of the system, such as threshold and maximum number of recognition attempts allowed per transaction. The operational values of such parameters should therefore not be altered for testing.

For tests in which the biometric device is newly installed, experimenters should allow sufficient time for test crew to become habituated before collecting data that is intended to predict the long-term performance of the device. This time may be stipulated either as a specific period (for example, 2 weeks), the time required to obtain a minimum average number of subject interactions, or by monitoring performance data to determine when the test crew "learning curve" flattens out indicating that sufficient habituation has been obtained.

NOTE 3 If one aim of the test is to measure the effect of habituation on performance, more detailed controls and reporting might be needed. See Annex F.

NOTE 4 If multiple devices (of the same or of different biometric modalities) are under test, test subjects might not be equally habituated to all device types. For example, a test subject might be habituated to placement-based fingerprint devices but not to swipe-based fingerprint devices. Further, certain biometric technologies are more heavily impacted by the degree of habituation than others.

6.3 Test Plan

6.3.1 General

Operational test planning is directed by the type of performance information an organization wishes to collect (addressed in Clause 6.1), and is constrained by characteristics of the application and the operational environment (addressed in Clause 6.2) that should not be altered for the purpose of testing.

The test plan shall specify:

— system implementation and configuration;

— test population;

— test transactions from which performance metrics will be derived.

In determining the target number of test transactions to be made, it is important to consider how accurate the results need to be and the level of confidence bounds to be used.

Annex C provides an outline test plan for an operational evaluation. Annex D provides guidance on statistical methods for determining population sizes etc. Annex E describes elements of a test plan to monitor and evaluate long-term and temporal trends in operational system performance. Annex F describes elements of a test plan for operational habituation testing.

6.3.2 System implementation and configuration

Collection of data for the purposes of operational testing should be done in a manner that influences as little as possible subject and public impressions of the system, and that has minimal impact on normal use and operation.

The system under test may be configured or instrumented to produce data supporting performance analysis (e.g., comparison scores, quality scores or sample images or features). The act of collecting such data will often impact performance of the system itself, particularly with respect to throughput. The experimenter shall (i) document any instrumentation and reconfiguration of the system; (ii) minimize the impact of these modifications on performance; and (iii) estimate and document the impact on performance attributable to the modifications.

EXAMPLE A system might be configured to log comparisons scores resulting from 1:1 comparisons, or to save recognition samples. If saving samples noticeably increases transaction times above those of the operational system, this could substantially affect subject interactions and hence error rates as well as throughput. If so, the saving of such data would be inadvisable.

The operational application might constrain (i) what data can be recorded, and (ii) the degree to which the operational system can be modified. For example, a physical access control system might be incapable of storing samples or providing visibility into comparison functions due to architecture constraints. Conversely, a border control application might mandate that samples and detailed transaction logs be retained. Moreover, operational test outputs are generally recorded on a transactional basis, which can limit one's ability to evaluate performance at multiple comparison score thresholds.

NOTE The issue of visibility into device operations can be important to operational testing. Depending on the output of the device, one might not know whether a rejected transaction was due to a biometric matching error, a time-out, a failure-to-acquire or invalidity of non-biometric data.

The test plan shall specify, for each component (or system) tested:

— types of outputs recorded by the biometric system, including but not limited to comparison scores, acceptance and rejection decisions, candidate lists, enrolment quality scores, sample quality scores;

— range of comparison scores and quality scores output by the biometric system as well as the actual threshold value(s) used for the test;

— method(s) through which outputs are provided by the biometric system (e.g., some results may be logged locally on the biometric device, other results saved centrally in the system).

6.3.3 Test population

6.3.3.1 Test subject selection

Operational testing typically utilizes employees, citizens, customers, or individuals otherwise associated with an organization as its test population. The experimenter shall report the relationship between test subjects and the biometric system operator. The experimenter shall report whether the subjects utilize the biometric system(s) being evaluated in the regular course of their interaction with the organization, or whether utilization takes place specifically for the purposes of the test. The test crew may also include a group of specially selected test subjects within a wider test population of regular system users.

