STANDARDIZED VERSUS PRACTICE

ISO/IEC 19795-2 COMPARED TO CASES

DRS. IR. F.F. DE WIT

5-12-2024

UNIVERSITY OF TWENTE.

COLOPHON

MANAGEMENT
Electrical Engineering, Mathematics & Computer Science (EEMCS), Data Management & Biometrics

DATE
5-12-2024

REFERENCE
Reference

VERSION
1.0

STATUS
Working Document

AUTHOR(S)
Drs. Ir. F.F. de Wit

EMAIL
f.f.dewit@utwente.nl

POSTAL ADDRESS
Zilverling 4057
P.O. Box 217
7500 AE Enschede

WEBSITE
www.utwente.nl

FILENAME
2024-12 Analysis 19795-2 versus cases (Working Document) V1

REGISTRATION DETAILS
Registration details

TABLE OF CONTENTS

Summary 3
1. Introduction 4
1.1 Motivation 4
1.2 Approach 4
2. Analysis results & observations 5
2.1 Test design 5
2.2 Assembling an appropriate test corpus 5
2.3 Performance measurement 5
2.4 Reporting 6
2.5 Other issues 6
3. Conclusions 7
Appendix 9

SUMMARY

Researchers from the Data Management and Biometrics group (DMB) have participated in evaluations of biometric performance for years. Recently we started a line of research into a practical framework for biometric performance evaluation.

The series of standards crafted by the international standards organisation ISO starting with ISO/IEC 19795-1 [1] intends to support systematic biometric performance testing and reporting; we analysed the 19795-2 standard by comparing the requirements and guidelines therein to our approach with several evaluation projects with two questions in mind: 1) *does/do our approach(es) comply with the relevant clauses given in ISO/IEC 19795-2?*, and 2) *does the standard provide us with a framework or recipe that we can execute?*

The analysis results show that overall we usually comply with the standard's relevant clauses; where we have diverged from the standards we usually have done so with a reasoned motivation. The main finding contributing to our research on a practical framework is that the standard does not provide a framework or recipe for biometric performance evaluation that can be executed directly.

1. INTRODUCTION

1.1 MOTIVATION

Several members of the Data Management & Biometrics group at the University of Twente have been engaged in evaluating the performance of biometric technology for a considerable time. To improve our practice and learn from our experience, we have started doing research into a practical framework for biometric performance evaluation. Part of that research is to evaluate existing approaches on their applicability and use for our evaluation practice.

One of the existing approaches to biometric performance evaluation is the series of standards crafted by the international standards organisation ISO starting with ISO/IEC 19795-1 [1]. Since our main efforts are on so called technology evaluation, we first focussed on the standard on that particular type of evaluations: ISO/IEC 19795-2 [2].

With this report we share the results of our analysis of the latter standard with respect to three major technology evaluation projects we have undertaken, i.e.:

1. *AFIS 2018*: performance evaluation of automated fingerprint identification algorithms
2. *Seamless flow*: performance evaluation of face recognition for identity verification and identification function for a seamless border concept
3. *NoQ3*: performance evaluation of face recognition for identity verification as part of a large scale automated border control (ABC) gate system

These three cases are described in more detail in [3].

To claim compliance with ISO/IEC 19795-2 [2] one needs to comply with the clauses specified in said standard. It goes without saying that we are both interested in the degree to which aforementioned projects comply with this standard. On the other hand, we would also like to know whether the standard provides a practical framework – one could say a recipe – that can be easily applied to set up and execute biometric technology performance tests.

1.2 APPROACH

First, we have made a table of all relevant clauses from [2]. With each clause we reflected on our approach with each of the three projects, and noted the observations in the table. This analysis then gives insight both into what degree the projects were compliant with the standard and the way we could address issues of non-compliance.
On the other hand, the analysis should also tell us whether strict adherence to the standard would help us focus on what is important in the performance evaluation projects we usually take part in, including whether it provides a framework or recipe for these efforts.

Section 2 discusses the analysis results and further observations, while section 3 draws the conclusions.

2. ANALYSIS RESULTS & OBSERVATIONS

The full analysis table can be found in the Appendix.

In this section we will go through the table and consider the results and observations for *test design* ([2] clause 6.1), *assembling an appropriate test corpus* (clause 6.2), *performance measurement* (claus 6.3), *reporting* (clause 6.4) and *other issues applicable to technology and scenario evaluations* (clause 8).

2.1 TEST DESIGN

There are no serious issues or failure to comply when considering test design for the three aforementioned projects. The main issue is that the standard requires most parameters of the design to be explicitly specified or reported; this is not our common practice up until now, but might contribute to more clearly documenting what our objectives and approach are and communicating those to other parties involved.

