International Standard

ISO/IEC 19795-10

Information technology — Biometric performance testing and reporting —

Part 10: Quantifying biometric system performance variation across demographic groups

First edition
2024-10

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

Partie 10: Quantification de la variation des performances du système biométrique selon les groupes démographiques

Reference number
ISO/IEC 19795-10:2024(en)

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).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

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.*

A list of all parts in the ISO/IEC 19795 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

As the use of biometric technology increases, so too does public interest in establishing whether the technology performs similarly for all individuals. Stakeholders are asking government and industry organizations that use biometric technology to establish whether these technologies vary in performance for different demographic groups. The intention of this document is to provide guidance on how to measure and report performance variation across demographic groups.[2]

This document is intended to help organizations evaluate demographic performance in biometric systems and report their results. Specifically, this document outlines how to measure and report biometric performance variations across demographic groups. It provides a set of metrics and best practices to facilitate such testing. However, this document does not provide guidance on how to establish specific causes for the observed variations. The following demographic variables are explicitly discussed in this document:[7][10][12]

- biological characteristics, such as:
  - sex, age, weight, height and skin lightness;
- social constructs, such as:
  - ethnicity, gender and language.

Many other variables can cause systematic changes in biometric characteristics or in how individuals interact with biometric systems. The following demographic variables are relevant although not explicitly discussed in this document:

- performance variations based on temporary states, such as:
  - self-styling (e.g. makeup, eyewear, mask-wearing, clothing, hairstyles),
  - behavioural or emotional states (e.g. intoxication),
  - behaviours (e.g. smiling, closing eyes, varying pose);
- performance variation caused by diseases or injuries, such as:
  - eye surgery, cataracts, vision correction,
  - stroke, cleft lip, Apert’s syndrome,
  - missing digits;
- performance variation caused by disabilities.

Demographic performance variation for applications other than biometric recognition, such as emotion, gender or age estimation, are not considered in this document.

Information technology — Biometric performance testing and reporting —

Part 10:
Quantifying biometric system performance variation across demographic groups

1 Scope

This document establishes requirements for estimating and reporting on performance variations observed when cohorts belonging to different demographic groups engage with biometric enrolment and recognition systems. In this context, performance refers to failure-to-enrol rate, failure-to-acquire rate, shifts in comparison score, recognition error rates, and aspects of response and processing time (throughput).

This document is applicable to the following:

- demographic group membership;
- using phenotypic measures;
- reporting on tests;
- stating statistical uncertainty estimates;
- operational thresholds settings;
- equitability;
- procurement agency activities.

This document also provides terms and definitions to be used when reporting performance variation across demographic groups.

This document is applicable to:

- technology evaluations of algorithms, subsystems and systems;
- scenario evaluations of systems;
- operational evaluations of fielded systems.

Application of this document does not require detailed knowledge of a system's algorithms but it does require specific knowledge of the demographic characteristics for the population of interest.

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*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 2382-37 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
differential performance measure
DPM
difference in biometric system measures across different demographic groups

EXAMPLE Differences in error rates [e.g. False Match Rate (FMR), False Non-Match Rate (FNMR)] between different demographic groups.

Note 1 to entry: ISO/IEC 2382-37:2022 term 37.09.28 defines "demographic differential" as the difference in "outcome of a biometric system". This definition is equivalent to this document's "differential performance measure". This document also recognizes other kinds of demographic differentials, such as *differential treatment* (3.7) and *comparison score differential measure* (3.4).

3.2
false negative differential performance
FND
difference in false negative error rates calculated across multiple demographic groups

EXAMPLE If Group A's false non-match rate is 10 %, and Group B's false non-match rate is 20 %, the false negative differential is 10 percentage points if viewed as a mathematical difference or a factor of 2 if viewed as a mathematical ratio (see 6.4).

3.3
false positive differential performance
FPD
difference in false positive error rates calculated across multiple demographic groups

EXAMPLE If Group A's false match rate is 1 %, and Group B's false match rate is 3 %, the false positive differential is 2 percentage points if viewed as a mathematical difference or a factor of 3 if viewed as a mathematical ratio (see 6.4).

3.4
comparison score differential measure
difference in system measures across different demographic groups represented through comparison score analysis

EXAMPLE Differences in mean comparison scores for different demographic groups (see 6.9).

