INTERNATIONAL STANDARD

ISO/IEC 30107-3

Second edition
2023-01

Information technology — Biometric presentation attack detection —

Part 3: Testing and reporting

Technologies de l'information — Détection d'attaque de présentation en biométrie —

Partie 3: Essais et rapports d'essai

Reference number
ISO/IEC 30107-3:2023(E)

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

Published in Switzerland

Contents

Page

Foreword

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

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

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

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

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

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

This second edition cancels and replaces the first edition (ISO/IEC 30107-3:2017), which has been technically revised.

The main changes are as follows:

— the relative impostor attack presentation accept rate has been added (13.4.4);

— information on roles in presentation attack detection testing have been added (Annex C);

— general technical clarifications and improvements have been made.

A list of all parts in the ISO/IEC 30107 series can be found on the ISO and IEC websites.

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

Introduction

The presentation of an artefact or of human characteristics to a biometric capture subsystem in a fashion intended to interfere with system policy is referred to as a presentation attack. The ISO/IEC 30107 series deals with techniques for the automated detection of presentation attacks. These techniques are called presentation attack detection (PAD) mechanisms.

As is the case for biometric recognition, PAD mechanisms are subject to false positive and false negative errors. False positive errors wrongly categorize bona fide presentations as attack presentations, potentially flagging or inconveniencing legitimate users. False negative errors wrongly categorize presentation attacks (also known as attack presentations) as bona fide presentations, potentially resulting in a security breach.

Therefore, the decision to use a specific implementation of PAD will depend on the requirements of the application and consideration of the trade-offs with respect to security, evidence strength and efficiency.

The purpose of this document is as follows:

— to define terms related to biometric PAD testing and reporting, and

— to specify principles and methods of performance assessment of biometric PAD, including metrics.

This document is directed at vendors or test laboratories seeking to conduct evaluations of PAD mechanisms.

Biometric performance testing terminology, practices and methodologies for statistical analysis have been standardized through ISO and Common Criteria. False accept rate (FAR), false reject rate (FRR) and failure to enrol rate (FTE) are widely used to characterize biometric system performance. Biometric performance testing terminology, practices and methodologies for statistical analysis are only partially applicable to the evaluation of PAD mechanisms due to significant fundamental differences between biometric performance testing concepts and PAD mechanism testing concepts. These differences can be categorized as follows.

a) **Statistical significance**

Biometric performance testing utilizes a statistically significant number of test subjects, representative of the targeted user group. Error rates are not expected to vary significantly when adding more test subjects or using a completely different group.

In PAD testing, many biometric modalities can be attacked by a large or indeterminate number of potential presentation attack instrument species (PAIS). In these cases, it is very difficult or even impossible to have a comprehensive model of all possible presentation attack instruments (PAIs). Hence, it could be impossible to find a representative set of PAIS for the evaluation. Therefore, measured error rates of one set of PAIs cannot be assumed to be applicable to a different set.

PAIS present a source of systematic variation in a test. Different PAIs can have significantly different error rates. Additionally, within any given PAIS, there is random variation across instances of the PAI series. The number of presentations required for a statistically significant test scales linearly with the number of PAIS of interest. Within each PAIS, the uncertainty associated with a PAD error rate estimate depends on the number of artefacts tested and the number of individuals.

EXAMPLE 1 In fingerprint biometrics, many potent artefact materials are known, but any material or material mixture that can present fingerprint features to a biometric capture device is a possible candidate. Since artefact properties such as age, thickness, moisture, temperature, mixture rates and manufacturing practices can have a significant influence on the output of the PAD mechanism, it is easy to define tens of thousands of PAIS using current materials. Hundreds of thousands of presentations would be needed for a proper statistical analysis, and even then, resulting error rates cannot be transferred to the next set of new materials.

PAI presentation can also be source of variation in a test. Variation in pressure, position or even PAI presenter characteristics can impact PAD performance.

b) **Comparability of test results across systems**

In biometric performance testing, application-specific error rates based on the same corpus of biometric samples can be used to compare different biometric systems or different configurations. Results can be used to unambiguously compare and assess system performance. By contrast, when using error rates to benchmark PAD mechanisms, interpreting results can be highly dependent on the intended application.

EXAMPLE 2 In a given testing scenario with 10 PAIS (presented 100 times), $System_1$ detects 90 % of attack presentations and $System_2$ detects 85 %. $System_1$ detects all presentations for 9 PAIS but fails to detect all presentations with the 10th PAIS. $System_2$ detects 85 % of all PAIS. Which is better? In a security analysis $System_1$ would be worse than $System_2$, because revealing the 10th PAIS would orient an attacker such that they could use this method to defeat the capture device all the time. However, if attackers could be prevented from using the 10th PAIS, $System_1$ would be better than $System_2$, because individual rates indicate that it is possible to overcome $System_2$ with all PAIS.

c) **Cooperation**

Many biometric performance tests address applications such as access control in which subjects are cooperative. Errors due to incorrect operation are an issue of a lack of knowledge, experience or guidance rather than intent. Significant uncooperative behaviour in a group is not part of the underlying "biometric model" and would render the determined error rates almost useless for biometric performance testing.

PAD tests include subjects whose behaviour is not cooperative. Attackers will try to find and exploit any weakness of the biometric system, circumventing or manipulating its intended operation. Presentation attack types, based on the experience and knowledge of the tester, can change the success rates for an attack dramatically. Hence it can be difficult to define testing procedures that measure error rates in a fashion representative of cooperative behaviour.

d) **Automated testing**

In biometric performance testing, it is often possible to test comparison algorithms using databases from devices or sensors of similar quality. Performance can be measured in a technology evaluation using previously collected corpora of samples as specified in ISO/IEC 19795-1.

In PAD testing, data from the biometric capture device (e.g. digitized fingerprint images) can in some cases be insufficient to conduct evaluations. Biometric systems with PAD mechanisms often contain additional sensors to detect specific properties of a biometric characteristic. Hence, a database previously collected for a specific biometric system or configuration is not necessarily suitable for another biometric system or configuration.

Even slight changes in the hardware or software could make earlier measurements useless. It is generally impractical to store multivariate synchronized PAD signals and replay them in automated testing. Therefore, automated testing is often not an option for testing and evaluating PAD mechanisms.

e) **Quality and performance**

In biometric performance testing, performance is usually linked directly to biometric data quality. Low-quality samples generally result in higher error rates while a test with only high-quality samples will generally result in lower error rates. Quality metrics are therefore often used to improve performance (dependent on the application).

In PAD testing, even though low biometric quality can cause an artefact to be unsuccessful, there is no reason to assume a certain quality level from artefacts in general. Samples from artefacts can exhibit better quality than samples from human biometric characteristics. Without a model of attacker skill, it seems valid (at least in a security evaluation) to assume a "worst case" scenario where the attacker always uses the best possible quality. That way, one can at least determine a guaranteed minimal detection rate for the specific test set while reducing the number of necessary tests at the same time.

It is then a matter of rating the attack potential of successful artefacts (effort and expertise for the needed quality) in order to assess the security level, as is the practice in Common Criteria evaluations.

Based on the differences in a) through e), the following general comments regarding error rates and metrics related to PAD mechanisms can be derived.

— In an evaluation, PAIS are analysed/rated separately.

— Attack presentation classification error rates other than 0 % for a PAIS only prove that the PAI can be successful. A different tester can potentially achieve a higher or lower attack presentation classification error rate. Further, training to identify the relevant material and presentation parameters could increase the attack presentation classification error rate for this PAIS. The experience and knowledge of the tester, as well as the availability of the necessary resources, are significant factors in PAD testing and are taken into account when conducting comparisons or performance analysis.

Error rates for PAD mechanisms are determined by the specific context of the given PAD mechanism, the set of PAIS, the application, the test approach, and the tester. Error rates for PAD mechanisms are not necessarily comparable across similar tests, and error rates for PAD mechanisms are not necessarily reproducible by different test laboratories.

Information technology — Biometric presentation attack detection —

Part 3: Testing and reporting

1 Scope

This document establishes:

— principles and methods for the performance assessment of presentation attack detection (PAD) mechanisms;

— reporting of testing results from evaluations of PAD mechanisms; and

— a classification of known attack types (Annex A).

Outside the scope are:

— standardization of specific PAD mechanisms;

— detailed information about countermeasures (i.e. anti-spoofing techniques), algorithms or sensors; and

— overall system-level security or vulnerability assessment.

The attacks considered in this document take place at the biometric capture device during presentation. Any other attacks are considered outside the scope of this document.

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 15408-1, *Information security, cybersecurity and privacy protection — Evaluation criteria for IT security — Part 1: Introduction and general model*

ISO/IEC 15408-2, *Information security, cybersecurity and privacy protection — Evaluation criteria for IT security — Part 2: Security functional components*

ISO/IEC 15408-3, *Information security, cybersecurity and privacy protection — Evaluation criteria for IT security — Part 3: Security assurance components*

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

ISO/IEC 30107-1, *Information technology — Biometric presentation attack detection — Part 1: Framework*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 19795-1, ISO/IEC 2382-37, ISO/IEC 30107-1 and the following apply.

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

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

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

3.1 Attack elements

3.1.1
presentation attack
attack presentation
presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system

Note 1 to entry: An attack presentation can be a single attempt, a multi-attempt transaction, or another type of interaction with a subsystem.

3.1.2
bona fide presentation
interaction of the biometric test subject and the biometric data capture subsystem in the fashion intended by the policy of the biometric system

Note 1 to entry: Bona fide is analogous to normal or routine, when referring to a bona fide presentation.

Note 2 to entry: Bona fide presentations can include those in which the user has a low level of training or skill. Bona fide presentations encompass the totality of good-faith presentations to a biometric data capture subsystem.