A *major observation* here is that the standard gives a generic description of a technology test execution and refers to some more generic execution sequences in its annex A.
These sequences are so generic, however, that they do not provide an actual practical “recipe”; they make sense but provide no guidance on how to apply the clauses, or how to execute the sequence.

2.2 ASSEMBLING AN APPROPRIATE TEST CORPUS

With respect to construction of a test corpus (i.e. collection of samples) we run into some difficulties. The three projects all used what we have come to call “operational” data. This is data,

1. selected from an existing collection of samples that is or has been used for the purpose of the system’s function (e.g. a database of fingerprints and fingermarks)
2. Collected through the system in use, from actual users (e.g. live images and passport pictures collected from travelers using an ABC gate system)

Such data has the advantage of having realistic properties, but said properties are almost irrevocably linked to the existing system. This may introduce a bias into the dataset; the data included in the corpus could be processed by the existing system after all, and might be easier to process by similar algorithms.

A more fundamental issue with operational data is the lesser measure of control one has when data is acquired with an operational system in realistic conditions compounded by the reduced access one has to information if the collection has to be done by external parties. This adds a layer of complexity which the standard does not consider.

On the other hand, the standard does highlight that we could be more systematic in documenting properties of the corpus and its construction, including collecting data on failures unrelated to the actual biometric recognition algorithms (which may introduce the aforementioned bias).

2.3 PERFORMANCE MEASUREMENT

There are quite a few non-compliance items and caveats for the evaluation projects with respect to the performance measurement clauses of the standard. This reinforces the impression that we could be more consistent and systematic in logging data on failures during score calculation as well.

However, there are some calculations that the standard prescribes which are useful when fully documenting a biometric performance test but may not be of use for interested parties. It is likely this is going to be a tradeoff between principled rigor and practical considerations.

2.4 REPORTING

As with performance measurement, the clauses on reporting highlight that we could use some help in being a little more thorough but we may have to strike a balance between reporting every detail and just enough for the interested parties.
Since these clauses also refer back to sections 6.1 through 6.3 this applies to the issues there as well.

2.5 OTHER ISSUES

The main “other issues” lie within the interaction with stakeholders.

In one case technology suppliers were asked to supply their own hardware to run the tests; this is not compliant with the standards, but might be justified given sufficient oversight by the testing organisation.

It is common for a commissioner to simply request some selection of algorithms to be tested. The standards urge us to try and give the criteria that were used to select the algorithms.

3. CONCLUSIONS

We set out to analyse standard ISO/IEC 19795-2 and compare it to our approach in three biometric performance evaluation projects of a reasonably large scale. Our two objectives were, 1) to assess where we could improve our practice using the standard, and 2) whether the standard offers a framework or “recipe” to set up a biometric performance test.

We conclude that we can indeed improve our practice by more systematically applying the clauses from the standard.

We also conclude that the standard, does not offer a way to design and execute a test.

APPENDIX

Part[1]	Clause	Item	Evaluation description	
2	5	*Only describes what the relative differences between technology and scenario evaluation are. No actual conformance items.*		
2	6			
	6.1	Test design		
	6.1.1	Goals: Evaluation is designed to evaluate a system's enrolment, acquisition and matching functions	Not all evaluations are designed to evaluate all these subsystems; technology evaluation often aims at just one of them, most often matching.	
		Evaluation is designed to evaluate a system's enrolment, acquisition or matching functions, or a combination thereof.	AFIS 2018:	Matching (identification; fingerprint)
			Seamless Flow:	Matching (verification/identification; Face)
			NoQ3:	Matching (Verification; Face)
	6.1.2	Application realism If the test intends to evaluate performance within an application/concept of operations, the test must mimic the functional and procedural aspects	N/A	
	6.1.3	Determination of appropriate performance measures Experimenters shall determine which performance measures are applicable to their evaluation, in addition to those listed at clause 6.3	AFIS 2018:	Custom metric via penalties for accuracy and efficiency
			Seamless Flow:	Because of the proposed application setup both verification and identification metrics were reported
			NoQ3:	N/A
		Test design ensures all required metrics can be generated	AFIS 2018:	Require delivery of necessary data by contenders in evaluation protocol

