TECHNICAL SPECIFICATION

ISO/IEC TS 19795-9

First edition
2019-12

Information technology — Biometric performance testing and reporting —

Part 9: Testing on mobile devices

Reference number
ISO/IEC TS 19795-9:2019(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

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

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

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

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

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

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

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

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

Introduction

The development of a mass-market in connected mobile devices, e.g. smartphones and tablets, has allowed users the convenience of accessing remotely a variety of services which previously needed face-to-face interactions or to have physical access to the service provider's infrastructure.

For some services, convenience should nevertheless remain secondary to the security needs. These services include for example remote payment on commercial websites, banking transactions or certified signing of official documents. To allow these trustful interactions, the need of reliable user authentication is of paramount importance.

One way to certify the user's identity is to implement biometric authentication ability in the device.

It is then important to properly evaluate the accuracy of biometric authentication to ensure that security is strong enough to allow mobile sensible transactions.

Several biometric modalities are widely utilized in consumer-focused mobile devices. Evaluation of biometric performance for all of these modalities should be consistent and follow the same guidelines, methodologies and requirements. Nevertheless, some modality specific considerations should also be addressed when conducting an evaluation. This document provides a general framework usable for all modalities as well as dedicated recommendations when needed.

ISO/IEC 19795-1 describes three types of biometric performance evaluations: technology, scenario and operational evaluations. ISO/IEC TR 30125[1] recommends scenario evaluation as the most proper type of evaluation for testing biometric performance on mobile devices.

A scenario evaluation is an "end-to-end" biometrics evaluation in which the full system is tested with a careful control of the surrounding conditions. However, when applying this type of evaluation to biometric systems working on mobile devices, testing and reporting methods should consider the particularities and constraints of these use cases.

Information technology — Biometric performance testing and reporting —

Part 9: Testing on mobile devices

1 Scope

This document provides guidance for performance testing of biometrics when this technology is used on mobile devices with local biometric authentication to improve authentication assurance.

This document aims to:

— Provide guidance for affordable and cost-efficient testing and reporting methods for performance assessment at a full system level of biometric systems embedded in mobile devices with offline evaluation of false accept rate (FAR) claims.

— Define modality-specific considerations of these methods.

This document is applicable to:

— verification use cases related to secure transactions.

This document is not applicable to:

— privacy aspects;

— secure authentication from mobile device to server;

— testing and reporting for presentation attack detection (PAD) mechanisms in mobile devices;

— performance testing of biometric sub-systems such as acquisition sub-system or comparison sub-system;

— continuous authentication.

2 Normative references

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

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

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*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 2382-37, ISO/IEC 19795-1 and the following apply.

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

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

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

3.1
mobile device
small, compact, handheld, lightweight computing device, typically having a display screen with digitizer input and/or a miniature keyboard

EXAMPLE Laptops, tablet PCs, wearable ICT devices, smartphones, USB gadgets.

3.2
authentication assurance
confidence in the authentication process

3.3
gender
classification as male, female or some other category based on social, cultural or behavioural factors

Note 1 to entry: This is determined through self-declaration or self-presentation and may change over time.

Note 2 to entry: Depending on jurisdiction recognition, this may or may not require assessment by a third party.

3.4
time limit
longest time before a biometric system returns a decision on accept or reject, success or failure to enrol, success or failure to acquire a biometric sample

Note 1 to entry: All decisions returned after time limit are discarded.

Note 2 to entry: Time limit set to 0 means no time limit. All metrics evaluated are reported with the longest time needed by the biometric system to return decisions.

4 General considerations for biometrics on mobile devices

4.1 Biometric authentication process

Currently there are two types of local authentications that can be executed by mobile devices:

— explicit authentication in which the user is aware that a biometric authentication is going to happen and presents voluntarily his/her biometric characteristic to the capture device;

— passive authentication in which the user is authenticated without active effort by the user.

NOTE This document does not cover biometric systems where the user is continuously authenticated by the system in the background.

Therefore, the definition of testing methods and protocols should consider both types of authentications. In particular, the test plan should consider the following aspects:

— what constitutes the biometric capture process;

— how the user should interact with the mobile device during this process;

— which are the policies to manage failure to acquire (FTA) failures.

4.2 Biometric capture sensor

Mobile devices may have two kinds of biometric capture sensors:

— Embedded sensors which are the generic sensors of a mobile device (e.g. front and back cameras, microphone, touchscreen) but that are used for collecting the biometric characteristic during a biometric authentication.

  For some modalities, biometric acquisition may rely on these embedded sensors, which have not been designed and optimized for this task. Image resolution or uncontrolled image post-processing may for example impede the accuracy of biometric algorithms designed for more controlled acquisition.

— Dedicated sensors which are specific sensors for collecting biometric characteristics (i.e. a fingerprint reader).

Even when dedicated sensors are embedded in the device, they need to be coupled with optimized software in order to take into account hardware and ergonomics constraints specific to a mobile use case. These constraints may include, e.g. lower usable area available for fingerprint, closer range acquisition for face recognition.

A full-system evaluation should be carried out including the biometric capture sensor. Where the same or equivalent capturing sensor and the same comparison algorithm is used, evaluation results for one device may be applicable to others.

For this reason, guidelines for the evaluation should include recommendations to decide the most appropriate sensor to use during the evaluation as well as how to expand test reports to include the characteristics of the evaluated sensor as well as its situation on the device. These guidelines should be defined per biometric modality.