3.1.3
attack type
elements and characteristics of a presentation attack, including presentation attack instrument (PAI) species, concealer or impostor attack, degree of supervision, and method of interaction with the capture device

3.1.4
test approach
totality of considerations and factors involved in presentation attack detection (PAD) evaluation

Note 1 to entry: Elements of a test approach are given in Clauses 6–10.

Note 2 to entry: A test approach refers to all processes, factors and aspects specified in the course of the evaluation.

3.1.5
item under test
IUT
implementation that is the object of a test assertion or test case

Note 1 to entry: The IUT is the equivalent of the "target of evaluation" (TOE) in Common Criteria evaluations.

3.1.6
presentation attack instrument species
PAIS
class of presentation attack instruments (PAIs) created using a common production method and based on different biometric characteristics

EXAMPLE 1 A set of fake fingerprints all made in the same way with the same materials but with different friction ridge patterns would constitute a PAIS.

EXAMPLE 2 A specific type of alteration made to the fingerprints of several test subjects would constitute a PAIS.

Note 1 to entry: The term “recipe” is often used to refer to how to make a PAIS.

Note 2 to entry: PAIs of the same species can have different success rates due to variability in the production process or in the PAI source.

3.1.7
PAI series
class of presentation attack instruments (PAIs) created using a common production method and based on the same biometric characteristics

EXAMPLE A set of fake fingerprints all made in the same way with the same materials and with the same friction ridge pattern.

Note 1 to entry: Depending on the experimental goals, an evaluation can potentially utilize multiple series, each with different production methods or sources. While tests involving several biometric sources can demonstrate generality of a PAI species, they add variation associated with individual human traits.

3.1.8
target of evaluation
TOE
IT product that is the subject of the evaluation within the context of the Common Criteria

Note 1 to entry: The TOE is the equivalent of the "item under test" (IUT) in Common Criteria evaluations.

3.1.9
attack potential
measure of the capability to attack a target of evaluation (TOE) given the attacker’s knowledge, proficiency, resources and motivation

3.2 Metrics

3.2.1
attack presentation classification error rate
APCER
proportion of attack presentations using the same presentation attack instrument (PAI) species incorrectly classified as bona fide presentations by a presentation attack detection (PAD) subsystem in a specific scenario

3.2.2
attack presentation classification error rate at the given attack potential
$APCER_{AP}$
attack presentation classification error rate (APCER) of the most successful presentation attack instrument (PAI) species within a given attack potential

3.2.3
bona fide presentation classification error rate
BPCER
proportion of bona fide presentations incorrectly classified as presentation attacks in a specific scenario

3.2.4
attack presentation acquisition rate
APAR
proportion of attack presentations using the same presentation attack instrument (PAI) species from which the data capture subsystem acquires a biometric sample of sufficient quality

3.2.5
attack presentation non-response rate
APNRR
proportion of attack presentations using the same presentation attack instrument (PAI) species that cause no response at the presentation attack detection (PAD) subsystem or data capture subsystem

EXAMPLE A fingerprint system can potentially not register or react to the presentation of a PAI due to the PAI's lack of realism.

3.2.6
bona fide presentation non-response rate
BPNRR
proportion of bona fide presentations that cause no response at the presentation attack detection (PAD) subsystem or data capture subsystem

3.2.7
impostor attack presentation accept rate
IAPAR
in a full-system evaluation of a verification system, proportion of impostor attack presentations using the same presentation attack instrument (PAI) species that result in accept

3.2.8
impostor attack presentation accept rate at the given attack potential
$IAPAR_{AP}$
in a full-system evaluation of a verification system, proportion of impostor attack presentations of the most successful presentation attack instrument (PAI) species at the given attack potential that result in accept

3.2.9
concealer attack presentation reject rate
CAPRR
in a full-system evaluation of a verification system, proportion of concealer attack presentations using the same presentation attack instrument (PAI) species that result in a reject decision

Note 1 to entry: CAPRR is transactional and involves both biometric comparison and presentation attack detection (PAD), whereas non-match rates are attributed to the comparison algorithm.

Note 2 to entry: A biometric concealer compelled to unlock a device might want to be rejected. In this case, they might use a PAI to obscure their biometric characteristic, leading to a rejection.

3.2.10
impostor attack presentation identification rate
IAPIR
in a full-system evaluation of an identification system, proportion of impostor attack presentations using the same presentation attack instrument (PAI) species in which the targeted reference identifier is among the identifiers returned, or, depending on the intended use case, at least one identifier is returned by the system,

Note 1 to entry: An attacker might be both an impostor (trying to match an existing non-self enrollee) and a concealer (obscuring their real biometric sample with a PAI).

3.2.11
concealer attack presentation non-identification rate
CAPNIR
in a full-system evaluation of an identification system, proportion of concealer presentation attacks using the same presentation attack instrument (PAI) species in which the reference identifier of the concealer is not among the identifiers returned, or, depending on the intended use case, in which no identifiers are returned

Note 1 to entry: In a negative identification system, such as a block-list, the concealer could intend that no identifiers are returned to avoid scrutiny by a human operator.

Note 2 to entry: CAPNIR is transactional and involves both biometric comparison and presentation attack detection (PAD) whereas false negative identification rates are attributed to the comparison algorithm.

3.2.12
relative impostor attack presentation accept rate
RIAPAR
sum of impostor attack presentation identification rate (IAPAR) and false reject rate (FRR) at a fixed decision threshold

3.2.13
PAD subsystem processing duration
PS-PD
duration required for the presentation attack detection (PAD) subsystem to classify PAD data

3.2.14
data capture subsystem processing duration
DCS-PD
duration required for the data capture subsystem to acquire a sample, inclusive of the presentation attack detection (PAD) subsystem processing duration (if applicable)

3.2.15
full system processing duration
FS-PD
duration required for the data capture subsystem and comparison subsystem to acquire and process a sample, inclusive of the presentation attack detection (PAD) subsystem processing duration (if applicable)

3.3 Test roles

3.3.1
biometric impostor
subversive biometric capture subject who performs a biometric impostor attack

Note 1 to entry: Biometric impostors include capture subjects using PAIs and capture subjects presenting intentionally damaged or altered characteristics.

[SOURCE: ISO/IEC 2382-37:2022, 37.07.13, modified — Original notes to entry deleted; new Note 1 to entry added.]

3.3.2
biometric concealer
subversive biometric capture subject who performs a biometric concealment attack

Note 1 to entry: Biometric concealers include capture subjects using PAIs and capture subjects presenting intentionally damaged characteristics.

[SOURCE: ISO/IEC 2382-37:2022, 37.07.21, modified — Note 1 to entry added.]

3.3.3
PAI presenter
individual or mechanism presenting the presentation attack instrument (PAI) to the biometric system

3.3.4
PAI source
individual or mechanism from which biometric samples are obtained for use in a presentation attack instrument (PAI), realized in a PAI series

3.3.5
PAI creator
individual or mechanism responsible for the conception, formulation, design and realization of a presentation attack instrument (PAI) species

4 Abbreviated terms

The abbreviated terms below are used in this document.

AP	attack potential
APAR	attack presentation acquisition rate
APCER	attack presentation classification error rate
$APCER_{AP}$	attack presentation classification error rate at the given attack potential
APNRR	attack presentation non-response rate
ATM	automatic teller machine
BPCER	bona fide presentation classification error rate
BPNRR	bona fide presentation non-response rate
BSI	German Federal Office for Information Security
CAPNIR	concealer attack presentation non-identification rate
CAPRR	concealer attack presentation reject rate
CCRA	Common Criteria Recognition Arrangement
DCS-PD	data capture subsystem processing duration
DET	detection error tradeoff
EAL	Evaluation Assurance Level
FTA	failure to acquire rate
FTE	failure to enrol rate
FAR	false accept rate
FNIR	false negative identification rate
FPIR	false positive identification rate
FRR	false reject rate

FSDPP	fingerprint spoof detection protection profiles
FS-PD	full system processing duration
IAPAR	impostor attack presentation accept rate
$IAPAR_{AP}$	impostor attack presentation accept rate at the given attack potential
IAPIR	impostor attack presentation identification rate
IUT	item under test
PAD	presentation attack detection
PAI	presentation attack instrument
PAIS	presentation attack instrument species
PS-PD	PAD subsystem processing duration
RIAPAR	relative impostor attack presentation accept rate
TOE	target of evaluation

5 Conformance

To conform to this document, an evaluation of PAD mechanisms shall be planned, executed and reported in accordance with mandatory requirements as follows:

— those of Clause 6 and subclauses 11.1 and 13.1;

— for evaluations of PAD mechanisms in enrolment, 11.2;

— for evaluations of PAD mechanisms in verification, 11.3;

— for evaluations of PAD mechanisms in positive or negative identification, 11.4;

— for PAD subsystem evaluations, 13.2;

— for data capture subsystem evaluations, 13.3;

— for full-system evaluations of verification systems, 13.4.2.1;

— for full-system evaluations of positive identification systems, 13.4.2.2;

— for full-system evaluations of negative identification systems, 13.4.2.3.

Due to the potential sensitivity of information such as species formulation or system-specific vulnerabilities, additional [derived/summary] reports for general dissemination may redact or generalize sensitive information.

6 Presentation attack detection (PAD) overview

This document describes two types of presentation attackers: biometric impostors (a.k.a. impersonators) and biometric concealers. These types of attackers differ in that biometric impostors typically need to defeat PAD subsystems, pass quality checks, and match through comparison subsystems, whereas biometric concealers do not need to match through comparison subsystems.

While the desired impersonation or concealment outcome can lend itself towards a sub-set of attack types, any type of PAI can be used by either type of attacker (see Annex A for additional information).

Evaluations of PAD mechanisms and resulting reports shall specify the type of presentation attacker (biometric impostor or biometric concealer) considered in an evaluation.