[1] Part of 19795 from which the conformance item derives
19795-1:2006 applies

			Seamless Flow	*Issue with one of the algorithms; data generated by stakeholder*
			NoQ3	Test design ensures all necessary data for metric calculation is generated
		Specify the type of comparison as: a) Verification b) Open-set identification c) Closed-set identification and explain the rationale.	AFIS 2018	Open-set identification; Each fingermark could be part of a mated pair (with a fingerprint set) or unmated. i.e. not every finger mark's "true" ID is enrolled.
			Seamless flow	Verification of live image against passport reference Open-set Identification against list of enrolled IDs (kiosk images); individuals imaged can be enrolees or not, so there is no closed ID set
			NoQ3	Verification of live image against passport reference
		The comparison functionality should be applicable to the algorithms under evaluation	This appears to imply that this standard only applies to recognition algorithms; other biometric algorithms are subject to other methods of testing (e.g. other metrics).	
		Note: Error rate calculation see 19795-1:2006 clause 7		
	6.1.4	Implementation primacy		
		The test plan does not dictate the methods by which functionality is implemented; it is the responsibility of the implementation how to perform its functions	N/A; Test plan gives freedom to perform given task, only considers a given supplier-prescribed API or considers only a specific implementation.	
	6.1.5	Policies on disclosure of information to suppliers		
		...formulate policies before testing about what information will be disclosed to suppliers a) Before a test b) At execution time	*AFIS 2018: We explained what information/data would be released when and what the contenders should expect as results/reporting.* *We did not explicitly formulate policies other than that.*	

			We did take measures to ensure a low likelihood that information about data-set composition was leaked to the contenders. *All other projects: N/A* *We could discuss what guidelines we think are necessary to make sure it is clear the formulated policies are met at the time of testing; i.e. that there are not just policies but compliance*	
	6.1.6	Non-interchangeability of identification and verification attempts		
		Comparison scores that result from a one-to-many identification search shall not be presented as results of verification attempts without justification	AFIS 2018	N/A
			Seamless flow	*The Kiosk vs live comparisons may have been executed as one against many on a technical level to generate genuine and imposter scores; i.e. a zero-effort pairing.*
			NoQ3	One algorithm was executed via a tool provided by the supplier. This limited our control with respect to how scores were generated and may have been subject to some form of score normalisation. The comparisons to generate genuine and imposter scores may therefore be similar to a one-to-many identification search.
	6.1.7	Acknowledgement of models		
		If a model, approximation or prediction of identification performance is reported in place of, or in addition to, an empirical trial, the model shall be verified to the extent possible with the available data and fully documented.	N/A	
	6.1.8	Sequential use		
		The test plan shall define to order of the use of the test data. This order shall be appropriate to the application. The implementation should process the test data in this sequence.	*AFIS 2018: protocol did not specify the order in which the data should be processed. Contenders were told to load all data onto their systems and report any issues.* Other projects: N/A	
	6.1.9	Pre-test procedures		
	6.1.9.1	Installation and validation of correct operation		

		The test organisation shall [...] ensure that the hard[- and] software is installed and configured appropriately and shall verify that the system is operating correctly.	*AFIS 2018: Test plan/protocol put the responsibility of a working system on the side of the contenders; they were tasked with delivering hardware and software so they could process the data and run the comparisons necessary for the task.* Other projects: N/A
	6.1.9.2	Data preparation	
		Any subject-identifying metadata or any data ordinarily not available to the application must be expunged from the test data.	N/A: Test corpus usually just contains biometric samples. identifiers and/or filenames do not refer to, nor are they traceable to demographic factors or other potentially identifying metadata.
	6.1.10	Generic test execution sequence	
		Generic description of technology test execution: - Convert enrolment samples to biometric references; may be stored in linear collection - Identification/verification samples are converted to biometric sample features - Comparisons are made i.e.: o Verification: direct comparison to sample features and biometric reference o Closed-set identification: search of enrolled population to return user identifier o Open-set identification: search enrolled database and ▪ Return one or more identifiers ▪ Return null identity i.e. indicate the test subject is not found in the database	N/A; We never prescribed a particular test execution and/or executed the test ourselves
		Note: annex A of 19795-2 contains execution sequences for specific types of technology tests;	The generic execution sequences are useful to set up a test at a high level of abstraction. They do not help in planning for more complex cases, nor do they provide a practical script for execution.
	6.2	Assembling an appropriate test corpus	
	6.2.1	General	
		Technology evaluation => evaluate one or more algorithms for enrolment/comparison performance.	*Not sure what this means;*