4.3 Uncontrolled environment

A full-system evaluation should control the conditions in which testing is going to be carried out. As described in ISO/IEC TR 30125, one of the major issues of mobile devices is the uncontrolled nature of the capture environment and the variability over time. To obtain realistic results, biometric performance should be analysed in numerous conditions. However, it is unfeasible to do it due to time and budget constraints. Testing conditions should be reduced to measure performance in a limited set of conditions.

Therefore, recommendations for selecting the most proper testing conditions and how to report it should be defined for ambient conditions present during the evaluation. Specific recommendations should be defined per biometric modality.

4.4 Challenges in storing references and generating comparison scores

In most current implementations of biometric authentication on mobile devices, the generation and the storage of the biometric references and samples are protected from external access. Devices are not designed to store multiple references and generate comparison scores from the submission of a probe against these references. Some solution is necessary to the problem of comparing one probe against multiple references on mobile devices, where typically only one reference is available, and scores are not accessible. Solutions could include development of test harnesses, of prototype devices, or of alternative operating modes.

A process for validating results from these alternative operating modes against standard operating modes will be required.

4.5 Adaptation of the biometric references

Finally, most biometric solutions implemented in mobile devices are able to adapt the biometric references over time with the aim to reduce the false rejection rates. To assess the improvement over time, the evaluation methodology should emulate this process.

5 Overview of full-system evaluation of mobile devices

5.1 General description

An evaluation of a biometric system shall conform to the requirements and best practices described in ISO/IEC 19795-1. This document considers mobile devices as a "full biometric system": an end-to-end biometric system, covering all the steps from biometric sample acquisition and biometric characteristics extraction to biometric comparison with a biometric reference. A "full biometric system" evaluation encompasses the testing of all process and subsystems in a realistic scenario. As such, an evaluation shall additionally conform with the requirements of ISO/IEC 19795-2:2007, Clauses 1 to 5, 7 and 8, which are relevant for a scenario evaluation of a verification system.

NOTE Some mobile devices allow the enrolment of several users for biometric identification in a small dataset related to low security features, e.g. device unlocking. This document only considers verification use cases related to secure transactions, which can vary depending on the risk, policy and/or legislation that applies to the transaction.

Mobile devices are short lifecycle products, which may have various versions, regular software updates or hardware specifications changes from one market to another or from one production series to another. Several devices from various manufacturers may integrate the same biometric sensors and software provided by a unique biometric system provider. The evaluator shall determine exactly what is the Target of Evaluation (ToE) and what the evaluation covers.

The evaluation of the ToE may be in-house testing performed by the system manufacturer, or a third-party evaluation. The third-party may, for example, be a certification body, whose main objective would be to assess if the ToE meets or exceeds performance requirements relevant for the certification scheme.

EXAMPLE A certification scheme can require that a mobile device has a false reject rate (FRR) below 1 % and a FAR below 0,01 %. The system provider will claim that its product meets the requirements. The evaluator will test the mobile device and determine if observed errors rates support the claim.

For most of the currently commercialized mobile devices with biometric capabilities, the biometric application is a black-box for security and privacy reasons. Biometric samples are stored in a secure environment and all computations occur in a secure execution environment, with no access to biometric data or any intermediate results. A third-party evaluation requires that the system provider delivers a customized version providing access to biometric data or to detailed transaction logs.

5.2 Considerations for time efficient evaluation

5.2.1 Factors that increase the time and cost of biometric performance evaluations

There are different aspects that increase the time and cost of a performance evaluation:

— The minimum error rates to be able to determine with a statistically significant level of confidence. Depending on the expected error rates and the statistical significance to achieve, the number of transactions to perform may increase considerably. This fact entails an increase in either the number of subjects that participate in the evaluation, the number of visits or the number of transactions to be conducted by test subjects.

— The inability to store large amount of biometric data and to get access to the captured biometric samples and/or stored biometric references. As mentioned in the Introduction, mobile devices allow a few biometric references to be saved, and most of the time it is not possible to get access to them

for an external application due to security reasons. Both circumstances increase the time and the cost of the evaluation because testing procedures must be carried out online.

— The number of conditions to evaluate. Mobile devices are used in a diversity of situations (i.e. sitting at a table, standing, walking, or driving) in which the ambient conditions are changing constantly so there are innumerable scenarios for which biometric performance testing should be analysed.

5.2.2 Reduction of the number of recognition transactions

5.2.2.1 Approaches

The main challenge of a biometric evaluation of a mobile device for the evaluator is to assess that the observed errors rates, FAR and FRR, support a claimed performance with a sufficient statistical certainty.

ISO/IEC 19795-1 recommends the use of Rule of 3 or Rule of 30 approaches to assess performance claims.

— Rule of 3 defines the minimum number of transactions required for the estimation of a minimum error rate at a 95 % confidence level when no errors are obtained during the evaluation.

— Rule of 30 states that to be 90 % confident that the true error rate is within ±30 % of the observed error rate, there should be at least 30 errors. Based on the predefined error rates and the number of errors, i.e. 30 errors, it is possible to define the minimum number of transactions.

The target requirement for an error will directly influence the number of independent tests required to have statistical significance, and thus the size of the test crew, the time spent and the cost of the evaluation. To reduce the duration and cost of executing a biometric performance evaluation, the number of recognition transactions can be constrained.

EXAMPLE Evaluating a 0,1 % FAR rate by Rule of 30 requires 30000 independent tests (more precisely, an observed error rate of 0,1 % would mean the true error rate is between 0,07 % and 0,13 % with 90 % confidence). For the same 0,1 % FAR rate, Rule of 3 would only require 3000 independent tests, but the claim is only verified (with 95 % confidence level) if no error is observed, i.e. a test is not conclusive if only one error occurs.