Evaluations of PAD mechanisms are classifiable as one of three general types, increasing in specificity, as follows:

— generic, broad evaluations of PAD mechanisms of any device for an unknown application;

— application-focused evaluations of PAD mechanisms in which the set/range of attack types is selected to be appropriate to the application, such as those discussed in Clause 11; or

— product-specific evaluations of PAD mechanisms, used to test a supplier's claim of performance against a specific category of attack types.

Evaluations of PAD mechanisms and resulting reports shall describe the type of evaluation conducted as well as the attack types to be tested.

7 Levels of evaluation of PAD mechanisms

7.1 Overview

The evaluation of PAD mechanisms is determined by the item under test (IUT). PAD evaluations and resulting reports shall fully describe the IUT, including all configurations and settings as well as the amount of information available to the evaluator about PAD mechanisms in place. IUTs shall be categorized as follows.

— PAD subsystem.

— Data capture subsystem.

— Full system.

A PAD subsystem is hardware and/or software that implements a PAD mechanism and makes an explicit declaration regarding the detection of presentation attacks. Results of the PAD mechanism are accessible to the evaluator and are an aspect of the evaluation.

EXAMPLE 1 A PAD subsystem could be a fingerprint device that logs a PAD score or decision when a PAI is presented.

A data capture subsystem, consisting of capture hardware and/or software, couples PAD mechanisms and quality checks in a fashion opaque to the evaluator. The evaluator does not necessarily know whether the data capture subsystem utilizes presentation attack detection. Acquisition can be for the purpose of enrolment or recognition, but no comparison takes place in the data capture subsystem.

EXAMPLE 2 A data capture subsystem could be an iris collection device that fails to acquire a sample from an iris artefact, where it is impossible to determine whether failure to acquire is due to a liveness check or quality check (the implementation does not provide this level of transparency).

For simplicity, the term "quality check" encompasses feature extraction, segmentation, or any other automated processing function used to validate the utility of a biometric sample

A full system adds biometric comparison to the PAD subsystem or data capture subsystem, comprising a full end-to-end system. This leads to additional failure points for the PAI beyond PAD mechanisms and quality checks. In a full system, there can be one or multiple PAD mechanisms at different points in the system.

Evaluations of PAD mechanisms and resulting reports shall specify the applicable evaluation level, whether PAD subsystem, data capture subsystem, or full system. The resulting reports should discuss how the evaluation level influenced PAD testing.

7.2 General principles of evaluation of PAD mechanisms

Evaluations of PAD mechanisms shall cover a defined variety of attack types by utilizing a representative set of presentation attack instruments and a representative set of bona fide test subjects.

For the set of presentation attack instruments, evaluations of PAD mechanisms should be based on the appropriate evaluation level and on relevant attack types. Not all PAD mechanisms are designed to address all possible presentation attacks.

EXAMPLE A PAD mechanism designed to recognize an artificial biometric characteristic is not likely to be effective for detecting an altered biometric characteristic.

Once the types are defined, the number and range of presentation attack instruments to be evaluated should be specified. Establishing whether a specific attack type reproducibly succeeds does not require a very large number of presentations.

The evaluator shall define the parameters of the attack presentation to fully characterize the range of PAI presenter interactions with the IUT, to include the temporal boundaries of the presentation.

A representative set of bona fide test subjects is required to determine the frequency with which the PAD mechanism incorrectly classifies bona fide presentations. This is a critical part of PAD testing since a PAD mechanism could erroneously classify bona fide presentations as attack presentations. A high classification error rate for bona fide test subjects would reduce system usability.

The representativeness of bona fide presentations should consider test subject selection and size as described in ISO/IEC 19795-1. Particularly, total numbers of bona fide presentations should exceed that required by the Rule of 30.

In an evaluation of PAD mechanisms, the evaluator shall define bona fide presentations and representative test subjects for the target application and population. The evaluator shall also provide a rationale for these definitions.

Defining “bona fide” presentations and representative test subjects can be a challenge in evaluations of PAD mechanisms. In some cases, the evaluator can define bona fide presentations as those that conform to vendor or implementer specifications. However, in certain applications, bona fide or representative test subject interaction with data capture devices can encompass a wide range of behaviours and conditions. For example, a vendor can define a conformant presentation to a fingerprint sensor as one conducted with clean fingerprints. While one could conduct a test in which all test subjects without perfectly clean fingerprints are excluded, it is reasonable to expect that operational systems have some tolerance for a range of regular, reasonable, or typical fingerprint conditions. Otherwise, operational systems would have excessively high false rejection or failure to enrol (FTE) rates. This is particularly relevant to PAD testing, because bona fide presentation classification errors can be most frequently encountered among test subjects whose interactions with data capture devices, while sufficient for enrolment or biometric recognition, are only marginally conformant with vendor specifications.

7.3 PAD subsystem evaluation

PAD subsystem evaluations measure the ability of the PAD subsystem to correctly classify both attack presentations and bona fide presentations. An effective attack presentation will be incorrectly classified as a bona fide presentation, resulting in the defeat of the PAD subsystem.

PAD subsystem evaluations can focus on the effectiveness of the biometric capture device (mostly hardware, or possibly internal firmware) in terms of refusal to acquire a sample, including cases with or without automated indications of refusal. Such evaluations focus on rejecting presentation attack instruments. The output of the PAD subsystem can be discrete, such as a pass/fail to each PAI utilized.

Alternatively, PAD subsystem evaluations can focus on the effectiveness of a PAD algorithm, as exemplified by LivDet.[4] This type of PAD subsystem evaluation can be performed offline with a corpus of samples; the PAD subsystem determines whether samples come from an attack. Such tests are typically based on a collected database, analogous to technology tests in biometric performance evaluations.

If the PAD subsystem returns a PAD score, false-negative and false-positive error rates can be expressed parametrically as functions of the decision threshold (e.g. through a detection error trade-off curve).

Clause 10 provides an overview of factors when designing a test for PAD subsystems designed to recognize artefacts.

7.4 Data capture subsystem evaluation

In data capture subsystems, presentation attacks can fail for reasons other than detection in the PAD subsystem. For example, the data capture subsystem can fail to respond to a presentation attack, or a quality subsystem can reject the presentation attack. In data capture subsystems where PAD mechanisms are not implemented or where the evaluator does not have access to the results of PAD mechanisms, outcomes are based on whether the data capture subsystem has successfully acquired a sample. An effective presentation attack defeats both the PAD subsystem (if present and active) and the quality subsystem, resulting in the capture of a biometric sample.

7.5 Full system evaluation

Full system evaluations add a comparison subsystem to the IUT, generating a comparison score or candidate list. This is illustrated in ISO/IEC 30107-1:2016, Figure 3.

Depending on the implementation, a full system evaluation can encompass the following.

— **PAD subsystem, data capture subsystem and comparison subsystem** (for IUTs in which results of the PAD mechanism are accessible to the evaluator). In this type of evaluation, testing corresponds to a scenario test with known attackers within the test crew. Presentation attacks are intended to subvert the PAD subsystem, data capture subsystem and comparison subsystem. A successful presentation attack defeats both the PAD subsystem and the data capture subsystem, resulting in the capture of a biometric sample. Subsequently the biometric sample will be submitted for processing by the comparison subsystem.

— **Data capture subsystem and comparison subsystem** (for IUTs in which PAD results are not accessible to the evaluator). In this type of evaluation, testing corresponds to a scenario test with known attackers within the test crew. Presentation attacks are intended to subvert the data capture subsystem and comparison subsystem. A successful presentation attack will defeat the data capture subsystem, resulting in the capture of a biometric sample. Subsequently the biometric sample will be submitted for processing by the comparison subsystem.

— **PAD subsystem and comparison subsystem** (for IUTs in which a corpus of samples is evaluated in an offline mode). In this type of evaluation, testing corresponds to a technology test with samples from presentation attacks in the corpus.

— **Comparison subsystem** (for IUTs in which comparator results and PAD mechanism results are indistinguishable).

The objective of the attacker becomes critical in full-system evaluations, because the outcome of the comparison subsystem dictates whether an attack was successful. Considerations are as follows.

— **Verification systems.** In the case of an impostor/access seeker attack, failure to match (i.e. rejection of the PAI by the comparator) is typically considered a successful outcome from the perspective of the system designer.

— **Positive identification systems.** In the case of an impostor/access seeker attack, failure to return a targeted identifier (i.e. the comparator does not match the PAI against a targeted enrolment) is typically considered a successful outcome from the perspective of the system designer.

— **Negative identification systems.** In the case of an identity concealer, returning an identifier associated with the concealed identity (i.e. the comparator matches the concealed characteristic against its enrolment) is typically considered a successful outcome from the perspective of the system designer.

NOTE In a block-and-allow system, if any returned identifier triggers an investigation that uncovers the attack, this is typically considered a successful outcome from the perspective of the system designer.

8 Artefact properties

8.1 Properties of PAIs in biometric impostor attacks

In biometric impostor attacks, the attacker intends to be recognized as an individual other than themselves.

For biometric impostor attacks in which the subject intends to be recognized as a specific, targeted individual known to the system, it is necessary to create an artefact with three properties.

— Property 1: The sample appears as a natural biometric characteristic to any PAD mechanisms in place.

— Property 2: The sample appears as a natural biometric characteristic to any biometric data quality checks in place.

— Property 3: A sample acquired by a capture device from the artefact contains extractable features that match against the targeted individual's reference.

Regarding Property 1, an evaluator will not necessarily have information on the PAD mechanisms in place for a given system. Understanding the PAD mechanisms implemented is likely to motivate the use of materials capable of appearing as natural biometric characteristics.

Property 3 is related to the signal processing and comparison mechanisms within the biometric system and not generally considered part of the PAD mechanism.

NOTE 1 These issues have been discussed in References [15], [16] and [17] and can require the use of new materials.