		Test planning is contingent on the type of data the experimenter wishes to generate.	*We generated the necessary data ourselves, instructed contenders in detail what data they needed to deliver, or had a third party generate the necessary data.* *Strictly speaking only the instruction to contenders was pre-planed and explicitly part of a "test plan"; other actions were implicitly planned or ad-hoc solutions.*	
	6.2.2	Unique enrolment		
		All corpus samples should correspond to real people.	All data we used was captured from real people. (Morphed images correspond to *two* real people of course but are excluded from this analysis.)	
		An evaluation design should not intentionally enrol different samples from the same individual as if they were from different individuals.	*Note that – strictly speaking - for operational fingermark data (like in AFIS 2018), due to uncertainty of the ground truth, we cannot exclude the possibility that some fingermarks are from the same individual, even though they have not been identified.* *However, this does not constitute enrolment of different samples from the same individual as if they were from different individuals.*	
		For tests where each identity corresponds to a different individual, the testing organisation shall report processes implemented to ensure this.	*AFIS 2018*	*We put in our best effort to de-duplicate based on the meta-data we had; there was at least one error discovered based on updated meta-data (i.e. a fingerprint set had been linked to an existing record/ID after meta-data export)*
			Seamless flow	*No processes in place to make sure there is a duplicate entry; Apparent assumption is that no one would volunteer twice to participate in the data acquisition for this project*
			NoQ3	*Data was de-duplicated before use; no duplicate entries should be left.*
		If it is possible for one individual to have multiple identities in the corpus, it may be "cleaned" if practical; otherwise the test should proceed under the assumption that each identity corresponds to a different individual.	*AFIS 2018*	*Testing corpus was "cleaned" by selecting just one "tenprint" reference set of an individual enrolled person, and unique mated pairs – i.e. one*

				fingermark matches with one tenprint, and a tenprint only matches with one fingermark.
			Seamless flow	*N/A*
			NoQ3	*Deduplication through unique identifiers available in the data*
	6.2.3	Recurrence of data acquisition		
		Depending on access/availability of test subjects one can decide to capture reference samples multiple times over time to measure biometric reference ageing.	N/A; Data acquisition was only "repeated" for some SOTAMD subjects over a single delay for at best a few months. This doesn't allow for accounting for biometric reference aging.	
	6.2.4	Test subject identification		
		The experimenter shall report information on how test subjects are identified, including at a minimum:	*This was not reported in a strict sense*	
		a) Types of identifiers used to identify test subjects b) Amount and type of personal data collected	AFIS 2018	Used existing personal identifiers; no personal data collected
			Seamless flow	Used personal identifiers provided by data-collector/stakeholder; no personal data available to us.
			NoQ3	Used personal identifiers provided in data; some personal data was collected and used; this was not explicitly mentioned.
	6.2.5	Provision of non-biometric information		
		If available (in corpus) metadata normally available to the deployed system shall be provided to the systems under test.	N/A; Systems involved are usually just the biometric matching algorithms, their provided API's or a test setup of the operational system is used making this void.	
		Test report states types and names of metadata variables provided.	N/A	
	6.2.6	Representativeness of the corpus		
		Consider whether data in corpus is appropriate for the goals of the test or the applications of interest	*Odd request; I would expect this to be part of the job of a testing expert involved.*	
		If data is acquired under supervision of test organisation, information pertaining experimenter-test subject interaction shall be recorded in the areas of acclimatisation, training, habituation, and guidance.	Noncompliant; when data was acquired by ourselves (SOTAMD/NoQ3) we did not we did not record said information. We used existing corpuses acquired in daily operation (AFIS 2018),	

			specially acquired from daily operation (NoQ3), collected on voluntary basis by a third party on site (Seamless Flow).
	6.2.7	Untainted corpus	
		The corpus may be considered "tainted when:	
		a) Any implementation supplier has had possession of the corpus b) Any implementation supplier has provided equipment used in collecting or processing the corpus, particularly if this activity influenced the nature or quality of the corpus such as by excluding samples c) A system being tested has previously been tested and tuned using the corpus	*An instance of item b – Any implementation supplier has provided equipment used in collecting or processing the corpus - may have occurred with AFIS 2018.* *The question is whether this applies to data collected during daily operations using a system supplied previously by any implementation supplier that is evaluated.*
		When use of a tainted corpus is unavoidable this fact shall be documented in the test report	*Unclear if we documented the abovementioned issue in our test report.*
	6.2.8	Retirement of corpus	
		Samples should not be used in an evaluation of any system which has been tuned on basis of performance measured in a previous test using said samples.	N/A; as far as I know no test we supervised/conducted ever used data that had been used previously.
	6.2.9	Corpus validation	
		Validation : process whereby test subject data is screened for the purpose of removal of data not suitable for the purpose of the evaluation	Data was either provided to us as-is or selected by us based on (partially uncertain) meta-data. Selection included de-duplication but the data available was all "suitable" for the evaluation. Data provided by stakeholder could not be validated.
		Validation may include that test subject data is present, that it is in the correct format, that the correct instance has been collected, and that ground truth errors are identified.	
		Experimenters shall report if test subject data has been validated	As far as I know we never explicitly reported whether our test corpus had been validated…
		If so, experimenters detail the method(s) applied in validation	…what methods we applied to do so…
		The proportion and criteria for data removal shall be reported	… nor the proportion and criteria of removal of data.
	6.2.10	Corpus collection environment	
		Environmental conditions at data collection may influence biometric performance.	