The degree to which the test subjects are representative of the target population shall be documented. The demographics of the test population should be reported if available.

NOTE An operational test of employees comfortable with biometric technologies might not generate results applicable to a target population.

6.3.3.2 Test subject management

The test plan shall specify information related to test subject management, including at a minimum the following:

— types of identifiers used to identify test subjects;

— methods used to establish ground truth in recognition or identification tests;

— impact of methods establishing ground truth on test subject interaction with system(s);

— method of determining the order in which test crew members interact with systems (when testing multiple systems or components); and

— amount and type of demographic data to be collected.

Operational tests of biometric systems often require the use of personal data. Sometimes such data (e.g., subject data and identifiers) needs to be processed in a manner that retains the subjects' anonymity and complies with relevant privacy constraints and data protection procedures.

6.3.4 Test transactions

6.3.4.1 General

The test plan shall specify a target number of transactions to be executed the operational test. The level of confidence required in the test results will establish a minimum test size. This can be determined using statistical measures; see Annex D for details. The number of transactions shall be reported.

For measurement of system rejection rate or system identification rate, the frequency and number of transactions per test subject should be commensurate with the normal interaction of subjects with the biometric system in the operational environment. The frequency with which test crew members execute transactions shall be reported. Such frequencies may be reported as the median and average number of transactions executed per day per test subject.

NOTE The target number of transactions per test subject need not be uniform across all test subjects.

Measurement of error rates (i.e., false accept rate, false reject rate, false-positive identification rate, false-negative identification rate) requires a set of transactions where the ground truth regarding the identity of the capture subject and the legitimacy of their interaction is established. These transactions may be distinct from the set of transactions used to measure system rejection rates or system identification rates, or the sets may overlap.

In operational tests that measure false accept rate and false reject rate care should be taken that the ground-truthed genuine and impostor transactions are generated under broadly equivalent conditions (e.g., in terms of threshold, number of attempts allowed, habituation of test subject). Any discrepancies should be reported.

For systems previously in use, any modifications to system interaction introduced in the course of the test shall be reported.

During testing of operational use, there should be no other concurrent activities on the system under test, or in its environment, that could affect or invalidate the results (e.g. fire drills, repairs, or system maintenance). Should unplanned events occur in the course of evaluation which might affect or invalidate the results, such

occurrences shall be reported, and the decision about inclusion or exclusion of transaction data shall be documented.

6.3.4.2 Data recording processes

The test plan shall specify the following information related to data collection:

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

— the manner in which test subject interaction with the system is recorded;

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

— criteria for excluding transaction data from performance analysis.

EXAMPLE Test subject interaction with systems might be video recorded, logged by a system, self-reported, or directly observed. As less robust observation of system interaction takes place, less performance data is available for review. For example, measuring presentation errors or failures-to-acquire might not be possible without direct observation.

The experimenter shall include in the test report examples of data collection elements such as spreadsheets and logs, whether as screenshots or reproduced forms. Data collection should be automated to the degree possible without impacting performance results. The test report shall clearly indicate which metrics were generated through automated data collection and which were generated through manual data collection.

6.3.4.3 Genuine transactions

Operational evaluations should incorporate methods by which the identity of test subjects can be confirmed without recourse to the biometric system under test. Any methods used to confirm the identity of test subjects executing transactions recorded as “genuine” shall be reported.

Operational evaluations designed to measure the false reject rate, or identification error rates, shall incorporate genuine transactions in which the ground truth about the identity of test subjects, and the legitimacy of their interactions, is established without recourse to the biometric system under test.

NOTE In an operational environment, it might be impossible to determine whether transactions intended to be genuine are in fact being executed by the correct test subject. Further, it might be impossible to determine whether transactions are being executed in good faith.

6.3.4.4 Impostor transactions

Operational tests may incorporate impostor trials. Any methods used to confirm the identity of test subjects executing transactions recorded as “impostor” shall be reported. Any impostor activity that can be estimated or measured should be reported.