EXAMPLE Animal proteins (see Reference [18]) can be used to defeat PADs found in fingerprint systems such as that shown in Reference [15]. If the sample does not appear as natural to the PAD, the sample can be treated temporarily to affect such an appearance.[17]

The most straightforward way to affect Property 3 is to create a copy of the targeted individual's biometric characteristic. In some cases, it is possible to produce a copy of a physical biometric characteristic in the form of an artificial biometric characteristic which can be used for a presentation attack. Alternatively, if a copy of the targeted individual's enrolled reference can be obtained, an attacker can potentially create an artefact capable of being acquired by the biometric capture device to produce a signal that can reach a match to that reference. Such artefacts can be required to pass biometric sample quality checks.

Regarding attacks by a biometric impostor, attackers can potentially acquire a test subject's biometric characteristic directly from the test subject. Such acquisition can be cooperative (e.g. the test subject provides a fingerprint to a sensor) or non-cooperative (e.g. the test subject leaves a fingerprint on a glass or a biometric capture device allowing the attacker to lift the fingerprint). Additionally, faces or voices can be recorded by attackers with a camera or microphone. Different attack scenarios are associated with cooperative and non-cooperative characteristic data capture. Artefacts created from cooperative acquisitions can be of higher quality than those from non-cooperative acquisition, which can in turn impact PAD rates and biometric performance rates.

NOTE 2 A subject can be coerced into submitting high-quality samples, in which case the cooperative/non-cooperative distinction is not applicable.

For biometric impostor attacks in which the subject intends to be recognized as any individual already known to the system, without regard to which individual, a sample acquired by a biometric capture device from the artefact should have characteristics that can match one or more stored references when processed. The most straightforward way to affect Property 3 is to have knowledge of some of the references stored in the system. Without this knowledge, it is still possible to experiment against

similar biometric systems using characteristics from the enrolled population or a general population model as a proxy. Such experiments can provide insight into the probability of successful identification against one or more enrolled references.

NOTE 3 Artefacts aiming at arbitrary subject impersonation can be referred to as “wolf artefacts”. Artefacts used during enrolment intended to achieve a high IAPAR can be referred to as 'lamb artefacts'.

If the biometric impostor intends to utilize the disguised or altered biometric characteristic multiple times, then multiple copies of the PAI should be manufacturable, or a single PAI should have a life-span sufficient for the duration of the intended use. This can impact the choice of material or production method.

8.2 Properties of PAIs in biometric concealer attacks

In the biometric concealer attack, the attacker seeks to conceal their own biometric characteristics, either using an artefact or through disguise or alteration of natural biometric characteristics.

Artefacts created for the biometric concealer attack are meant to appear as a natural biometric characteristic to any PAD mechanisms and any biometric quality checks in place. Such artefacts should contain extractable features that can be compared to stored references. In addition to Properties 1 and 2, artefacts in biometric concealer attacks should also have the following property.

— Property 4: The extractable features should not match any stored references.

Property 4 is related to the signal processing and comparison mechanisms within the biometric system and is not part of the PAD mechanism.

Artefacts unable to generate features capable of further processing by the biometric system can trigger a “failure to acquire” signal within the system, leading to additional sample acquisition attempts or triggering an “exception handling” process. Both of these outcomes are undesirable from the perspective of the attacker

NOTE 1 Artefacts aiming at achieving high failure to acquire rate (FTA) or concealer attack presentation reject rate (CAPRR) can be referred to as “goat artefacts”.

In an identification system, conformance to Property 4 is a function of the number of stored references and the identification thresholds and policies in place.

8.3 Properties of synthesized biometric samples with abnormal characteristics

If a biometric system produces unusually high false match rates when presented with certain abnormal biometric characteristics, this can necessitate specific evaluation techniques. Examples of abnormal characteristics could include those with unusually large or small numbers of features. Such characteristics are not necessarily representative of any human biometric characteristic but could be synthesized and copied to an artefact. Such a characteristic might match against a wide range of enrolees. An evaluation can seek to determine whether synthesized biometric characteristics with abnormal properties are accepted by the biometric system and can result in higher-than-normal IAPAR against bona fide enrolee references.

Evaluations of PAD mechanisms and resulting reports that examine the efficacy of synthesized biometrics samples with abnormal properties shall detail the following:

— findings for acceptance of synthesized biometric samples with abnormal properties, and

— the degree of impact on IAPAR when using synthesized biometric samples with abnormal properties.

9 Considerations in non-conformant capture attempts of biometric characteristics

9.1 Methods of presentation

Capture subjects can intentionally change their biometric characteristics or the presentation of the characteristics in an attempt to avoid recognition or to impersonate an enrolee. For biometric modes such as voice and dynamic signature, capture subjects can intentionally modify their behaviours. For biometric modes such as fingerprint, a capture subject can intentionally manipulate the presentation of their characteristic to the capture device in order to produce a non-conformant captured sample. When the test subject behaves in this way, the presentation shall be considered an attack, not a bona fide presentation, and the test subject shall be denominated as an attacker.

Artefact detection techniques are not designed to detect non-conformant bona fide presentations.

9.2 Methods of assessment

All biometric characteristics are susceptible to capture subject-induced changes caused by capture subject behaviours. To determine the sensitivity of error rates to deliberate, capture subject-induced changes in biometric characteristics or presentation, evaluators can conduct a representative test of such changes' effect on error rates such as FTA and false non-match rate. If evaluation resources and time allow, sufficient trials can also be run to determine the effect on the false match rate.

10 Artefact creation and usage in evaluations of PAD mechanisms

10.1 General

Evaluations of PAD mechanisms can be designed to answer the following questions.

— How consistently does a specific artefact subvert a biometric system?

— Which factors influence the efficacy of an artefact-based biometric system attack?

— Which attack types with the lowest attack potential succeed in subverting the biometric system?

Artefact creation, provenance, usage and handling, from creation to utilization, are central to the evaluation of PAD mechanisms.

10.2 Artefact creation and preparation

In an evaluation of PAD mechanisms, one or more PAIS will be selected. When creating and preparing artefacts according to a selected PAIS, the following factors and parameters should be considered.

— Artefact creation process: artefact creation (or fabrication) can be based on multiple materials whose production, treatment and handling can impact artefact efficacy. Artefacts are not necessarily machine-generated finished products, and human factors can impact artefact performance.

— Artefact preparation process: artefacts can require treatment or preparation between creation and utilization.

— Effort required to create and prepare artefacts: for example, skills required, technical know-how, creation time, difficulty of procuring material and equipment to be used.

— Artefact creation consistency: a "production run" of artefacts, whether comprising several artefacts created in succession or created over a long span of time, can result in artefact-over-artefact efficacy variations. This can be due to variation in materials composition, handling anomalies, or environmental factors.

— Artefact customization for a specific PAI presenter: a given artefact can be intended for use by a specific PAI presenter for whom it has been custom-designed, or whose biometric characteristics are congruous with those of the artefact.

— Artefact customization for a specific system: a given artefact can be intended for use against a specific model or class of sensor, based on an analysis of the sensor's artefact detection properties. Evaluations of artefact efficacy can be designed to assess a given artefact, artefact series, or artefact species against a specific sensor model or class.

— Biometric characteristic sourcing: artefacts can be based on direct or indirect representations of biometric samples or characteristics, on modified or manipulated biometric samples or characteristics, or on synthetic samples or biometric characteristics. The efficacy of derived artefacts can be a function of the performance of biometric samples or characteristics.

— Artefact creation and preparation cost: creation of an artefact will involve cost for sourcing the materials required and for manufacturing. A cheaper, reliable artefact that can be easily manufactured is in some cases preferred.

Evaluations of PAD mechanisms and resulting reports shall describe how artefacts were created and prepared, addressing the following:

— creation and preparation processes;

— effort required to create and prepare artefacts (e.g. technical know-how, creation time, difficulty of collecting artefact materials, creation instruments and preparation instruments);

— ability to consistently create and prepare artefacts with intended properties;

— customization of artefacts for specific PAI presenters;

— customization of artefacts for specific systems;

— sourcing of biometric characteristics;

— availability of public information on the creation and preparation process;

— changes in the artefact creation or preparation processes over the course of the evaluation.

10.3 Artefact usage

In evaluations of PAD mechanisms where artefact-based presentation attack instruments are in use, the following factors and parameters should be taken into consideration.

— Artefact presentation training and habituation: the amount of training necessary to utilize and present an artefact, and the amount of training and habituation provided to the artefact presenter, can impact artefact efficacy. Certain types of artefacts can require little presentation training and habituation, such as a replay of an audio recording. Others can require substantial training and habituation, such as presentation of an artefact to a fingerprint swipe sensor.

— Artefact presentation durability: certain types of material-based artefacts have a finite utilization lifespan, such that their efficacy decreases after one or more presentations. Conversely, an ideal artefact would be infinitely reusable. Artefacts can be characterized by differences in time and number of presentations that result in acceptance of an artefact (e.g. a silicone fingerprint PAI is a more durable artefact than a gelatine PAI). See Annex B for additional information.

— Covert use of the artefact: successful use of artefact can depend on whether the application is supervised, and if so the degree of scrutiny that might be applied during artefact usage.

Evaluations of PAD mechanisms and resulting reports shall describe how artefacts were used in the evaluation, addressing the following:

— level of PAI presenter training and habituation;

— artefact durability, including the number of presentations associated with each artefact;

— level of scrutiny or oversight applied during artefact usage; and

— changes in PAI presentation processes over the course of the evaluation.

NOTE The reported number of presentations associated with each artefact can be a range or an estimate.

10.4 Iterative testing to identity effective artefacts

Based on the creation, preparation and usage considerations above, an evaluator can evaluate presentation attack instruments with a special effort on those found to be initially effective. The analysis can take place in two phases. After a first phase of tests, the evaluator can extensively test each PAI misclassified as bona fide in a second phase of tests. APCER can then be measured for each selected PAI. If APCER exceeds a fixed threshold for one PAIS, the PAI can be deemed successful. Additionally, if a PAI does not cross the APCER threshold in the second phase of evaluation, the evaluator should still put special effort into determining if PAI effectiveness can be increased by refining the creation process or improving the presentation method.