		Available information pertaining to environmental conditions during corpus acquisition relevant to the modalities under evaluation should be reported	Most of the time the conditions were not known to us, and were not under our control….	
		If applicable experimenters should report that such information was not available	…so we should have reported as such but we did not.	
	6.2.11	Failure at source		
		Offline tests use stored biometric samples that may have been gathered with or without a biometric system in the acquisition process.		
		The test report shall disclose any known information about how the data was processed in any stage before use in the test	N/A; Usually we have no details how data has been processed, if at all.	
		Particularly when samples were discarded/rejected then the Failure At Source (FAS) rate shall be reported	N/A; Given data collected outside of our direct supervision FAS is unknown or uncertain. For NoQ3 we had data on failures at the gate but it was unclear whether this referred to discard/rejection of samples or technical failures.	
	6.3	Performance measurement		
	6.3.1	Enrolment	(We generally do not measure enrolment performance)	
		Offline test shall record the proportion of test subjects for whom an implementation elects to reject enrolment of their designated enrolment samples in the corpus as Failure to Enrol rate (FTE).	Uncertain	
			AFIS 2018	We decided not to report FTE because we had reasons to expect some failures to be due to data quality issues outside of the control of the contenders (i.e. we could not ensure the failure is due to the implementation)
			Seamless flow	We did not report FTE
			NoQ3	Accounted for FTE but did not report
		The experimenter shall specify the minimum number of samples required, and maximum number of samples permitted, for successful enrolment	N/A; we never considered enrolment performance. Corpora usually contain one (reference) sample for enrolment and one probe sample	
		For each biometric system tested, the experimenter should calculate: a) Distribution of enrolment quality scores	*Is this a requirement or a guideline?*	
			AFIS 2018	Not calculated either item
			Seamless flow	Not calculated either item

		b) Failure to enrol for different demographic groups, or associated with different environmental conditions, or for other logical segments of the corpus	NoQ3	Calculated item b) no information on item a) available.
	6.3.2	Failure to acquire		
		Offline tests shall record the proportion of verification or identification attempts for which the system fails to capture or locate an image or signal of sufficient quality as the Failure to acquire rate (FTA).	AFIS 2018	Similar to enrolment; we chose not to report this because of data issue uncertainties.
			Seamless flow	No data collected related to FTA; not reported
			NoQ3	Attempted to account for FTA by counting number of probe samples where template generation failed
		The formula for calculating FTA can be found in ISO/IEC 19795-1		
	6.3.3	Verification metrics		
		For each verification system tested, the experimenter shall compute the following:	We don't generally calculate these metrics unless these are used directly. AFIS 2018: only calculation of penalties, none of these Seamless flow: calculation of FMR/FNMR and FRR/FAR for appropriate comparisons	
		a) False match rate (FMR) and false non-match rates (FNMR)	When appropriate	
		b) False reject rate (FRR) and false accept rate (FAR), unless test design is such that false accept rate and false reject rate are identical to false match rate and false non-match rate	When appropriate	
		c) Number of genuine and impostor comparisons executed	Not a standard part of a report	
		d) For genuine test subjects, distribution of time elapsed between enrolment and acquisition of sample features, if available	Usually not available for the system under test; matching algorithms are mostly executed on different hardware for testing	
		e) Uncertainty of test results, as well as basis and formulae for estimating uncertainty	Not a standard part of reporting; we used Credible intervals in AFIS 2018 and Seamless flow to varied effect.	
		FMR and FNMR as well as FAR and FRR, may be rendered into receiver operator characteristic (ROC) or detection error trade-off (DET) curve.	Generally we used DET curves to compare overall comparison performance.	
		Note: for systems that return match/non-match decisions as opposed to comparison scores, performance may be reported at a single operating point on the ROC or DET.	N/A; Seamless flow: calculation of operating point compliant with norms	