For FAR evaluation, while the independence criteria would require that one test subject is only involved in one impostor transaction, it is commonly agreed that the statistical loss of computing all possible cross-comparisons between test subjects is acceptable. This approximation shall be considered relevant for a mobile device evaluation (i.e. with N test subjects, N*(N-1)/2 impostors tests can be made). Table 1 gives examples of the number of test subjects required for various FAR targets. In general, all possible cross-comparisons can only be executed offline.

Table 1 — Number of tests required in an evaluation for various FAR targets following Rule of 3 or Rule of 30

FAR target	Rule of 3		Rule of 30	
	Minimal number of tests required	Minimal number of test subjects	Minimal number of tests required	Minimal number of test subjects
1 %	300	25	3000	78
0,1 %	3000	78	30000	246
0,01 %	30000	246	300000	776
0,001 %	300000	776	3000000	2450
0,0001 %	3000000	2450	30000000	7747

As noted in 5.1, a third-party evaluation of FAR would quickly be impractical and time consuming if the evaluator only has access to an unmodified mobile device. The main drawback would be the impossibility to enrol more than one person on the mobile device at a time, meaning that each verification transaction must be done separately and that the enrolled person should be changed regularly.

A mobile device customized for evaluation could address these issues by off-line testing:

— the reference and probe biometric templates acquired on the device can be exported from the customized mobile device and compared off-line using a dedicated SDK.

— alternatively, the mobile-device manufacturer could provide a dataset and a dedicated SDK for off-line testing. The evaluator should then assert the relevance of the provided data.

As FRR targets are commonly higher than FAR targets. FRR evaluation is more practicable from a third-party point of view even without customized hardware or access to off-line testing.

Table 2 — Number of tests required in an evaluation for various FRR targets following Rule of 3

FRR target	1 %	2 %	3 %	5 %
Number of genuine tests required for Rule of 3	300	150	100	60

Table 2 gives a few examples of the number of tests required for Rule of 3 approach: as an example, for 1 % FRR target, a test crew of 300 people is sufficient. Nevertheless, as noted earlier, Rule of 3 implies that no error occurred, therefore an evaluation has a chance of being inconclusive. Using Rule of 30 could avoid the risk of an inconclusive result, but the number of tests required would increase tenfold.

One way to limit the increase of size of the test crew and to achieve affordable and time efficient evaluations is to allow multiple transactions to be executed by the same test subject. However, this will decrease test independence and increase statistic uncertainty. To help mitigate the decrease of independence, one test subject should never proceed to more than five mated transactions.

Table 3 gives examples of the number of test subjects required when one test subject is involved in more than one transaction for an FRR target of 1 %.

Table 3 — Number of test subjects required to follow Rule of 30 for 1 % FRR target, when one test subject can be involved in several genuine comparisons

Number of probe transactions per subject	1	3	5
Number of test subjects required	3000	1000	600

Similarly to FAR testing, a customized mobile device, allowing, for example, off-line testing, could make an FRR evaluation involving multi-transactions more practical.

To increase the independence of transactions, acquisitions should be performed in separated sessions with a time gap of at least one day between each one:

— For mated comparison trials, transactions should be performed in separate sessions. If genuine transactions are performed in the same session the measured FRR will be significantly underestimated. If it is not feasible to perform verification transactions on separate days:

  — the test subject shall at least disengage from the device and perform another activity between transactions,

  — the test report shall acknowledge that it is the same day FRR which was measured and is underestimating the real FRR of the system.

— For non-mated comparison trials, testing does not need to follow this recommendation.

NOTE 1 While the impact varies depending on modality, same day comparison for genuine transaction will in particular significantly neglect the impact of ageing on recognition accuracy. Most mobile devices do have reference adaptation mechanisms to update enrolment regularly and mitigate ageing effects, however the loss of validity of results for same day comparison is especially relevant for face modality.

NOTE 2 Carrying out separate sessions increases the duration and the cost of the evaluation and this mainly affects the FRR rate as it is addressed in 5.2.3.

In addition to Rule of 3 and Rule of 30, bootstrapping (sampling with replacement) is also a method to determine confidence particularly when a distribution cannot be assumed. Bootstrapping allows the use of the test data to determine the confidence depending on the test scenario, e.g. number of subjects and number of attempts (or transactions) per subject.

When attempting to establish whether an error rate is below a threshold, it is appropriate to choose a one-sided upper bound. In this method, the statistical analysis does not focus on the mean, but rather whether it is below a specific performance target. For example, see Figure 1 (below). This compares Rule of 30 with a one-sided upper bound confidence interval using bootstrapping.