The evaluator could report the number of tests performed in the second phase and the threshold used for APCER. A very stringent methodology would use a 0 % threshold for APCER, meaning every presentation attack which demonstrates capability to be misclassified at least two times is deemed successful, as the PAI already succeeded at least once in the first phase.

11 Process-dependent evaluation factors

11.1 Overview

Processes for enrolment, identification and verification can impact evaluation design. Evaluations of PAD mechanisms and resulting reports shall describe whether the evaluation design considered enrolment, identification, and/or verification processes, or alternatively whether the evaluation design considered a generic biometric sub-system independent of a specific process.

11.2 Evaluating the enrolment process

Biometric systems have special vulnerabilities during the process of enrolment, often necessitating the implementation of PAD mechanisms. These include:

— enrolment by a data subject of the biometric characteristics of a different individual;

— enrolment of synthetic biometric characteristics not from any individual;

— enrolment of "universal" biometric characteristics common to all or many individuals; and

— enrolment of biometric characteristics that can be altered in a consistent fashion.

Enrolment processes are often more time-consuming than identification and verification processes, involving validation of documents or other materials used to establish evidence of identity. Enrolment processes are often supervised or monitored such that the use of artefacts or non-conformant capture attempts are discoverable by an operator. Such discovery can be made through visual inspection of the capture subject or through the review of biometric data shown to the operator (e.g. on a computer screen).

EXAMPLE A test can involve personnel acting as operators who determine whether potentially suspect presentations are taking place.

Enrolment processes can also implement more rigorous biometric quality checks than identification or verification processes, increasing the likelihood that a presentation attack is detected. Lastly, enrolment processes often involve presentation of a given biometric characteristic multiple times. This

has implications for the longevity and visual plausibility of the artefact or non-conformant capture attempt.

Evaluations of PAD mechanisms and resulting reports that apply to enrolment processes shall describe the following:

— use of enrolment-specific quality thresholds or presentation policy;

— parameters of the enrolment transaction, including number and duration of presentations;

— level of operator oversight present in the process;

— manner in which operator functions were applied or emulated in the evaluation;

— whether the IUT checks sample quality and provides feedback to the test subject (e.g. "finger too wet", "move to a quieter room").

11.3 Evaluating the verification process

Verification processes are less likely to be attended than enrolment processes, with implications for artefact usage and non-conformant capture attempts. Artefacts do not necessarily require a high level of visual plausibility, and test subjects can potentially experiment with different levels of non-conformant capture attempts to induce false matches.

Evaluations of PAD mechanisms and resulting reports that apply to verification processes shall describe the following:

— use of quality thresholds and presentation policy;

— parameters of the verification transaction, including the number and duration of presentations;

— level of operator oversight present in the process;

— manner in which operator functions were applied or emulated in the evaluation;

— whether the IUT checks sample quality and provides feedback to the test subject (e.g. "finger too wet", "move to a quieter room");

— policy after failing all attempts (e.g. asking for a PIN, a password, or waiting for 30 s before attempting again);

— whether the IUT provides feedback after a failed attempt; and

— if the IUT provides feedback, a list of the feedback messages.

11.4 Evaluating the identification process

Identification processes, like enrolment processes, are sometimes supervised or monitored such that the use of artefacts or non-conformant capture attempts can potentially be discovered by an operator. However, the level of scrutiny applied to a capture subject during identification processes is likely to be less than that applied during enrolment. This can impact the level of visual plausibility that the artefact or non-conformant capture attempt needs to achieve.

An identification system can be designed to return candidates above a score threshold, though such a search will not necessarily return any candidates. Alternatively, an identification system can return the strongest candidate regardless of comparison score. The latter type of identification system requires higher degrees of non-conformance to induce a false negative identification.

Evaluations of PAD mechanisms and resulting reports that apply to identification processes shall describe the following:

— use of quality thresholds and presentation policy;

— parameters of the identification transaction, including the number and duration of presentations;

— configuration of system to perform negative or positive identification;

— whether test subjects were enrolled in the databases against which identification took place;

— level of operator oversight present in the process;

— whether and how an operator adjudicates candidate identities returned by the system; and

— manner in which operator functions were applied or emulated in the evaluation.

11.5 Evaluating offline PAD mechanisms

Some outcomes of PAD mechanisms can potentially not occur immediately after presentation, but offline at a later time. This can be necessary for a number of reasons, as follows.

— PAD mechanisms can be time-consuming, meaning that the real-time processing of results is not feasible. Results could occur hours or days after the biometric presentation.

— Newer or different PAD mechanisms can be run across previously captured biometric samples.

— Subsequent events can suggest or confirm that a presentation attack has occurred. This can require evidence in the form of original biometric sample(s) to be retained for forensic analysis to detect and confirm PAD mechanism results and/or for court purposes.

Evaluation of offline PAD mechanisms might benefit from PAD mechanism data produced during a presentation and retained; ISO/IEC 30107-2 establishes requirements on such data.

Reports that evaluate offline PAD mechanisms shall describe their implementation in the overall processing scheme.

12 Evaluation using Common Criteria framework

12.1 General

The Common Criteria Parts 1-3 (ISO/IEC 15408-1, ISO/IEC 15408-2 and ISO/IEC 15408-3) and the Common Evaluation Methodology (ISO/IEC 18045) are relevant to the independent security evaluation of IT products. The independent evaluation and certification of IT products according to these International Standards is widely used in many different areas.

The Common Criteria is defined in three parts:

— Part 1 contains the “Introduction and general model”;

— Part 2 contains the “Security functional components”;

— Part 3 contains the “Security assurance components”.

PAD evaluations based on the Common Criteria shall follow the methodologies given in ISO/IEC 15408-1, ISO/IEC 15408-2 and ISO/IEC 15408-3, and the Common Evaluation Methodology (ISO/IEC 18045).

The Common Evaluation Methodology (CEM)[1] is a companion document to the Common Criteria and defines the minimum actions to be performed by an evaluator in order to conduct a Common Criteria evaluation, using the criteria and evaluation evidence defined in the Common Criteria.

Within the Common Criteria, the target of evaluation (TOE) is the IT product that is the subject of the evaluation. This corresponds to the IUT as referred to in this document. The TOE is characterized through the security target, a document that identifies the security functional requirements and security assurance requirements and can refer to one or more protection profiles. A protection profile is used to describe a class of IT products that share a certain scope and can be used to solve a certain

security problem. A security target, on the other hand, describes the security characteristics of a concrete IT product and how it fulfils all security requirements.

Security functional components as defined in ISO/IEC 15408-2 are the basis for the security functional requirements expressed in a Protection Profile or Security Target. A Protection Profile or Security Target contains a set of security functional requirements to describe the security functionality of the TOE in a semi-formal language. The fact that the security functionality of a TOE is not only described in natural language but also in a semi-formal language serves to make different evaluations comparable.

The security assurance components determine the level of depth during evaluation. Every security assurance component from ISO/IEC 15408-3 stands for one task of the evaluator during the evaluation. The seven predefined evaluation assurance levels (EAL1–EAL7) correspond to increasing efforts for design verification and testing as shown in Table 1.

Table 1 — EALs and their description

Evaluation Assurance Level (EAL)	Depth of evaluation
EAL1	Functionally tested
EAL2	Structurally tested
EAL3	Methodically tested and checked
EAL4	Methodically designed, tested, and reviewed
EAL5	Semi-formally designed and tested
EAL6	Semi-formally verified design and tested
EAL7	Formally verified design and tested

Each EAL includes a vulnerability assessment. A higher EAL reflects a more rigorous vulnerability assessment and a higher attack potential to be performed in penetration testing. Attack potential is a measure of the effort expended in the preparation and execution of the attack. The Common Evaluation Methodology (ISO/IEC 18045) gives general guidance on calculating attack potential as a function of required time, expertise, knowledge of the TOE, window of opportunity and equipment.

A Protection Profile or Security Target includes the set of security assurance components predefined for an EAL, possibly augmented by additional assurance components.

The Common Criteria Recognition Arrangement (CCRA) has international certificate authorizing members and is further described in Reference [28]. Protection Profiles for biometric systems are also listed in Reference [28].

The Common Criteria framework is a pure security evaluation standard. In principle, the Common Criteria only focus on the question whether an IT product provides the security functionality required for a certain use case/environment and whether sufficient trust can be laid into the implementation of this security functionality.

12.2 Common Criteria and biometrics

12.2.1 Overview

Biometric systems can be evaluated according to the Common Criteria as any other IT product. Biometric systems have certain characteristics that need special consideration during an evaluation, including the following.

— **Biometric performance error rates:** Biometric authentication does not work as deterministically as other means for authentication or identification of users. Some biometric performance error rates (e.g. according to ISO/IEC 19795-1) have an impact on the security of the system and need to be considered during a security evaluation.

— **PAD:** It is well known that some biometric systems (e.g. PAD subsystem, data capture subsystem, or full system) can be vulnerable against presentation attacks. The evaluation of the capability to

detect and defeat these attacks can belong in the scope of a Common Criteria evaluation depending on the use case of the TOE.

— **Vulnerability assessment:** Biometric systems in general can be subject to special kind of attacks (such as hill climbing) that will need consideration during a security evaluation.

For these areas, special guidance is required in order to facilitate a comparable evaluation in all laboratories of the Common Criteria schema worldwide. Special characteristics of biometrics in Common Criteria evaluations are dealt with in the form of guidance for the evaluator performing an evaluation and the developer of a biometric system. The ISO/IEC 19989 series provides such guidance, but the most important aspects are summarized in 12.2.2-12.2.5. Other approaches to security evaluation of biometrics are given in ISO/IEC 19792.

12.2.2 General evaluation aspects

The Common Criteria poses requirements on a wide variety of aspects of the TOE, starting from the development (including the development environment) up to the delivery of the TOE to the customer. Most aspects can be applied to biometric systems as to any other IT product. However, in some areas, specific guidance is given to the evaluator on how to evaluate these aspects. For example, the description of the design of a biometric system refers to specific aspects of the technology.