Operational evaluations designed to measure the false accept rate shall incorporate impostor transactions in which the ground truth about the identity of test subjects is established without recourse to the biometric system under test.

Generation of impostor trials is a difficult aspect of operational testing. The following are potential methods through which impostor transactions with known ground truth can be generated.

— **In-line with operational use of the system**: Test crew members may execute impostor transactions on the operational system. The impostor transaction may utilize an ID or token for the subject being impersonated against which to execute a comparison. This provides a direct reflection of the ability of impostors to match in a system. In-line testing might not be viable for reasons of security, throughput, or process flow.

— **Through dedicated operational systems**: Test crew members may execute impostor transactions at a dedicated time, or through a dedicated system, utilized solely for the purpose of impostor testing. Such testing would have no impact on primary path operations. A test crew member may execute several impostor transactions against one or more biometric references. Experimenters shall ensure that the process and character of interaction with dedicated systems is consistent with that of operational devices.

— **Through reuse of genuine transaction data, in real-time or offline**: The appropriateness of this hybrid approach will depend on the operational environment and protocols, as well as the ability to save samples, references, or features.

— **In a scenario environment**: Test crew members may execute impostor transactions in a non-operational environment configured with the same thresholds as operational systems. Executing such testing and synthesizing results with genuine test subject data as gathered in operational systems requires that experimenters shall ensure that the method of interaction with dedicated systems is consistent with that of operational devices.

NOTE 1 Offline testing of the false accept rate using samples saved in genuine transactions can be strongly biased. For example: (i) the saved sample might be the one most closely resembling the claimed reference, while an impostor claim would likely have resulted in the saving of a very different sample; (ii) some systems develop sample features from multiple sequential samples, but save only a single sample which might not be representative of all the samples presented; (iii) some systems only save samples from successful transactions. In such cases, reuse of genuine transaction data can grossly underestimate the likelihood of a successful impostor transaction.

NOTE 2 It might not be viable to maintain zero-effort impostor attempts as in an operational environment; a test crew member will typically know whether genuine or impostor transactions are being executed. If an experimenter wishes to maintain the zero-effort impostor methodology for verification systems, one method is to provide a token against which a comparison will take place whose ID is unknown to the test crew member. Another possibility is to have an attendant provide a claimed identity unknown to the subject through entry of an ID number or PIN.

NOTE 3 Vulnerability testing with “active impostor” transactions exploiting potential vulnerabilities is not within the scope of this part of ISO/IEC 19795.

6.4 Performance measurement

6.4.1 Throughput

The test design shall determine and define the metrics to be used for throughput.

— One approach may define throughput as the average number of transactions completed over an appropriate unit of time (minutes, hours) under conditions of a full queue. With this approach measured values can depend upon waiting time in the queue, as subjects observe other transactions and prepare to adjust their behaviours when finally encountering the equipment accordingly.

— An alternative approach may define throughput as the expected (average) time of a transaction for a single person when encountering the system without a queue. This approach shall require determination of the interaction boundaries for the transaction, which will necessarily contain some level of arbitrariness.

Different definitions would be expected to lead to different measured values of the throughput.

Methods by which transaction times are measured shall be reported. Even when timing data is collected through direct observation by biometric operational personnel or by test personnel (e.g., with stop watches), supporting timing data can be collected automatically by the system at several points in the process. It is quite reasonable to expect that automatic timing points will not coincide with the timing points required to determine throughput.

In the case of recognition, transaction durations for enrolled subjects will often differ from those for unenrolled subjects, and should therefore be measured and reported separately.

6.4.2 Enrolment analysis

The experimenter shall report if enrolment was within the scope of the tests.

NOTE Enrolment testing might not be within the scope of an operational evaluation, for example when enrolments were conducted prior to the evaluation using a system outside the scope of the evaluation.