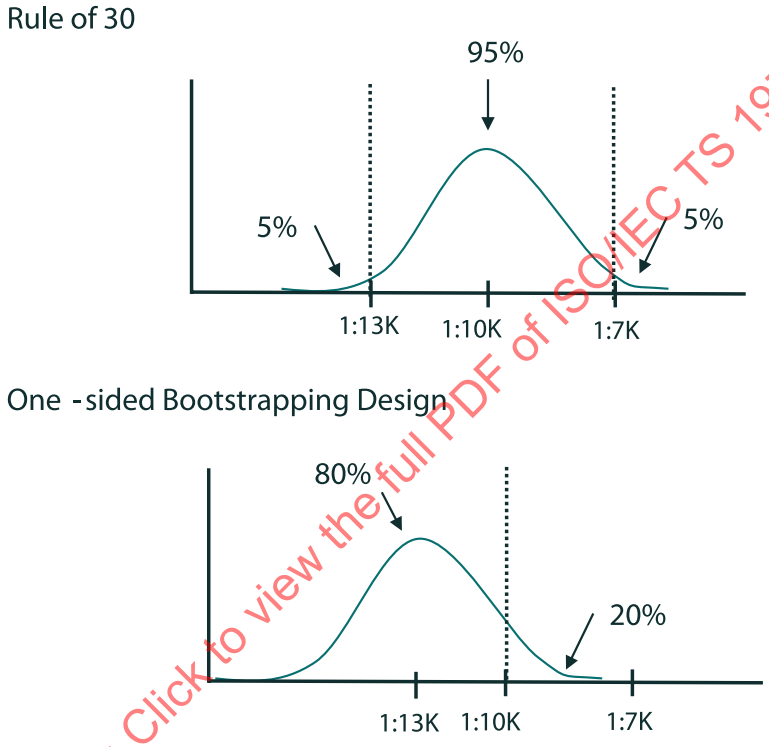

Figure 1 — Comparison of Rule of 30 and one-sided bootstrap design

Additionally, depending on the scenario, the number of subjects may be less than required for Rule of 30.

EXAMPLE Simulations of the bootstrapping process were performed using the following settings: 245 Subjects (n), 1 enrolment per subject, 5 verification transactions(m), which resulted in 298900 total impostor comparisons from N = nm(n–1). Errors were randomly distributed across the 298900 comparisons, 5000 bootstraps created using the method in ISO/IEC 19795-1. When the Upper Bound (UB) of the Confidence Interval of the bootstrap distribution is set to 0,01 %, the mean FAR is necessarily below 0,01 %. Table 4 below provides the number of errors associated with 68 %, 80 %, and 95 UB confidence intervals. For example, in this simulation, the number of errors is 17 out of 298,900 in order to achieve a 95 % upper bound.

Table 4 — Example of Upper Bound (UB) confidence intervals and required number of errors

UB of confidence interval Set to 1:10,000	Number of errors to achieve UB
68 %	27 (out of 298,900)
80 %	23 (out of 298,900)
95 %	17 (out of 298,900)

5.2.2.2 Impact on the test methods and results

The recommendations of 5.2.2.1 directly affect two key aspects of the evaluation.

— The method to obtain the false accept rate. The calculation of the FAR should be done offline applying the cross-comparison strategy to achieve the maximum number of transactions. This means that for the evaluation:

  — a prototype which includes the biometric capture sensor and which operates as the biometric system under test should be created. This prototype should be able to save biometric references and samples.

  — a test harness which works as the biometric system under test and which will be able to compute offline error rates from the stored information should be developed.

  — equivalence tests between online and offline results.

  — it is necessary to define what is a conformant transaction and which data should be stored to replicate a transaction offline.

  — the minimum number of test subjects that should compose the test crew should be defined, especially for those modalities for which is possible to obtain different biometric characteristics from the same test subject.

— The method to obtain the uncertainty of the metrics. The Rule of 3 assumes that no errors will happen. But this is not always true. This rule proceeds from the assumption that transactions are independent and identically distributed. However, biometrics errors are not identically distributed across the population as well as transactions are not independent as several transactions will come from the same test subject. As a consequence, errors may happen and it will be necessary to calculate the uncertainty of the metrics.

To achieve this, ISO/IEC 19795-1 proposes different approaches: parametric and non-parametric. Due to the assumptions that are made when applying the parametric approaches, it is recommended to use a non-parametric approach, such as the bootstrapping method for the calculation of the confidence intervals.

NOTE ISO/IEC 19795-1 describes the bootstrapping method as an estimation method that reduces the need to make assumptions about the underlying distribution of the observed error rates and the dependencies between attempts.

If the bootstrapping approach is used, at least 1 000 bootstrapping samples shall be used. 95 % upper bound should be used for confidence interval.

5.2.3 Reduction of the number of conditions to evaluate

5.2.3.1 Approaches

Another aspect to achieve efficient and affordable evaluations in the case of mobile devices is to limit the test to calculate biometric performance in specific situations. Otherwise, addressing all possible combinations of testing conditions, such as the ambient conditions and the position of the mobile device relatively to the user, would be too complex to test in a cost-efficient evaluation.

Therefore, it is recommended to restrict the test to a scenario that will be easy to generate and control during the evaluation among test subjects.

NOTE While a biometric system is traditionally deployed in a closed environment where the system manager can enforce a policy able to optimize the system behaviour, biometric verification with a mobile device implies that no such policy can be applicable. Depending on the biometric modality used, an unconstrained acquisition environment can have a limited or severe impact on the overall system performance.

The conditions suggested for this scenario are indoor conditions, with no noise in which the mobile device is hand-held by the user.

5.2.3.2 Impact on the results

As performance testing is only calculated for one or a limited number of conditions, the results may differ from the reality.

These results should be considered as a baseline. For analyzing the influence of ambient conditions or the test subject interaction on biometric performance, other standards, e.g. ISO/IEC 29197[2] and ISO/IEC 21472[1)][3], should be applied.

5.2.4 Reduction of the number of visits

5.2.4.1 Approaches

Another aspect that may contribute to reducing the cost and duration of the evaluation is the limitation of the number of visits to just one visit. ISO/IEC 19795 addresses to conduct multiple visits for including effects such as ageing and/or the acquisition conditions on the measurements of performance rates.