3.5
mated comparison score differential measure
difference in the statistics of mated score distributions observed for different demographic groups

EXAMPLE If the mean mated comparison score for subjects in Group A is 10 and the mean mated comparison score for subjects in Group B is 5, then the mated comparison score differential measure is a mean difference of 5 (see 6.9).

3.6
non-mated comparison score differential measure
difference in the statistics of non-mated score distributions observed for different demographic groups

EXAMPLE If the mean non-mated comparison score for subjects in Group A is 10 and the mean non-mated comparison score for subjects in Group B is 5, then the non-mated comparison score differential measure is a mean difference of 5 (see 6.9).

3.7
differential treatment
different set of actions for a biometric enrolee or biometric capture subject based on their demographic group

EXAMPLE Implementing a system in which one machine learning model recognizes male faces and a different machine learning model recognizes female faces.

3.8
categorical demographic variable
demographic variable of an individual that is nominally or ordinally described

EXAMPLE A data subject's gender or ethnicity.

3.9
continuous demographic variable
demographic variable of an individual that is observable, measurable and not necessarily constrained to discrete categories

EXAMPLE An individual's age or the measurement of a phenotypic trait, such as an individual's skin lightness.

3.10
intersectional demographic variable
demographic group that is the combination of multiple categorical demographic variables.

EXAMPLE A data subject's gender-ethnicity.

3.11
demographic group
value of a continuous, categorical or intersectional demographic variable associated with a data subject

EXAMPLE A data subject that has self-reported their gender as female has a demographic group of female for the categorical demographic variable of gender.

3.12
demographic reference database
database comprising biometric references annotated with demographic variables and groups

3.13
aggregate equitability measure
AEM
performance measure that combines multiple measures of differential performance into an aggregate measure of overall differential performance

3.14
confidence interval
interval estimator (T_0, T_1) for the parameter θ with the statistics T_0 and T_1 as interval limits and for which it holds that $P[T_0 < \theta < T_1] \geq 1 - \alpha$

Note 1 to entry: Unless otherwise stated, the threshold for statistical significance, α, is 0.05, which equates to a 95 % probability that the parameter is within the interval limit.

[SOURCE: ISO 3534-1:2006, 1.28, modified — original Notes to entry have been removed and replaced by a new Note 1 to entry.]

3.15
effect magnitude
statistical measure of the size of an observed differential

EXAMPLE 1 A mathematical difference of 20 percentage points in false non-match rates between two demographic groups (e.g. 5 % vs. 25 %).

EXAMPLE 2 A mathematical ratio of 5 between false non-match rates between two demographic groups (e.g. 5 % vs. 25 %).

4 Conformance

To conform to this document, a biometric evaluation assessing performance variation across demographic groups shall be planned, executed and reported in accordance with the requirements contained in Clauses 5 to 7.

5 Planning the evaluation

5.1 Identifying the scope of the evaluation

This subclause establishes the experimental methods for designing evaluations to measure demographic differences in the performance of biometric systems. In general, experimental design includes setting the objectives of an evaluation and determining the statistical properties and design of the evaluation to match the objectives. This document applies specifically to evaluations in which one of the objectives is to calculate differential performance measures (DPMs) or calculate comparison score differential measures in biometric systems across different demographic groups. Prior to executing the evaluation, the tester shall prepare a test plan describing the evaluation.

- Test plans shall describe the objectives as well as any models or hypotheses of the evaluation, including:
  - demographic variables and groups of interest;
  - biometric performance measures of interest;
  - demographic differential performance measure(s) and/or comparison score differential measure(s) of interest;
  - effect magnitude(s) of interest.
- Test plans shall describe the data that will be gathered to test these models or hypotheses, including:
  - how demographic variables are to be measured or otherwise collected;
  - manipulated, fixed or blocked factors, including counterbalancing factors where appropriate;
  - controls for non-tested factors;
  - target sample size requirements, including the rationale for cohort selection when generating mated and non-mated trials and the target type I and type II error.
- Test plans should describe what analyses will be performed on these data and what inductions will be attempted, including:
  - the expected uncertainty around differential performance measures;
  - statistical tests to be performed.