For tests in which enrolment is within the scope, the experimenter shall report the following:

— throughput for enrolment transactions,

— number of test subjects and transactions used to derive throughput,

— failure-to-enrol rate,

— number of test subjects and transactions used to derive the failure-to-enrol rate, and

— the proportion of the test subjects utilizing pre-existing enrolments (for which enrolment transactions are not executed as part of the operational test).

If the test utilizes previously enrolled test subjects, or obtains enrolled references from another system, such that enrolment transactions are not conducted within the system under test, information on the test subjects' prior enrolment should be reported if available.

Suggestions (non-mandatory) for additional reporting of enrolment performance are listed in Annex A.1.

6.4.3 Recognition analysis

6.4.3.1 General

For each verification or identification system tested, the experimenter shall report:

— throughput for recognition transactions;

— numbers of test subjects and recognition transactions used to derive throughput;

For each verification system tested, the experimenter shall report:

— system rejection rate;

— numbers of test subjects and recognition transactions used to derive the system rejection rate;

For each identification system tested, the experimenter shall report:

— system identification rate;

— numbers of test subjects, identification transactions, and enrolled individuals used to derive the system identification rate.

Suggestions (non-mandatory) for additional reporting of recognition performance are listed in Annex A.2.

6.4.3.2 Recognition error rate analysis

For tests where the test plan provides a means to establish ground truth to enable calculation of comparison error rates the experimenter shall report the following:

— for verification systems: false reject rate and false accept rate;

— for identification systems: false positive identification rates and corresponding false negative identification rates;

— the number of test subjects, recognition transactions and number of enrolled individuals used to derive error rates;

— whether ground-truthed transactions were generated in-line with system operations, or in a dedicated environment or by alternative means such as offline comparisons;

— for genuine and identification transactions, distribution of time lapsed between acquisition of enrolment and comparison data;

— statistical significance of test results based on number of errors, error rates, test population, and number of transactions executed.

NOTE Generally it is not possible to report a detection error trade-off curve indicating performance at thresholds not used in the test as (i) system data collection can stop when the operational threshold is reached and (ii) subjects can be expected to tune their behaviour to the threshold in use.

Suggestions (non-mandatory) for additional reporting of recognition performance are listed in Annex A.2.

6.5 Reporting

6.5.1 Reporting planned test results

The test report shall include

— the purpose and scope of the evaluation including description of the system under test (see reporting requirements in Clause 6.1);

— description of the application characteristics (see reporting requirements in Clause 6.2);

— a copy of the test plan (see reporting requirements in Clause 6.3);

— details of any deviations from this test plan, including the amount of data excluded in calculation of performance metrics; and observations logged by test personnel and operational personnel justifying deviations from test plan;

— the performance values measured in the evaluation (see reporting requirements in Clause 6.4);

— estimation of statistical significance of these results.

The test report shall reflect if a requirement was out of scope or not applicable.

If a requirement was not addressed due to information being unavailable, the test report shall state that the applicable data is unknown. The test report shall explain why the data is unknown. In this fashion a test report reader understands what data was not recorded and why.

6.5.2 Reporting additional analyses

In the course of running an operational evaluation, additional analyses may be conducted beyond those of the original test plan. The results of such additional analyses should be reported separately from those the evaluation set out to measure. Reporting of additional analyses should consider potential biases in the results.

6.5.3 Reporting observations

In the course of running an operational evaluation, test personnel, operational personnel and test subjects may make or record observations, e.g. on possible causes for failure of individual transactions. If reported, such observations shall be reported separately from the results the evaluation set out to measure and, if possible, a statistical estimate of validity should be included.

6.5.4 Report structure

An example outline for the structure and content of the test report is provided in Annex G.

6.6 Record keeping

In addition to ISO/IEC 19795-1 record keeping requirements, records retained in operational tests shall include, but not be limited to:

— photographic images of the operational environment sufficient to clearly indicate the relative positioning of devices and subjects during testing;

— communications with suppliers pertaining to system configuration and operation;

— spreadsheets and matrices used for data entry.

If these records are not retained, the test report shall include a statement to that effect. The test report shall also indicate whether non-retention was for privacy reasons.