5.2.4.2 Impact on the results

However, the impact of this assumption mainly affects the false rejection rate which will be optimistic. Nevertheless, later evaluations could be carried out increasing the number of visits to analyse the aforementioned effects.

6 Guidance for testing and reporting

6.1 Data collection

6.1.1 General procedures

The test crew involved in the evaluation shall be familiar with the type of mobile devices used (i.e. smartphone or tablet). The hypothesis is that the user is willing to use biometrics for secure transaction using a device they already possess and use daily. The evaluator should determine and shall report if some training is authorised with the biometric capabilities of the mobile device prior to the evaluation.

NOTE If the evaluation plans to use Rule of 3 to assess a claimed performance, as discussed in 5.2.2, the risk of having an inconclusive evaluation caused by a single error can be reduced by allowing a significant training phase.

The test crew shall be informed that they shall make their best effort when using the system and shall follow any guidance provided by the biometric system as closely as possible (i.e. the test crew shall never purposefully try to test the limits of the system).

6.1.2 Test crew size and characteristics

Test crew size will be influenced by considerations discussed in 5.2.2 about the number of tests required to have statistical significance. Depending on the biometric modality evaluated, one single test subject may provide biometric samples from several instances to be used in mated and inter-individual non-mated comparisons. For example, for an evaluation of a mobile device using fingerprint technology, one single person could provide biometric data coming from several fingers. Evaluator shall report if this possibility was used.

1) To be published. Current stage: 40.20.

The evaluator shall report test crew size. Age and gender distributions of the test crew shall be reported. Ethnicity should be reported. If the mobile device has an identified market target, the test crew should be as representative as possible of this target population.

For gender, there shall be no less than 40 % of male or female and no more than 60 % of male or female in the test crew. There are no other gender constraints, as long as those on male and female are respected.

6.1.3 Test subject interaction

Evaluation of biometric acquisition on mobile devices shall try to emulate an unsupervised environment by having limited intervention from the evaluator. The evaluator shall remain in an observer role only while the test subject interacts with the mobile device; any guidance shall be provided only by the mobile device, for example through its graphical user interface or audible instructions.

6.1.4 Modality specific consideration

6.1.4.1 Fingerprint

If the biometric system is designed to acquire and verify one fingerprint, one subject may provide biometric samples coming from several fingers. For non-mated comparison trials, intra-individual comparisons shall not be included. The number of samples given by one single person should be up to four different fingers and should be constrained to the index, thumb, or middle fingers from both hands.

If the biometric system is designed to acquire and verify several fingerprints simultaneously (i.e. through direct-view acquisition using regular camera), one subject may provide biometric samples coming from the left and right hands. For non-mated comparison trials, intra-individual comparisons shall not be included.

6.1.4.2 Face

Auxiliary means, such as selfie sticks, should not be utilized when collecting self-made face images.

6.1.4.3 Iris

If the biometric system is designed to acquire and verify one iris, one subject may provide biometric samples coming from the left and right irises. For non-mated comparison trials, intra-individual comparisons shall not be included. If the mobile device acquires and verifies both irises, one person corresponds to one test subject only.

If the mobile device does not have dedicated near infrared (NIR) acquisition module and instead relies on a regular camera to perform visible wavelength iris acquisition, the subject's eye colour will have a considerable impact on performance. The evaluator shall in this case report eye colour distribution. In cases where the iris-based biometric system generates comparison or capture errors, the evaluator should examine the occurrence of such errors as a function of eye colour.

6.1.4.4 Voice

If the biometric system is designed to acquire and verify the voice, the following aspects regarding data collection should be considered:

— The scenario in which biometric data will be collected are not only defined by the environmental conditions, but also the way in which users interact with the biometric sensor, which is embedded on the mobile device. There are a variety of positions, such as handheld, or over a table, and every position may involve different orientations and different distances from the speaker to the microphone.

  Therefore, the conditions in which the ToE is evaluated should describe not only the environmental conditions such as the type and level of noise but also the position of the phone and its orientation relative to the speaker.

— Test subjects characteristics. Specifically, relevant factors such as the native language and accent of the speaker should be considered as representative factors of the target population.

— The operation mode of the biometric system. In speaker recognition the operation mode of the system may be classified into text-dependent and text-independent.

In a text-dependent system, users are enrolled by the repetition of the same phrase several times. Then they just have to say the passphrase to be recognized. This passphrase can be chosen from a predefined catalogue or created freely by each user.

In a text-independent system, users are not forced to use a particular speech or passphrase. They are enrolled by speaking naturally for about 30 sec, and recognized by speaking in the same way but just for a short time, about 3 sec.

As a consequence, the type of phrases, the length and the content of the text for the data collection should be defined according to the operation mode of the ToE and the target application.

Test subjects shall be in good health with a normal sounding voice ensuring that their normal voice is registered. If a test subject is sick and has either lost their voice or has a hoarse voice, they shall not participate in the test.

6.1.4.5 Gestures, signatures, or signs

The test report shall describe whether the gestures were used to enter passcodes, write signatures, or draw other signs and whether the gestures were made via finger or stylus.

If the biometric system is designed to acquire data from the touch screen, accelerometer and gyro while the user is instructed to write on the touch screen, the following aspects regarding data collection should be considered:

— A test subject shall use the same finger for writing on the screen during the evaluation.

— A test subject with excessively sweaty hands shall not participate in the test.

— The test subject shall hold the mobile in a comfortable position and in the same orientation as much as possible during the evaluation.

— Mobile devices shall not have a screen protector and/or case protector.

In case the system requires the test subject to write in a specific language it is desired that the test subject will be a native speaker of this language.