12.2.3 Error rates in testing

When it comes to testing a biometric system in the context of a Common Criteria evaluation, the security-relevant error rates are a very important aspect of the functionality to be considered. According to the guidelines, the evaluator will perform the following steps.

— **Identify the relevant test approach:** Various test approaches are available starting from a database-based technology test of a biometric algorithm to an evaluation of the performance of the biometric system under operation. The correct test approach highly depends on the definition of the TOE.

— **Identify the security-relevant error rates:** As the Common Criteria focuses on the security-relevant error rates only, not all error rates of the biometric system are relevant. The identification of the security-relevant error rates is performed based on the type of the biometric system and its use case as defined in the Security Target.

— **Plan the execution of the test:** The actual test execution has to be planned and described within the test documentation in advance.

— **Estimate test size:** Collecting test data represents a significant amount of the effort of the overall test. It is essential to develop an idea about the amount of test data that is required before starting the actual process of test data acquisition.

— **Document the test plan:** It is essential to plan the required documentation for the test in advance of the test itself.

— **Acquire test crew:** For the quality of results, it is essential that the evaluator utilizes a test crew not known to the developer of the system beforehand.

— **Perform test:** The test is carried out under the sole control and responsibility of the evaluator.

— **Evaluate test results:** After testing, results will be evaluated and reported according to defined metrics.

12.2.4 PAD evaluation

The Common Criteria itself does not require a biometric system under evaluation to provide PAD mechanisms. The requirement for PAD mechanisms is dependent on the intended environment of the biometric system.

For example, a border control system under the strong and constant control of a border control officer does not necessarily require PAD, while an ATM that uses biometrics as the only means for authentication is more likely to require PAD. The guidelines for the evaluation of biometrics, however, specify that PAD mechanisms, if existing, belong to the security functionality of the system and therefore are to be evaluated. In other words, it is not possible to evaluate a biometric system according to Common Criteria without consideration of its PAD functionality.

PAD mechanisms can be viewed from two perspectives.

— PAD mechanisms belong to the security functionality of the biometric system and are functionally tested. Guidelines direct the evaluator on how to plan, conduct, document and evaluate such a functional test.

— PAD mechanisms also fall into the area of vulnerability assessment, as the use of a PAI against the biometric system is an attempt to circumvent the security functionality of the TOE.

The differences between the two perspectives can best be visualized using a concrete example. In the area of functional testing, the evaluators' concern regarding PAD is to verify that the TOE meets certain detection performance requirements. The PAD mechanism has to perform within a certain range of detection performance. Testing can be achieved by the use of a standardized toolbox. Beside some dedicated requirements on testing and documentation, this situation is very close to the situation in biometric recognition performance testing. Having passed the test from a functional perspective is a prerequisite to starting the vulnerability assessment. If the PAD mechanisms would not work within sufficient detection performance limitations, any kind of vulnerability assessment would be useless. In the vulnerability assessment, the evaluator will then try to circumvent the PAD mechanism, working within the limitations of the attack potential of the current evaluation. This can lead to a situation in which a TOE passes the functional test but where the evaluator can build a so-called "golden fake" that reproducibly breaches the security functionality of the TOE. If this happens, the TOE fails the security evaluation even though it showed good detection performance during functional testing.

As a basic rule it can be said that one repeatably successful attack against a TOE (always under consideration of the maximum attack potential) will make the security evaluation fail. This is one of the major differences between a security evaluation compared to a pure performance test.

12.2.5 Vulnerability assessment

12.2.5.1 Typical attack scenarios

Specific kinds of attacks against biometric systems exist. Presentation attacks are only one very prominent example. A biometric system can also be vulnerable against a hill-climbing attack, for example.

It is important that the evaluator considers typical and well-known presentation attacks during the evaluation of a biometric system. While the system is not necessarily vulnerable to all attacks, as a starting point for a vulnerability analysis it is important that all typical attacks are considered. These can be seen as a minimum list of attacks to be considered. They do not claim to be complete and the evaluator will in any case develop additional attack scenarios during evaluation.

12.2.5.2 Rating attacks

Guidance for the security evaluation of biometric systems introduces a dedicated scheme to rate the attack potential of attacks against biometric systems as minimal, basic, enhanced-basic, moderate, high or beyond high. The level chosen for the vulnerability analysis is one of the most important aspects of the chosen EAL. This decision basically answers the question against which attack potential a TOE is expected to be resistant.

The evaluator will perform their vulnerability assessment and penetration testing only up to the chosen level. Common Criteria uses a dedicated list of criteria to classify an attack in general. To reflect the dedicated characteristics of attacks against biometric systems, an extension and interpretation of

the standard attack rating scheme has been proposed by the European BEAT (Biometric Evaluation and Testing) project.[5] This scheme uses the characteristics of elapsed time, expertise, knowledge about the TOE, access to the TOE/window of opportunity, access to the biometric characteristic and success rate. The scheme utilizes a system of points to establish a numerical value for each attack. It also distinguishes between effort required to prepare/identify an attack and to exploit the attack. Such dedicated schemes for rating attacks have been proposed for other technical areas, namely smart cards and similar devices, in the past and are well accepted in the Common Criteria community.

EXAMPLE The FIDO Alliance applies the BEAT framework[5] in the context of mobile device evaluation.[6] Knowledge of the TOE and access to the TOE are assumed. PAI factors including time, expertise, equipment and biometric characteristic source are considered across three levels.

12.2.5.3 Previous approaches in fingerprint PAD protection profiles

Many aspects of the methodology outlined above have their origin in an approach developed by the German Federal Office for Information Security (BSI). The BSI developed two dedicated Fingerprint Spoof Detection Protection Profiles [7][8] in order to describe the security characteristics of a biometric system with PAD mechanisms. Both protection profiles define the identical set of security functional requirements that are required be met by a TOE claimed to be conformant to the protection profile, namely:

— PAD (i.e. spoof detection),

— audit for security relevant events,

— protection of residual information, and

— management of security functions.

Along with the protection profiles, a guideline has been developed[9] that provides guidance for the evaluator on how to evaluate PAD mechanisms. In the meantime, this guidance has been used as the foundation of the ISO/IEC 19989 series.

13 Metrics for the evaluation of biometric systems with PAD mechanisms

13.1 General

PAD mechanism performance can be expressed in terms of classification error rates, non-response rates, and other rate-based metrics. Such metrics could be utilized in security evaluations, academic evaluations, systematic technology or product development processes, or as quick-look benchmarks by an end user. Clause 13 provides metrics used in such tests. ISO/IEC 19795-1 provides an overview of the reporting requirements for a biometric performance test for bona fide presentations.

Evaluations of PAD mechanisms shall report the following:

— number of presentation attack instruments, PAIS and PAI series used in the evaluation;

— number of individuals involved in the testing, including PAI presenters unable to utilize artefacts and test subjects unable to present non-conformant characteristics;

— number of PAI sources from whom artefact characteristics were derived;

— number of presentation attack instruments created per PAI source for each PAIS;

— number of tested materials;

— description of output information available from PAD mechanism;

— ordering of subject presentations with and without PAI, and whether PAI presenters or test subjects were reused;

— ordering of presentations to PAD-enabled and disabled system;

— whether test subjects were reused.

To account for the full range of distinct and potentially overlapping roles in a PAD test, the experimenter shall, in the test report:

— define the purpose and responsibilities of the following roles in a PAD test:

  — test subject (conducts bona fide presentations and non-conformant capture attempts),

  — PAI presenter,

  — PAI source,

  — PAI creator;

— state whether the role was material to test results and provide a basis for this assertion;

— indicate the number of individuals who occupied each role (e.g. five individuals were PAD sources in the test);

— for each role, describe individuals' level of experience with presentation attacks; and

— document occurrences in which individuals occupied multiple roles, e.g. PAI sources were also PAI presenters.

In certain tests, it is necessary to enrol the PAI source as a bona fide reference.

For a full-system evaluation of impostor attacks, the PAI presenter shall not conduct presentations in which they are enrolled as a bona fide reference. This could result in PAI presenter's bona fide characteristics being compared beneath the PAI, undermining test results.

For a full-system evaluation of concealer attacks, the PAI presenter shall be enrolled. This is necessary in order to determine if the PAI presenter's biometric characteristic is correctly concealed.

Test reports shall describe any use of machines or automated mechanisms as PAI presenters or PAI sources.

NOTE Performance metrics discussed in Clause 13 can fail to achieve statistical significance due to limitations in sample size.

13.2 Metrics for PAD subsystem evaluation

13.2.1 General

PAD subsystem evaluations (see ISO/IEC 30107-1:2016, Figure 4) measure the ability of PAD subsystems to correctly classify presentation attacks.

13.2.2 Classification metrics

Both APCER and BPCER are reported in PAD subsystem evaluations.

In PAD subsystem evaluations, performance metrics for presentation attacks shall be calculated and reported as APCER. The evaluator shall report on the manner in which PAD decisions and scores were used to classify presentations.

The APCER for a given PAIS shall be calculated using Formula (1):

$$\mathrm{APCER}_{\mathrm{PAIS}} = 1 - \left(\frac{1}{N_{\mathrm{PAIS}}} \right) \sum_{i=1}^{N_{\mathrm{PAIS}}} Res_i \tag{1}$$

where

N_{PAIS} is the number of attack presentations for the given PAI species;

Res_i takes value 1 if the ith presentation is classified as an attack presentation, and value 0 if classified as a bona fide presentation.

Evaluations of PAD mechanisms shall report the number of artefact presentations correctly and incorrectly classified: total, by PAIS, by PAI series, by test subject and by source.