Annex A
(informative)

Non-mandatory performance metrics and reporting

A.1 Enrolment metrics

In addition to the mandatory reporting of failure-to-enrol rate and enrolment throughput the test plan may require reporting of the following results:

— proportion of test subjects unable to enrol due to lack of biometric characteristic (a subset of the failure-to-enrol rate determined above);

— for systems in which multiple presentations, attempts, transactions or sequences, are permitted or required to enrol, the proportion of individuals able to enrol at each effort level from lowest effort level to highest effort level;

— proportion of subjects unable to attempt enrolment (e.g. because they cannot physically reach the device);

— the median, mean and standard deviation for time to enrol for those test subjects able to enrol;

— the median time to enrol of the subset of test subjects to whom attendants provided guidance during enrolment;

— distribution of enrolment quality scores;

— enrolment results as available by for pre-determined subsets of transactions, e.g. by enrolment locations, demographic grouping of subjects, etc.

A.2 Recognition metrics

In addition to the mandatory reporting of system rejection and throughput etc, the test plan my also require reporting of the following results:

— failure-to-acquire rate — the number of presentations used to arrive at this rate, and the point at which a failure-to-acquire was declared, should be reported (if failures-to-acquire are known);

— distribution of comparison scores (if comparison scores are available) for recognition transactions and for genuine attempts and impostor attempts (if ground truth is known);

— detection error trade-off curve plotting the relationship between false accept rate and false reject rate (or false-positive and false-negative identification-error rates) over different decision criteria;

— median, or mean and standard deviation for duration of a recognition transaction — these times should be reported separately for enrolled subjects and unenrolled subjects;

— number of recognition transactions executed per day.

The test plan may also require that recognition results (i.e., system rejection rates, system identification rates, throughput rates, numbers of successful and unsuccessful recognition transactions) are reported with a breakdown by predetermined groups of transactions, e.g.:

— by test subject;

— by age, gender or other demographic grouping;

— by location (where the system is in operation at multiple sites);

— by time period (such as hour of the day or month of the year);

— by operational personnel.

Annex B
(informative)

Sub-transaction events in operational testing

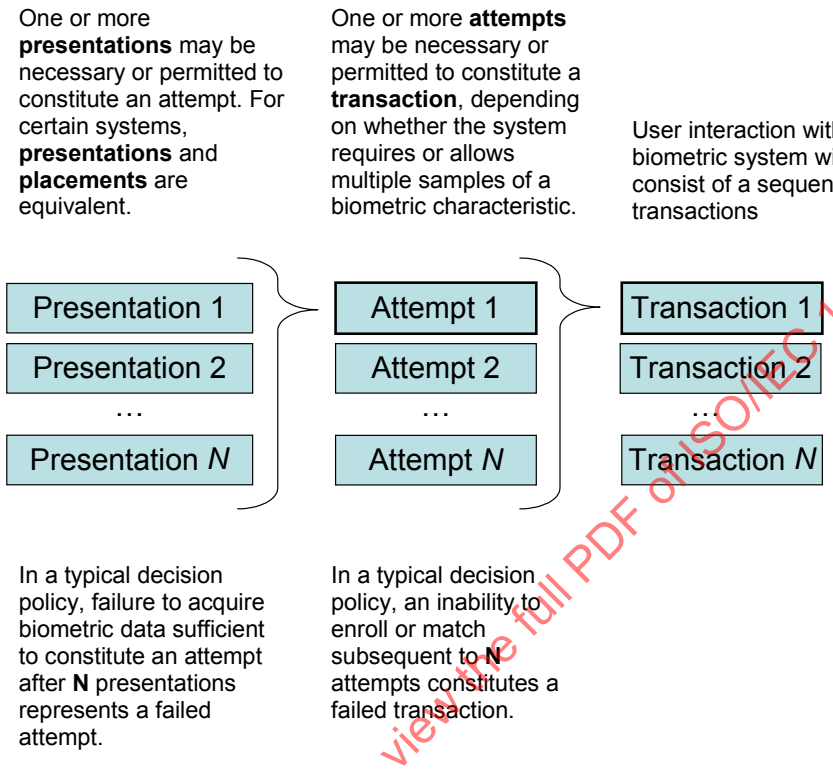

Figure B.1 — Relationship between presentations, attempts, and transactions within operational tests