The balance between the internal and external validity of the evaluation should be considered and explained. Evaluations with high internal validity, such as technology tests, focus on specific components of biometric systems and are well controlled: many factors are considered, documented and manipulated in a controlled fashion or fixed at pre-determined levels. Evaluations with high external validity, such as operational tests, are not necessarily able to attribute the observed differentials uniquely to the factors of interest due to uncontrolled variation in the test environment. These evaluations therefore have lower internal validity. Any deviations between the test design and the envisioned operational conditions for the system shall be noted and reported as these efforts to control variation may change the effect magnitude observed relative to the target environment within which the biometric system operates (see 7.2).

This document does not specify what constitutes an acceptable amount of differential performance. To inform the design of the evaluation, regulators or procurement guidelines can specify allowable differential performance, where appropriate, calculated according to at least one of the methods described in 6.4 to 6.9.

When specifying that a level of differential performance is not acceptable, regulators or procurement guidelines can utilize benchmarks as described in 7.4.3.

5.2 Demographic variables

5.2.1 Ground truth requirements

Evaluations of biometric performance have strict requirements for establishing the ground truth identity of data capture subjects. This is to ensure the validity of any metrics derived from these classifications, such as false match and false non-match rates. Evaluations of biometric performance across demographic groups have three additional constraints:

- Demographic evaluations shall specify the demographic variables of interest.
- Each demographic variable shall be comprised of defined demographic groups which shall be associated to individual data capture subjects.
- Demographic group membership for demographic variables of interest and other metadata should be collected at the same time as the corpus samples to avoid errors in inference.

Evaluations to measure performance variation across demographic groups should involve focused data collection where demographic groups are recorded and where ground-truth identity information is established.

Demographic group membership should not be inferred directly from biometric samples. An example of this is assigning the value of ethnicity or the value of gender from a face sample. Demographic groups are properties of a data subject or a data subject's biometric characteristic. They are not properties of a biometric sample. Estimating demographic groups from biometric samples can introduce spurious correlations between biometric performance and demographic variables. For example, if the width of a face or eyelid palpebral aperture used to estimate the demographic group is measured from the same sample used for biometric comparison, any lens distortion can affect both the biometric and the demographic outcomes. If it is not possible to establish demographic group membership independent of the biometric sample, other techniques should be applied (see 5.2.2 and 5.2.3). In this case, the tester should carefully consider and shall document any correlations and impacts between demographic variables and the biometric sample collection technique.

Many demographic variables are categorical. Categorical demographic variables are those that take a distinct, limited number of possible values, such as gender and ethnicity. Other demographic variables are continuous and have an infinite number of possible values. These can be combined into demographic groups for the purpose of analysis. In some practical applications, continuous demographic variables such as age and height are bound by natural limits and should be reported in appropriate granularity.

5.2.2 Categorical demographic variables

5.2.2.1 Sex

Sex is defined as the state of being male or female as it relates to biological factors such as DNA, anatomy and physiology. Sex typically consists of two categories, "male" and "female". Female individuals generally possess two copies of the X chromosome. Male individuals generally possess one copy each of an X and a Y chromosome. Important exceptions do occur and complicate binary classification. The tester should establish appropriate categories for sex. If necessary, the tester can extend the general binary classification model of male/female.

When sex is included in the evaluation, it shall be determined through the collection and analysis of DNA or by self-report. In evaluations that include sex, the tester shall prepare a statement that documents how sex was determined (see 7.3).

NOTE If sex was determined by self-report, gender can also be recorded.

5.2.2.2 Gender

Gender is defined as the classification of individuals as male, female or additional categories based on social, cultural or behavioural qualities. An individual's gender identity can consist of multiple, distinct categories. An individual's gender can also change over time. When gender is included in the evaluation, gender should be determined through self-reporting. Gender self-reporting options presented to the capture subject shall be documented. Gender should not be assigned by the tester conducting the evaluation.

In some evaluations that include gender, it is not always possible to obtain self-reported gender information. In this case, the tester shall prepare a statement describing why self-reporting was not possible and potential inferential errors this can cause (see 7.3).

5.2.2.3 Ethnicity

In the context of biometric evaluations, ethnicities are classifications of individuals within a society based on shared qualities that are generally considered distinct within that society. Categories can reflect common physical characteristics, ancestry, language, community, religious affiliation, cultural heritage or other common qualities.