6.1.4.6 Vascular

Vascular biometrics is technology that uses the vein structure within the hand, palm, or finger.

If the biometric system is designed to acquire and verify one vascular pattern, one subject may provide biometric samples coming from several body parts. For non-mated comparison trials, intra-individual comparisons shall not be included. One person may provide biometric samples of right hands and left hands for the back of the hand and palm, or biometric samples of thumbs, indexes and middle fingers from both hands for finger.

In the technologies using a finger vein, if the biometric system is designed to acquire and verify several vascular patterns simultaneously, one subject may provide biometric samples coming from the left and right hands. For non-mated comparison trials, intra-individual comparisons shall not be included.

6.1.4.7 Palm

If the biometric system is designed to acquire and verify one palm image, one subject may provide biometric samples coming from the left and right hands. For non-mated comparison trials, intra-individual comparisons shall not be included.

6.2 Test method

6.2.1 Enrolment

The enrolment procedure shall follow the requirements in ISO/IEC 19795-1.

6.2.2 Iterative and multi session enrolment

Some mobile devices integrate biometric systems which use an iterative enrolment strategy. After the initial enrolment transaction, each verification transaction may be used to improve and/or update the reference biometric data.

The evaluator shall emulate this process by enforcing a defined number of verification transactions before evaluating FAR and FRR. The number of training verification transactions may follow a best practice guideline defined by the system manufacturer. The evaluator shall report how many training transactions were made, the time-lapse between each of them and if the acquisition environment changed between them. After the training transactions sessions are finished, the evaluator shall switch off the updating of the reference template to have a fixed reference biometric data when evaluating the system.

6.2.3 Verification

As an embedded full biometric system, a mobile device may integrate acquisition and comparison subsystems in an interconnected way.

— For some biometric systems, acquisition and comparison are sequential operations. A verification presentation finishes when a sample with a good enough quality has been acquired. The resulting biometric sample is then compared with the biometric reference.

— For other biometric systems, the acquisition and comparison subsystem are operated simultaneously. The verification presentation will continuously provide biometric probes to the comparison subsystem which will compare them with the biometric reference. The acquisition process will continue as long as a positive verification has not been obtained, or as long as a time limit has not been reached. For such a system, if FAR is for example evaluated off-line and as the expected result for an impostor comparison is a timeout, all verification transactions intended to be used for impostor tests shall reach the timeout and be fully stored for future usage.

6.3 Performance measurement

6.3.1 Metrics

An evaluation of a mobile device shall at least report:

— FTE;

— FTA;

— FRR;

— FAR.

6.3.2 Optional technology evaluation for lower FAR claims

For evaluation of FAR claims below 0,01 %, the evaluator may proceed to a technology evaluation of the matching component.

The ToE may then be tested in two stages:

— Stage 1: Scenario evaluation: the evaluator shall evaluate FAR with the data collected from the test crew. The test crew dataset shall be large enough to have at least statistical significance for evaluating a FAR claim of 0,01 % or below.

— Stage 2: Technology evaluation: If a FAR claim of 0,01 % or below is verified at stage 1, the evaluator may use a pre-existing dataset not acquired from the test crew for this second stage. This pre-existing dataset may be provided by the supplier and shall be acquired with the same sensor. The pre-existing dataset should be large enough to allow using Rule of 30 when evaluating the FAR claim. The evaluator shall validate that the pre-existing dataset is consistent with the test crew dataset by sub-sampling the pre-existing dataset to the size of the test crew dataset and evaluating FAR by bootstrapping. The mean of the bootstrap distribution shall be below or equal to the value measured at stage 1.

6.3.3 Guidance for target requirements evaluation

Table 5 gives recommendations on the approach to follow based on the FAR requirements to have statistical significance in a cost-affordable way.

Table 5 — Recommendation for statistical approach and test size for a given FAR target

FAR target	Statistical approach	Test size
FAR target > 0,1 %	Evaluator should use Rule of 30	Test size should follow Rule of 30 requirements (see Table 1)
0,1 % ≥ FAR target ≥ 0,01 %	Evaluator should use Rule of 3, if applicable Evaluator may use bootstrapping	Test size should follow Rule of 3 requirements (see Table 1)
0,01 % > FAR target	Evaluator should use two-stage evaluation (see 6.3.2)	Test size shall at least be 246 (requirement for Rule of 3 at 0,01 %) If used, pre-existing dataset shall at least be 530

6.4 Considerations for third party evaluation

6.4.1 General

This clause gives recommendations and requirements specific for an evaluation of a mobile device by a third party. Typically, the most common use case would be an independent evaluation of a mobile device by a laboratory accredited to deliver certifications by checking that the system under test meets some accuracy and functional requirements. As discussed in 5.1, the system provider shall deliver customized mobile devices and evaluation software development kit (SDK) to make such an evaluation possible.

6.4.2 Specifications for the system under test

6.4.2.1 Querying capabilities

The system under test should be able to be queried regarding:

— operating points supported;

— how many distinct biometrics are allowed to be enrolled;

— whether it does template adaptation;

— its rate limiting parameters.

The evaluator should use this capability to perform its evaluation.

6.4.2.2 Configuration capabilities

The system under test should be able to:

— have its operating point configured;

— turn off its adaptation ability to export verification transactions;

— have its rate limiting ability turned off;

— be configured to limit the maximum number of attempts per transaction.