Annex C
(informative)

Sample operational test specification

Table C.1 — Sample outline of an operational test specification

1	**Background & Scope**		
1.1	Objectives	Describes the evaluation's goals, target audience, and the manner in which it is differentiated from other tests in its field.	6.1.2 6.1.4
1.2	Performance metrics	Describes the evaluation criteria, the manner in which comparative or absolute performance statements are generated and the anticipated reliability of results.	6.1.5
2	**Systems evaluated**		
2.1	Acquisition devices	Describes acquisition device(s) evaluated, to include identification, build, and version as applicable, as well as criteria for participation and/or selection.	6.1.3
2.2	Application software	Describes enabling software through which acquisition devices were implemented, to include identification, build, and version as applicable.	
2.3	Enrolment and comparison algorithms	Describes the biometric enrolment and comparison algorithm(s) processing engines, to include identification, build, and version as applicable.	
2.4	Software developer's kits (SDKs)	Describes toolkit(s) and development kit(s) used to implement the device and application software, to include identification, build, and version as applicable, as well as the manner in which the toolkits were utilized to execute testing.	
3	**Application characteristics**		
3.1	Operational narrative	Describes the evaluation's concept of operations, including the operational setting under evaluation, modalities(s) tested, and scale.	6.2.2 6.2.5
3.2	System policies	Described policies on enrolment and recognition transactions, guidance and training for subjects, acclimatization etc. that will be followed in the operational evaluation.	6.2.3 6.2.4 6.2.7
3.3	Environment	Describes test environment, and any differences from the operational application.	6.2.6
3.4	Habituation	Describes habituation levels for the evaluation, and any difference from the operational application.	6.2.8
4	**Test plan**		
4.1	Test system implementation	Describes configuration of the system to enable data collection during the evaluation, types of outputs etc.	6.3.2

4.2	Test population	Describes the number of test subjects, selection method, and demographic controls (if any), and management of test subjects such as use of identifiers.	6.3.3.1 6.3.3.2
4.3	Test transactions	Describes: – the number and frequency of test transactions – methods of recording outcomes of test transactions – methods for establishing ground truth for genuine and impostor transactions	6.3.4.1 6.3.4.2 6.3.4.3 6.3.4.4
4.4	Criteria for exclusion of transactions from analysis	Describes processes for validating performance data and criteria under which data collected in a test transaction would be excluded from analysis	6.3.4.2

Annex D
(informative)

Methods to determine test size

D.1 General

Performance evaluations should plan to use a test size (i.e., number of test transactions) that will be sufficient for the purpose of the test. The size of the test influences the accuracy of the performance results obtained, how low an error rate can be measured; the likelihood of finding a significant result when one exists (the *power* of a test), and the likelihood of misinterpreting random effects as being significant (*significance level* of the test).

The relationship between number of test transactions, and the statistical significance of results depends upon hypotheses to be tested by the evaluation. In biometric system evaluation the objective is often to compare proportions such as error rates, rejection rates, or failure-to-enrol rates, e.g.:

— Is there a difference in rejection rates between two subsystems using different sensors?

— Does changing the environmental conditions improve the rejection rates?

— Is the system performing within the specified performance bound for false rejection rate?

The examples below show the general approach for determining test size in such cases.

D.2 Test size to compare proportions

The evaluation investigates whether the system rejection rate differs for two different sensors (A and B). The experiment will sample n genuine verification transactions using each sensor. Let p_A and p_B denote the true probability of rejection with each sensor. The test statistic $\overline{D}$ will be the observed difference in rejection rates between the two samples, and the null hypothesis ($p_A = p_B$) will be rejected if $|\overline{D}|$ exceeds the test criterion c.