			NoQ3: calculation of performance (+ uncertainty) at operationally relevant score threshold	
		For verification systems the experimenter should calculate:		
		a) Distribution of comparison scores for genuine test subjects and impostors b) Verification results for different demographic groups or associated with different environmental conditions or for other logical segments of the corpus	We did not calculate distributions on any of our reports	
			Seamless flow	N/A; No data on demographic groups environmental conditions or other logical segments
			NoQ3	Compliant; results calculated for different demographic groups, different environmental condition (i.e. locations), and document issue year.
	6.3.4	Identification metrics		
		For all identification systems tested, the experimenter shall calculate the uncertainty of test results, as well as basis and formulae for estimating uncertainty	We have not calculated distributions for any of the evaluations	
		For closed-set identification systems, the experimenter shall calculate the following:	N/A	
		a) Cumulative match characteristics (CMC) b) Number of searches executed	N/A	
		For open-set identification systems the experimenter shall calculate the following:		
		a) False positive identification rates (FPIR) and corresponding false negative identification rates (FNIR), preferably over a range of thresholds	Calculated when appropriate over a range of thresholds (Seamless flow, second stage) including DET (like) curve	
		b) Bining error rate and penetration rate if binning is used	N/A	
		For identification systems the experimenter shall calculate the following:		
		a) Identification results for different demographic groups, or associated with different environmental conditions, or for other logical segments of the corpus	Data not available to either segment the results or segment the data subjects.	

	6.3.5	Generalized error rates including failure to enrol and failure to acquire	
	6.3.5.1	General	
		The immediate output of an offline test – i.e. paired false match/non-match rates – shall be combined with the measured values of failure to acquire and failure to enrol.	Point of interest/action; a way to intuitively combine FMR/FNMR with FTA/FTE rates
		Note: if FTE/FTA are zero this should be noted; if either is non-zero then FAR and FRR should be computed as they differ from FMR/FNMR.	*I'm not sure if we make this distinction like this i.e. if FAR/FRR are different from FMR/FNMR if and only if either FTE/FTA (or both) are nonzero.*
		Note: Samples that resulted in failure to acquire or enrol may be released to suppliers for further study	
		Note: For systems that return match/non-match decisions performance may be reported at a single operating point on the ROC/DET	
		Note: A large number of failures to enrol decreases the generalized FAR (GFAR) but increases the generalized FRR (GFRR).	
	6.3.5.2	Single-attempt transactions	Unclear whether this applies to any of our evaluations
		For each implementation under test the experimenter shall determine the generalized FAR (GFAR) and generalized FRR (GFRR) for single-attempt transactions.	N/A - unclear
		When transactions consist of single attempts the GFAR is the proportion of impostors who are acquired and matched at some operating point threshold, t: GFAR(t) = (1-FTA) FMR(t) (1-FTE) Similarly the GFRR is the proportion of genuine users who are unable to be acquired, who can acquire but can't enrol or who are able to acquire and enrol but are falsely rejected, at some operating point threshold t: GFFR(t) = FTA + (1-FTA) FTE + (1-FTA)(1-FTE)FNMR(t) These equations hold only for n = 1 i.e. single-attempt transactions.	
		Note: different formulae may be needed if it is acceptable to not enrol certain individuals	(Noted)

		Note: One can avoid explicit failure to enrol and acquire measurements by specifying that all verification comparisons will result in a comparison score. A supplier can satisfy this requirement by internally recording a falure-to-enrol or failure-to-acquire condition and report suitably low values when such a biometric reference is used in a one-on-one comparison. This will correctly include FTA and FTE in the DET characteristic.	(Noted)
		Note: Alternatively GFAR and GFRR can be determined by: - Including failed (neither accept nor reject) impostor transactions and impostor transactions for individuals whose enrolment failed in the total number of impostor transactions - Including failed (neither accept nor reject) genuine transactions and genuine transactions for individuals whose enrolment failed in the total number of genuine transactions and - Counting failed (neither accept nor reject) genuine transactions and genuine transactions for individuals whose enrolment failed as false rejects	(Noted)
	6.3.5.3	Multi-attempt transactions	
		In case transactions consist of multiple attempts, GFAR and GFRR calculations should be generated on a test-specific basis	
	6.3.6	Throughput performance	*In general we do not measure throughput performance given that the technology tested is run on different hardware as used in the implementation.*
			We therefore skip this section.
	….		
	6.4	Reporting	
	6.4.1	General	
		The results from the evaluation shall be presented in a test report.	Complied
		A test report shall document the entire test process.	*Unclear; Test process is reported as completely as possible and necessary to explain the results.*
		All reporting requirements in clause 6.1 through 6.3 shall be documented in the test report.	*Not/not completely compliant; See above for details*