When considering how well a PAD subsystem performs in detecting the PAIS of a specified attack potential (AP), the APCER of the most successful PAIS within this attack potential should be used as shown in Formula (2):

$$\mathrm{APCER}_{\mathrm{AP}} = \max_{\mathrm{PAIS} \in A_{\mathrm{AP}}} \left(\mathrm{APCER}_{\mathrm{PAIS}} \right) \tag{2}$$

where A_{AP} is a subset of PAI species with attack potential at or below AP.

Attack potential should be calculated based on the ISO/IEC 19989 series.

The max-based formula reflects the vulnerability of a PAD system to at least one attack at a tested attack potential level. This is a good assumption for applications when measuring PAD error rates for the purpose of making security decisions, where the expected attacker would attack the system using the PAI most likely to be effective (within their means). In addition, attackers can use PAIs that were not tested, and using the max error rate observed among a suite of tested PAIs is a more reliable security metric. In operations, this APCER statement would apply if attackers had the same attack potential as deployed by the test laboratory's experimenters. In operational cases, where attackers had less knowledge than the test laboratory and selected presentation attack instruments randomly, or on basis of ease of production, a lower APCER rate would be achieved than given in Formulae (1) and (2).

At the PAD subsystem level, performance metrics for the set of bona fide presentations captured with the evaluation target shall be calculated and reported as BPCER. BPCER shall be calculated as shown in Formula (3):

$$\mathrm{BPCER} = \frac{\sum_{i=1}^{N_{\mathrm{BF}}} Res_i}{N_{\mathrm{BF}}} \tag{3}$$

where:

N_{BF} is the number of bona fide presentations;

Res_i takes value 1 if the ith presentation is classified as an attack presentation, and value 0 if classified as a bona fide presentation.

Evaluations of PAD mechanisms shall report the number of bona fide presentations correctly and incorrectly classified: the total and by test subject.

If the PAD subsystem returns a multi-valued PAD score, the frequency distributions of the PAD scores should be reported for each PAIS and for bona fide presentations.

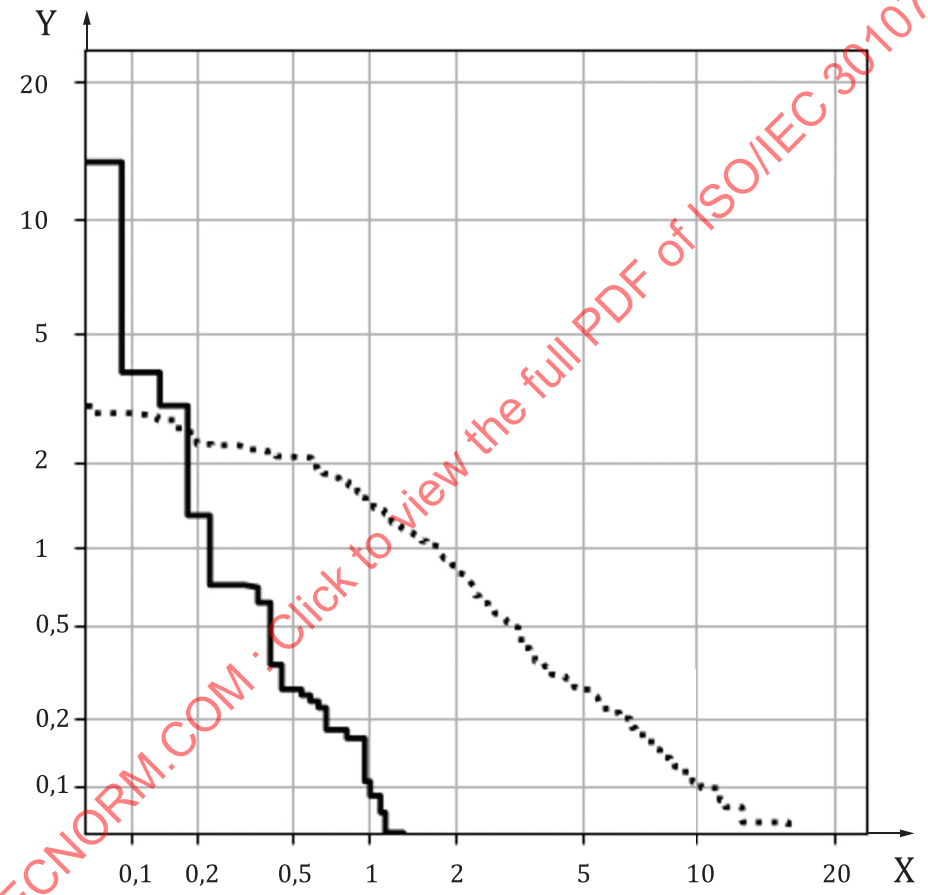
Reporting the aggregate of APCER and BPCER (e.g. half-total error rate) is not conformant with this document.

The classification performance of a PAD mechanism may be reported in a single figure as the BPCER at a fixed APCER.

EXAMPLE When $APCER_{AP}$ is 5 %, BPCER can be reported as $BPCER_{20}$.

When interpreting the performance of a PAD subsystem, it is important to recognize that there can be presentation attacks types, PAIS and other factors which have not been tested. Therefore, the reported performance of a PAD subsystem does not provide any information regarding its effectiveness in detecting presentation attacks which have not been tested.

The performance of a PAD subsystem over a range of decision thresholds may be graphically represented using a DET plot. DET plots are threshold-independent, allowing detection performance comparison of different systems under similar conditions, or of a single system under differing conditions. The DET plot shall indicate the security-related metric (APCER) on the horizontal axis and the convenience-related metric (BPCER) on the vertical axis as illustrated in Figure 1. The ideal combination of low APCER and low BPCER occurs at the bottom left corner of the plot.



Key

X	APCER (%) – security metric
Y	BPCER (%) – convenience metric
.....	PAD algorithm 1
———	PAD algorithm 2

Figure 1 — PAD performance shown on a DET plot

13.2.3 Non-response metrics

Taking into account supplier recommendations and the intended use case scenario for the PAD subsystem, the evaluator shall define what constitutes a non-response and specify conditions under which a non-response contributes to the classification error rate.

EXAMPLE An evaluator might define a non-response as no appearance of a biometric image for 5 s after presentation of a biometric characteristic or PAI.

The evaluator shall report non-response rates for the PAD subsystem using the following metrics:

— for each PAIS, attack presentation non-response rate (APNRR) and the sample size on which the computed rate is based;

— bona fide presentation non-response rate (BPNRR) and the sample size on which the computed rate is based.

NOTE A high APNRR value can be seen as a positive outcome from the perspective of the system designer. A PAD subsystem that responds to an attack presentation could give an attacker insight into its techniques, such that a non-response can be preferred.

13.2.4 Efficiency metrics

Time-sensitive applications can be adversely affected by increased transaction time. The evaluator should report PAD subsystem processing duration (PS-PD) as mean duration. PS-PD should be reported separately for attack presentations and bona fide presentations. Non-responses are not included when calculating PS-PD. PS-PD may be determined by direct observation. Alternatively, the average processing duration change due to the PAD subsystem may be estimated by recording a number of presentations with and without PAD enabled and analysing the differences in processing durations.

13.2.5 Summary

Table 2 lists performance metrics for PAD subsystem evaluation.

Table 2 — PAD subsystem performance metrics

Subsystem	Metric	Type of presentation	Reporting
PAD subsystem	APCER	Attack	Mandatory
	BPCER	Bona fide	Mandatory
	APNRR	Attack	Mandatory
	BPNRR	Bona fide	Mandatory
	$APCER_{AP}$	Attack	Optional
	PS-PD	Bona fide or attack	Optional

For PAD subsystems that return a multi-valued PAD score, PAD score frequency distributions are recommended for each PAIS and for bona fide presentations.

13.3 Metrics for data capture subsystem evaluation

13.3.1 General

Data capture subsystem evaluations measure the ability of the subsystem to correctly reject presentation attacks and to correctly acquire bona fide presentations. In this context, correctly reject means to not acquire.

13.3.2 Acquisition metrics

Data capture subsystem acquisition performance for presentation attacks is calculated and reported as the attack presentation acquisition rate (APAR). APAR is the proportion of attack presentations using the same PAIS from which the data capture subsystem acquires a biometric sample of sufficient quality.

The evaluator shall report acquisition rates of the data capture subsystem as follows:

— for each PAIS, the attack presentation acquisition rate (APAR);

— for bona fide test subjects erroneously not acquired by the data capture subsystem, the FTA or FTE as defined in ISO/IEC 19795-1; and

— the sample sizes on which the above computed rates are based.

FTE is reported for evaluations with an enrolment component. FTA is reported for evaluations with a recognition component.

Taking into account supplier recommendations and the intended use-case scenario for the device, the evaluator shall specify conditions under which a non-response contributes to APAR, FTE and FTA.

13.3.3 Non-response metrics

The evaluator shall report non-response rates of the data capture subsystem as follows:

— for each PAIS, the APNRR;

— for bona fide test subjects, the BPNRR; and

— the sample sizes on which the above computed rates are based.

Taking into account supplier recommendations and the intended use-case scenario for the device, the evaluator shall define what constitutes a non-response.

13.3.4 Efficiency metrics

The evaluator should report the data capture subsystem processing duration (DCS-PD) as mean duration. Data capture subsystem processing duration should be reported separately for attack presentations and bona fide presentations. Non-responses are not included when calculating DCS-PD.

NOTE Statistical evaluation can provide zero-normalized duration scores for each subject as well as for over the whole test crew population.

13.3.5 Summary

Table 3 lists performance metrics for data capture subsystem evaluation.

Table 3 — Data capture subsystem performance metrics

Subsystem	Metric	Type of presentation	Reporting
Data capture subsystem	APAR	Attack	Mandatory
	FTE	Bona fide	Mandatory for enrolment
	FTA	Bona fide	Mandatory for recognition
	APNRR	Attack	Mandatory
	BPNRR	Bona fide	Mandatory
	DCS-PD	Bona fide or attack	Optional

13.4 Metrics for full system evaluation

13.4.1 General

Full system evaluations shall report comparison subsystem results in addition to PAD or data capture subsystem results.