In order to calculate a desired test size the experimenter should decide:

— a value δ which represents the size of difference that is considered substantial;

— the desired probability P' of obtaining a significant result if the true difference is δ or greater (common values for P' are 0.80, 0.90 or 0.95);

— the significance level α of the test (common values for α are 0.10, 0.05, or 0.01).

For the purposes of this example, assume that the verification transactions are statistically independent. Then $\overline{D}$ will be approximately normally distributed, with mean $p_A - p_B$, and standard deviation of $\approx \sqrt{[2p(1-p)/n]}$ (where $p = (p_A + p_B)/2$ might be estimated from observation of the operational system).

The example will use values δ = 0.06, P' = 0.80, α = 0.10 and p = 0.12. That is:

a) if $|p_A - p_B| \geq 0.06$ then test should have *at least* an 80% chance of finding $|\overline{D}| > c$ and rejecting the null hypothesis;and

b) if $p_A = p_B$ then the test should have *at most* a 10% chance of finding $|\overline{D}| > c$ and rejecting the null hypothesis.

Note that

— $\overline{D}$ is approximately normal, with mean $(p_A - p_B)$, standard deviation $\sqrt{[2(0.12)(0.88)/n]} = \sqrt{[0.21/n]}$,

— 10% of the standard normal distribution lies outside the interval (–1.645, 1.645),

— 80% of the standard normal distribution lies within the interval (–0.842, ∞),

Requirements a) and b) give:

$$0.06 - 0.842\sqrt{(0.21/n)} > c > 1.645\sqrt{(0.21/n)}$$

For there to be a feasible test criterion c we must have:

$$n > (1.645 + 0.842)^2(0.21)/(0.06)^2 = 363.$$

So, for our example, we would need an evaluation with at least 363 test transactions per sensor.

The general formula for the number of test transaction for comparison of proportions (for a two-tailed test, and with independent transactions) is

$$n > (Z_\alpha + Z_\beta)^2[2p(1-p)/\delta^2]$$

where Z_α is the normal deviate for two-tailed significance level α, Z_β is the normal deviate corresponding to the one-tailed significance level $\beta = (1 - P')$, and p is (an estimate of) the proportion of interest averaged over both samples. Note that smaller values for δ, higher values for P', or smaller significance levels α will increase the number of test transactions required.

The cited method depends on the normality and independence assumptions. The method will tend to underestimate the number of transactions required if transaction results are positively correlated (e.g., if a high percentage of transactions are conducted with relatively few test subjects). Caution is also needed if, using the values estimated for p and n, $(n \times p) < 10$ (or if $n \times (1-p) < 10$), when the normal approximation may be insufficient for a good estimate of n.

D.3 Comparison of rejection rate against specification

Assume that the required performance is to achieve a rejection rate below p_0. The test statistic X will be the observed number of rejections, and the system will be accepted as performing within specification if X is no greater than a test criterion c.

The significance level α means that if the true rejection rate exceeds p_0 the chance of incorrectly accepting the system as meeting performance is at most α.

Again values P' and δ are chosen where it is desired that if the rejection rate is below $p_0 - \delta$, then there is at least a chance P' that the system will be accepted as meeting the requirement. Then, assuming that transactions are independent, the number of transactions n must satisfy:

$$\sum_{i=0}^{c}\binom{n}{i}{p_0}^i(1-p_0)^{n-i} \le \alpha \quad \text{and} \quad \sum_{i=0}^{c}\binom{n}{i}(p_0-\delta)^i\left(1-(p_0-\delta)\right)^{n-i} \ge P'$$

These can be solved to determine n and c, either by using the normal approximation to the binomial distribution (similar to example D.2), or for small values p_0, using tools such as Excel.

NOTE Using Excel, c=CRITBINOM(n, p_0, α) -1 gives the test criterion for accepting the system as meeting the specification. The number of transactions n must be large enough for BINOMDIST(c, n, p_0-δ, TRUE) to exceed P'.