		If a requirement was out of the scope or not applicable, the report shall state that (the requirement was out of scope or was not applicable).	*Not compliant*	
		The report may be released to different audiences in separate sections on different timelines.	*(Not a requirement) varies between projects*	
	6.4.2	System information		
	6.4.2.1	Specifications		
		For the biometric system(s) tested, the experimenter shall report:		
		a) For acquisition devices: manufacturer, model, version, and firmware as applicable. If the acquisition device's core acquisition components are integrated within a third party device, then manufacturer, model, version and firmware of the core acquisition components shall be reported	N/A	
		b) For comparison algorithms: provider, version, revision	AFIS 2018	We did not have version or revision of the algorithms run; we did not report
			Seamless flow	We reported the information for the algorithms we ran on our systems; information was not reported on algorithm which was run by third party
			NoQ3	We reported the information for the algorithms we ran on our systems; information was not reported on algorithm which was run via software tool from provider
		c) Specification of the platform through which systems were tested, including but not limited to, platform, OS, processing power, memory, manufacturer, database type, database size and model.	We did not report the specification of the platform we used to test algorithms.	
	6.4.3	Data collection process		
		The experimenter shall report the following information related to data collection:		
		a) Methods of recording data for each performance element, including those not logged by the system(s)	*Not sure;* *In case we run algorithms on our own systems, we may not specify how we collect the necessary data.* *In case third parties/stakeholders run algorithms we usually are*	

			provided with the data they generate to our specification, but we don't know how they gathered it.
		b) Processes for auditing and validating performance data collection, including those not logged by the system(s)	Not sure/questionable; I think we need to consider whether we have validation and auditing measures in place for performance data.
		The experimenter shall provide examples of data collection elements such as spreadsheets and logs whether as screenshots or reproduced forms	*We have not provided examples of performance data collection elements, mainly because data usually is collected automatically only*
	6.4.3.1	Architecture	
		For the biometric system(s) tested, the experimenter shall report:	
		a) Biometric data acquisition, processing and storage architecture	N/A; *may be a point of interest for NoQ*
		b) Data flow between system components	N/A
	6.4.3.2	Outputs	
		For biometric system(s) tested the experimenter shall report:	
		a) Types of outputs the system reports, including but not limited to: comparison scores, accept/reject decisions, candidate lists, enrolment quality scores, sample quality scores;	*Not sure we reported this*
		b) Range of comparison scores system is capable of reporting, as well as supplier-specified thresholds	Not reported
		c) Range of enrolment quality scores system is capable of reporting as well as supplier-specified thresholds	*N/A; not reported*
		d) Range of sample quality scores system is capable of reporting as well as supplier-specified thresholds	*N/A; not reported*
		e) Method(s) through which outputs are provided by the system	*N/A; not reported*
	6.4.3.3	Method of implementation	
		For each biometric system tested, the experimenter shall report system implementation information corresponding to each of the following:	
		a) Method of biometric and platform system acquisition b) Level of supplier involvement in system implementation	*N/A; not reported*
	6.4.4	Disclosure	
	6.4.4.1	External reporting	

		A test plan shall disclose what input, intermediate and output material is to be made available to non-suppliers on what schedule and to whom	*N/A; Unclear whether this is relevant to our evaluations, assuming this does not apply to agreements on delivery of reports to the stakeholder/client*	
	6.4.4.2	Sample properties disclosure		
		The test plan shall specify what sample related information will be provided to suppliers, and on what schedule. This may be modified in response to formal comments from suppliers	AFIS 2018: we clearly specified what information the contenders would get (and also which information they would not). Other projects: N/A	
	6.4.5	Report structure		
		The following sections shall be incorporated in the test report:		
		- Executive summary	Complied	
		- Characteristics of corpus data	Mostly complied; data is described in general terms	
		- Specific test purposes	Complied	
		- Data collection	Complied	
		- Data analysis	Mostly complied; we do not analyse the data other than calculate metrics and visualisations	
		- Record keeping	Mostly not complied	
		- Performance results	Complied	
		- Full test plan	Not complied	
	8	Other issues applicable to technology and scenario evaluations		
	8.1	Parties to a test		
		An evaluation shall be conducted by a tester.		
		The biometric system under test shall be provided by one or more suppliers		
		If the tester and supplier are the same entity, or are affiliated, or are otherwise not independent, then this shall be documented in the test report	AFIS 2018	Algorithm suppliers were allowed to set up their own hard- and software under strict supervision, making them testers; this was described in the test plan and report
			Seamless flow	Supplier of the operational system provided us with score values for one algorithm; this was not clearly noted
			NoQ3	*No comments*