NOTE Depending on the IUT, PAD or data capture subsystem results can be unavailable.

13.4.2 Accuracy metrics

13.4.2.1 Evaluation of verification systems

For verification systems, for each PAIS, at least one of the following shall be reported:

— IAPAR and the sample size on which this computed rate is based;

— CAPRR and the sample size on which this computed rate is based.

If the evaluation includes both biometric impostors and biometric concealers, then both IAPAR and CAPRR shall be reported.

NOTE To defeat recognition, biometric concealers desire a high CAPRR as well as a high APCER. To be falsely recognized, biometric impostors desire a high IAPAR as well as a high APCER.

For a given IUT, the IAPAR of the most successful PAIS with attack potential AP may be reported as $IAPAR_{AP}$.

Comparison subsystem results shall be reported as FAR/FRR, including calculations and the basis of the results obtained.

13.4.2.2 Evaluation of positive identification systems

For positive identification systems, for each PAIS, impostor attack presentation identification rate (IAPIR) and the sample size on which the computed rate is based shall be reported.

NOTE To be falsely recognized, biometric impostors desire a high IAPIR as well as a high APCER.

Comparison subsystem results shall be reported as FPIR, including calculations and the basis of the results obtained.

13.4.2.3 Evaluation of negative identification systems

For negative identification systems, for each PAIS, concealer attack presentation non-identification rate (CAPNIR) and the sample size on which the computed rate is based shall be reported.

NOTE To defeat recognition, biometric concealers desire a high CAPNIR as well as a high APCER.

Comparison subsystem results shall be reported as FNIR, including calculations and the basis of the results obtained.

13.4.3 Efficiency metrics

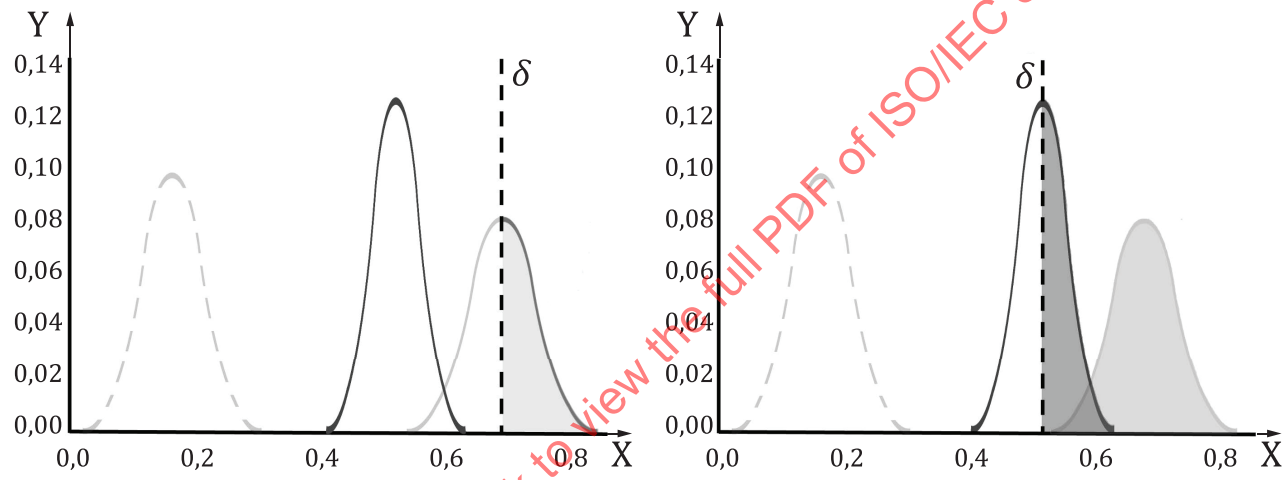
The evaluator should report the full-system processing duration (FS-PD). Increases in FS-PD due to PAD can be important in high throughput and other time-sensitive applications. The time required for PAD characteristics to be processed by the signal processing subsystem can be different than for bona fide biometric characteristics. FS-PD accounts for changes in signal processing durations due to PAD mechanisms along with durations accumulated across all other subsystems.

FS-PD with PAD mechanisms enabled and disabled should also be reported. In some cases, FS-PD can be determined by direct observation. In other cases, an aggregate average processing duration increase due to PAD mechanisms can be estimated by recording a number of transactions with and without PAD mechanisms enabled and analysing the differences in processing durations.

13.4.4 Generalized full-system evaluation performance

The IAPAR is directly dependent on the decision threshold value $\hat{o}$ of the comparison subsystem. For a presentation attack to be meaningful in a biometric impostor use case, its recognition rates need to be higher than those of a zero-effort attack, as discussed in 8.1. In such cases, the distribution of comparison scores from PAIs will overlap with mated comparison scores to some significant degree. This is in contrast to random non-mated comparison scores, which typically do not have significant overlap with mated comparison scores. Comparison scores are numerical values that result from a comparison.

IAPAR can potentially be as low as 0 % for a system with a poorly-adjusted decision threshold as shown in Figure 2 a), where it would be much higher for a system with a properly-adjusted decision threshold as shown in Figure 2 b).



a) Decision threshold with lower IAPAR but poor comparison performance

b) Decision threshold with higher IAPAR and acceptable comparison performance

Key

X	comparison scores
Y	probability density
δ	decision threshold
———	bona fide mated
– – –	bona fide non-mated
———	attacks

Figure 2 — Decision thresholds with lower/higher IAPAR and poor/acceptable comparison performance

Figure 2 a) shows a poor recognition configuration with IAPAR of 0 % and FRR of 65 %. Figure 2 b) shows a recognition configuration with an IAPAR of 41 % and a more acceptable FRR of 1 %.

In order to achieve a generalized metric with respect to the FRR of the system, the experimenter shall report the sum of IAPAR and FRR. This can be expressed as the relative impostor attack presentation accept rate (RIAPAR), expressed as shown in Formula (4):

$$a(\tau) = b(\tau) + c(\tau) \tag{4}$$

where

a is the RIAPAR;

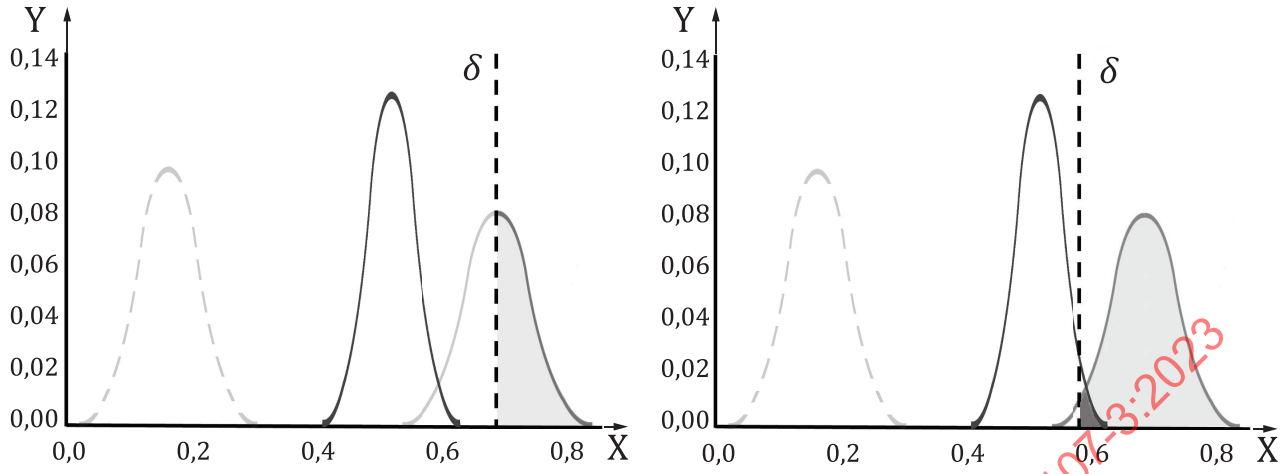
b is the IAPAR;

c is the FRR

(τ) is the threshold.

In contrast to IAPAR, RIAPAR also takes into account the FRR of the biometric recognition system. RIAPAR reflects both the vulnerability and the convenience of a system. The resulting RIAPAR is shown as 65 % in Figure 2 a) and 42 % in Figure 2 b).

The decision threshold can be optimized to improve RIAPAR. Figure 3 a) shows a decision threshold a with robust IAPAR (0 %), but which significantly increases the FRR at the cost of unacceptable RIAPAR. Figure 3 b) shows an optimized configuration with a threshold that results in an IAPAR of 3 % and an FRR of 4 %. The RIAPAR for this configuration 7 %.



a) Decision threshold with suboptimal RIAPAR

b) Decision threshold with optimized RIAPAR

Key

X	comparison scores
Y	probability density
δ	decision threshold
———	bona fide mated
– – –	bona fide non-mated
———	attacks

Figure 3 — Decision threshold with suboptimal and optimized RIAPER

RIAPAR applies to full system performance, irrespective of whether PAD is in operation.

13.4.5 Summary

Table 4 lists performance metrics for full-system evaluation.

Table 4 — Full-system performance metrics

Subsystem (recognition type)	Metric	Type of presentation	Reporting
Comparison subsystem (verification)	FAR / FRR	Bona fide	Mandatory
	IAPAR	Attack	Mandatory for biometric impostors
	RIAPAR	Attack and bone fide	Mandatory for biometric impostors
	CAPRR	Attack	Mandatory for biometric concealers
	$IAPAR_{AP}$	Attack	Optional
	FS-PD	Attack or bona fide	Optional
Comparison subsystem (positive identification, applicable to biometric impostors)	FPIR	Bona fide	Mandatory
	IAPIR	Attack	Mandatory
	FS-PD	Attack or bona fide	Optional