The system under test shall include all functionality required for biometrics: the Biometric Data Capture, Signal Processing, Comparison, and Decision functionality, whether implemented in hardware or software. A system under test shall be provided for each Allowed Integration, e.g. different thickness of glass. The integration manual is provided for reference to the evaluator, it shall be coherent with the configuration and operation of the Test Harness.

The system under test shall be provided to the evaluator from the Vendor in the form of a Common Test Harness which is set up to allow the evaluator to perform the testing efficiently and identify the components of the Test Harness as being of the system under test.

6.4.2.3 Export transaction capability

The system under test should be able to export enrolment transactions and verification transactions.

6.4.2.4 Offline matching capability

The system under test should provide an offline matcher that utilizes exported transactions.

6.4.2.5 Test harness description

To follow the recommendations of 6.4.2.1 to 6.4.2.4, the test harness should consist of:

a) Configurable Enrolment system which:

   1) Selects the operating point(s) to be evaluated.

   2) Has enrolment hardware/software as will be executed by the mobile device.

   3) Includes a biometric data capture sensor and enrolment software.

   4) Can clear an enrolment.

   5) Can store an enrolment from acquired biometric sample(s) for use in on-line verification evaluation.

   6) Can provide enrolment templates from acquired biometric sample(s) defined as “user’s store reference measure based on features extracted from enrolment samples” for use in off-line verification evaluation.

   7) Indicates a failure to enrol.

b) Configurable Verification on-line system which:

   1) Selects the operating point(s) to be evaluated.

   2) Has verification hardware/software as will be executed by the mobile device.

   3) Includes a biometric data capture sensor, a biometric matcher, and a decision module.

4) Captures features from an acquired biometric sample to be compared against an enrolment template.

5) Makes accept/reject decision at a specific operating point.

6) Indicates an on-line failure to acquire.

7) Indicates an on-line decision (accept or reject).

8) Provides a set of acquired biometric sample(s) from an on-line verification transaction (this is called a stored verification transaction). This will be used for off-line verification.

c) Configurable Verification off-line software, which:

1) Selects the operating point(s) to be evaluated.

2) Has verification software as will be executed by the mobile device.

3) Accepts an enrolment template and the stored verification transaction and performs matching in off-line batch mode.

4) Provides a decision (accept or reject).

6.4.3 Consistency of system under test online and offline

The evaluator shall utilize a method to check the consistency of results from the online and offline matchers from the system under test and report the conclusion of this analysis.

6.4.4 Checking a system provider self-attestation

A system provider may claim that the system under test achieves a better performance than what can be measured in the third-party evaluation process. The evaluator may then check that its results are consistent with those of the system provider, even if not actually reproducing the results.

EXAMPLE A certification scheme assesses that a system achieves 0,03 FRR for 0,001 FAR. The system provider, using internal datasets, measures that its system actually achieves 0,03 FRR for 0,0001 FAR. The laboratory responsible for the test collects enough data to check the 0,001 FAR requirement, but cannot evaluate a 0,0001 FAR claim. The vendor provides its bootstrapping distributions to the evaluator which checks them against its own bootstrapping distributions. The evaluator makes a statement that the results look consistent.

6.5 Reporting

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

a) Evaluator details:

1) company/organization,

2) date evaluation completed,

3) date test report published;

b) System and test participant details:

1) system specifications:

i) manufacturer,

ii) model,

iii) version,

iv) firmware,

v) biometric modality;

2) Application software (if applicable):

i) provider,

ii) title,

iii) version,

iv) build;

c) Summary of performance results:

1) time limit,

2) FTE rate,

3) FTA rate,

4) FRR,

5) FAR.

The report shall also include the following test details (according to ISO/IEC 19795-1):

a) details of the system(s) tested, concept of operations;

b) details of the test environment;

c) test subject information and instructions;

d) test subject training;

e) attended/unattended testing (enrolment and verification);

f) guidance to test subject;

g) test order and acclimatization;

h) test subject identifiers;

i) enrolment level of effort;

j) verification level of effort;

k) reference adaptation;

l) appropriateness of levels of effort;

m) multiple visits and transactions;

n) executing genuine and impostor trials;

o) time separation between enrolments and test transactions;

p) details of abnormal cases and data excluded from analysis;

q) estimated uncertainties (and method of estimation);

r) data collection;

s) test crew general;

t) test crew habituation;

u) test crew composition;

v) test subject management;

w) performance measurement;

x) enrolment;

y) failure to acquire;

z) verification metrics;

aa) interim analyses;

bb) deviations from the guidelines.

Annex A
(informative)

Sample test report

This Annex gives an informative implementation of 6.5 reporting in the form of a sample test report. This sample test report consists of details on the test laboratory and test participant, summary of fundamental and verification system metrics, and test details.

A.1 Evaluator details

Table A.1 presents the details of the test laboratory.

Table A.1

Company/Organization	*[Evaluator provides relevant information]*
Date evaluation completed	*[Evaluator provides relevant information]*
Date test report published	*[Evaluator provides relevant information]*

A.2 Test participant details (ISO/IEC 19795-2:2007,7.4.2)

Table A.2 presents the details of the test participant including system specifications, and application software.

Table A.2

System specifications	
Manufacturer	*[Evaluator provides relevant information]*
Model	*[Evaluator provides relevant information]*
Version	*[Evaluator provides relevant information]*
Firmware	*[Evaluator provides relevant information]*
Biometric modality	*[Fingerprint, Face, Iris, Voice, Gestures, Vascular, Palm, other]*
Application software *(if applicable)*	
Provider	*[Evaluator provides relevant information]*
Title	*[Evaluator provides relevant information]*
Version	*[Evaluator provides relevant information]*
Build	*[Evaluator provides relevant information]*

A.3 Summary of performance results

Table A.3 presents test results according to the criteria specified in ISO/IEC 19795-1:2006, 10.2 and this document.