		Supplier involvement in technology and scenario evaluations is restricted to the supply, installation and configuration of the software and/or hardware.	This seems to contradict with aforementioned. (This may be resolved by having a strict supervision of any tester that is also a supplier)	
		The testing organisation executes enrolment and comparison tests without supplier input.	This seems to contradict with aforementioned. :	
			AFIS 2018	Testers/Suppliers executed enrolment and comparison tests under strict supervision of testing organisation. One tester/supplier reported to have been in contact with their organisation, which was explicitly prohibited.
			Other tests	Compliant
		Note: If it is imperative that an evaluation is construed to be a supplier's own best effort with no possibility of error by the testing organization, an alternative type of test in which tester and supplier roles differ from those enumerated in 8.1 can be conducted. Known as a supplier self-test, this type of test allows the supplier to provide, configure, and operate their own system on tester supplied materials. The tester is absolved of blame if the results are claimed to be deficient. Such an evaluation should use a client-server paradigm. Such tests are problematic in terms of expense, gaming and sample privacy.	AFIS 2018 appears to be a "supplier self-test" with the exception that it was also executed on supplier owned hardware. This would be ameliorated by direct supervision of the testing organisation, and by having each supplier set up their hardware in isolated rooms at testing organisation's site without direct contact to the outside.	
	8.2	Fairness		
		A competitive test shall not be designed to favour particular suppliers	We have no indication that the test design or execution favoured any particular supplier in any of the projects	
		Note: answers to questions by suppliers ("fishing for information") should be made public.	*Questions were handled by the testing organisation (not us) on AFIS 2018 (the only relevant case). Answers were released to all suppliers.*	
		Note: With technology tests usually representative sample data is released in the format to be used in the test.	For AFIS 2018 (the relevant case) there was no sample in advance of the test. However, suppliers were required to set up their systems in advance of the test and then load the dataset to estimate if they were capable of running the test.	

		Experimenters shall document any involvement on the part of the test organisation in the configuration, modification, refinement or adaptation of the implementation under test.	N/A	
		Experimenters shall document intellectual or physical input on the part of the test organisation that materially affects any outcome of the evaluation	N/A	
		In case of multiple components or systems are tested, examiners should report whether computing systems were tested on equivalent hardware and operating systems or whether images of operating systems were re-installed prior to each test segment in a system by system testing manner.	*We required the contenders to supply their own hardware as required to run their solution. They did not communicate this with us, and conferred with the test organisation to make sure the necessary resources were available.*	
	8.3	The experimenter shall report address the basis by which algorithms and systems are included in technology and scenario evaluations.	Interpreted as saying: "the experimenter's report shall address…"	
			AFIS 2018	We did not address the selection criteria by which the contenders were pre-selected; contenders took part by their own volition and were not selected on anything but that intention.
			Other tests	Algorithms were selected by commissioner of test
		Inclusion of algorithms or systems in the evaluation might be on the basis of: a) An open invitation to participate b) Selection by a test organisation, in which case the selection criteria shall be reported c) A contract with a supplier or a 3rd party to test a particular system	AFIS 2018	Option a
			Other projects	Option b; we did not report the selection criteria
		Technology and scenario evaluations can incorporate a single biometric system or multiple biometric components or systems. Technology and scenario evaluations can also incorporate like combinations of multiple biometric components or systems. Testing multiple systems provides the advantage of potentially establishing a range of performance against which different systems can be evaluated. Anomalous performance can be difficult to gauge from a single-system test. The number of systems teste may be constrained by budgetary constraints, availability of suitable technologies, or time required to acquire sample or process data.	Noted	

	8.4	Use of frequently asked questions	
		In a competitive technology or scenario evaluation, a frequently-asked-questions document may be maintained as a mechanism of communication between a test organisation and the suppliers. The author of each question should be suppressed	Noted
	8.5	Legal issues	
		Legal issues in technology and scenario test design, execution and reporting may need to be addressed.	Noted
		It may be necessary to enact a non-disclosure agreement between suppliers and test organisation	Noted
		Certain jurisdictions require a data privacy agreement between test subject and test organisation	Noted
	8.6	Release of source code	
		Depending on the type and purpose of the test, it may be appropriate to release test source code to suppliers	Noted
	8.7	Supplier comment on test report	
		Depending on the type and purpose of the test, it may be appropriate to allow suppliers to comment on a pre-release version of the report as prescribed by the testing organisation.	Noted

UNIVERSITY OF TWENTE
Drienerlolaan 5
7522 NB Enschede

P.O.Box 217
7500 AE Enschede

P +31 (0)53 489 9111

info@utwente.nl
www.utwente.nl