Table A.3

Fundamental and verification system metrics (ISO/IEC 19795-1:2006, 10.1 and 10.2 and 6.3.1 of this document)	
Time limit	*[Evaluator provides time limit]*
Failure-to-enrol (FTE) Rate	*[Evaluator provides FTE]*
Failure-to-acquire (FTA) Rate	*[Evaluator provides FTA]*
False Reject Rate (FRR)	*[Evaluator provides FRR]*
False Accept Rate (FAR)	*[Evaluator provides FAR]*

A.4 Test details

Table A.4 presents test details according to the criteria specified in this document or in ISO/IEC 19795-2:2007 or ISO/IEC 19795-1:2006 when adequate.

Table A.4

#	Relevant technical clause	Scenario test parameter	Test data specification/ Possible values
1	ISO/IEC 19795-2 7.1.1.1	Details of the system(s) tested, Concept of operations (scenario/application description)	*[Evaluator provides narrative]*
2	This document 5.2.3	Details of the test environment, Evaluation environment	*[Recommended environment is indoor conditions, with no noise in which the mobile device is hand-held by the user. Evaluator provides basis and narrative if different.]*
3	ISO/IEC 19795-2 7.1.2.1	Test subject information and instructions	*[Evaluator provides narrative.]*
4	This document 6.1.1	Test subject training	*[Evaluator provides narrative.]*
5	This document 6.1.3	Attended/unattended testing	*[Evaluator shall not interact, and reports if he is present during test.]*
6	This document 6.1.3	Guidance to test subject	*[Evaluator provides description of guidance provided by mobile device GUI.]*
7	ISO/IEC 19795-2 7.1.2.5	Test order and acclimatization	*[Evaluator provides narrative.]*
8	ISO/IEC 19795-2 7.1.2.6	Test subject identifiers	*[Evaluator provides narrative.]*
9	This document 6.2.1	Enrolment level of effort	*[Evaluator provides basis and narrative for choice of number of attempts by transaction and number of presentations by attempts. Evaluator reports statistics about the required number of attempts and presentations.]*
10	This document 6.2.3	Verification level of effort	*[Evaluator provides basis and narrative for choice of number of attempts by transaction and number of presentations by attempts. Evaluator reports statistics about the required number of attempts and presentations.]*

Table A.4 *(continued)*

#	Relevant technical clause	Scenario test parameter	Test data specification/ Possible values
11	This document 6.2.2	Reference adaptation	*[Evaluator describes reference adaptation mechanisms, if any, and describes how they were emulated.]*
12	ISO/IEC 19795-2 7.1.3.4	Appropriateness of levels of effort	*[Evaluator provides basis and narrative based on provider policy.]*
13	This document 5.2.2.1 and 5.2.4.1	Multiple visits and transactions	*[Evaluator provides basis and narrative for choice of number of visits.]*
14	This document 5.2.2	Executing genuine and impostor trials Number of transactions per test subject at each visit	*[Evaluator provides basis for choice of number of transactions per visits and reports estimates on impact for tests independence.]*
15	This document 5.2.4	Time separation between enrolments and test transactions	*[Evaluator provides basis for choice of time separation and narrative for impact on result validity.]*
16	ISO/IEC 19795-1 10.5	Details of abnormal cases and data excluded from analysis	*[Evaluator reports abnormal cases and basis for excluding from analysis.]*
17	ISO/IEC 19795-1 Annex B	Estimated uncertainties (and method of estimation)	*[Evaluator provides basis and narrative.]*
18	ISO/IEC 19795-2 7.1.6	Data collection	*[Evaluator provides examples of data collection elements.* *EXAMPLE Logs or spreadsheets.]*
19	ISO/IEC 19795-2 7.2	Test crew general	*[Evaluator provides narrative, number of subjects.]*
20	This document 6.1.1	Test crew habituation	*[Evaluator validates and reports test crew familiarity with tested devices.]*
21	This document 6.1.2	Test crew composition	*[Evaluator reports age and gender distribution and may report ethnicity.]*
22	ISO/IEC 19795-2 7.2.4	Test Subject management	*[Evaluator provides basis and narrative.]*
23	This document 6.3	Performance measurement	*[Evaluator provides basis and narrative for choice of Rule used to assess requirements and elaborates on reliability of results.]*
24	This document 6.2.1 and 6.3.1	Enrolment	*[Evaluator reports FTE rate, number of test subjects, transactions, attempts and presentations]*
25	This document 6.2.3 and 6.3.1	Failure to acquire	*[Evaluator reports FTA rate, number of test subjects, transactions, attempts and presentations]*
26	This document 6.3.1	Verification metrics	*[Evaluator reports FAR and FRR rate, number of test subjects, number of mated and non-mated transactions, attempts and presentations. Evaluator reports time distribution for transactions. If bootstrapping used, evaluator reports bootstrap distribution. Evaluator may also report FNMR and FMR if available, include DET curves and distribution of comparison scores.]*

Table A.4 *(continued)*

#	Relevant technical clause	Scenario test parameter	Test data specification/ Possible values
27	ISO/IEC 19795-2 7.3.7	Interim analyses	*[Evaluator provides basis and narrative for management of interim analysis.]*
28	ISO/IEC 19795-2 7.4.1	Deviations from the guidelines	*[Evaluator provides basis and narrative if deviations did occur.]*