When ethnicity is included in the evaluation, the tester shall prepare a statement that documents the method for determining ethnicity. If utilizing self-reporting to establish ethnicity, the tester shall prepare a statement that documents the ethnicity self-reporting options presented to the data subject (see 7.3). In technology and scenario evaluations, ethnicity shall be recorded and associated with collected samples through voluntary self-reporting. In operational tests ethnicity shall be established through voluntary self-reporting or from available ID data. Whenever possible, data subjects should be given the opportunity to select multiple ethnicities to designate multi-ethnic identities and the evaluation plan should enumerate the presented choices. Ethnicity shall not be assigned by the tester, for example by inspecting samples (see Reference [5] for further details).

In case of any deviations from these requirements, the tester shall prepare a statement that documents the specific reasons for use of alternate means of ethnicity determination and potential inferential errors this can cause (see 7.3). Finally, establishing ground truth for ethnicity can be inherently complex. The population of individuals that identify with specific ethnic groups can change across different societies and within a society over time. Because of this complexity, testers shall not use other demographic variables as proxies for ethnicity.

5.2.2.4 Birthplace

Birthplace refers to the geographic location (e.g. a region or country) where an individual was born. When birthplace is included in the evaluation, birthplace shall be established through voluntary self-reporting or from available ID data or documents. In evaluations that include birthplace, the tester shall prepare a statement that documents the method for determining birthplace. If utilizing self-reporting to establish birthplace, the tester shall prepare a statement that documents birthplace self-reporting options presented to the data subject. If birthplace is recorded more finely than nation state (e.g. by a region within a country), the tester shall prepare a statement that documents how this granularity was established (see 7.3). Birthplace is a distinct demographic variable from ethnicity and shall not be used as a proxy for ethnicity.

EXAMPLE Birthplace can be retrieved from an identity credential in an operational evaluation.

NOTE ID data or documents are preferred to voluntary self-report due to their reliability. However, self-report has advantages in terms of data minimization.

5.2.2.5 Place of residence

Place of residence refers to the geographic location (e.g. a region or country) where an individual lives or resides. When place of residence is included in the evaluation, place of residence shall be established through voluntary self-reporting, from available ID data, or documents. In evaluations that include place of residence, the tester shall prepare a statement that documents how place of residence was determined and categorized (see 7.3). Place of residence is a distinct demographic variable from ethnicity, since it can change quickly and arbitrarily. Place of residence shall therefore not be used as a proxy for ethnicity or birthplace.

5.2.2.6 Native language

Native language refers to language that an individual acquires fully through extensive exposure during the critical period of development and presently understands. When native language is included in an evaluation, the tester shall prepare a statement that documents the method for determining native language. If utilizing self-reporting to establish native language, the tester shall prepare a statement that documents native language self-reporting options presented to the data subject (see 7.3).

NOTE Native language can be a sensitive, complex and multi-dimensional topic, such as in cases where multiple languages are spoken at home.

5.2.3 Continuous demographic variables

5.2.3.1 Age

The age of an individual is the quantity of time that has elapsed since the moment of the individual's birth. Age is commonly expressed in months or years. When age is included in the evaluation, age shall be established through self-reporting. Age can be subsequently verified via identity documents (e.g. a driver's licence, passport, birth certificate, etc.).

5.2.3.2 Weight

In the context of a biometric performance test, weight refers to the mass of an individual relative to their gravitational conditions. Weight is expressed in Newtons but, for evaluations conducted under Earth's gravitational field, it is commonly expressed in kilograms. When weight is included in the evaluation, weight shall be established through either self-reporting or measurement using a weight scale.

In some evaluations, it is not possible to obtain measured or self-reported weight information. In this case, the ground-truth weight shall not be assigned by the tester.

5.2.3.3 Height

In the context of a biometric performance test, height refers to the measurement from the base (i.e. the feet) to the top (i.e. the top of the head) of an individual. Height is commonly expressed in centimetres. When height is included in the evaluation, height shall be established through either self-reporting or measurement using a height scale. Height can be further verified via identity documents (e.g. a driver's licence or passport).

In some evaluations, it is not possible to obtain measured or self-reported height information. In this case, the ground-truth height shall not be assigned by the tester.

5.2.3.4 Skin lightness

Skin lightness is the perceptual lightness or darkness value of an individual's skin. Skin lightness or darkness is primarily determined by the amount of melanin in an individual's skin cells. Skin lightness or darkness, or the amount of melanin in skins cells, can be impacted by ethnicity as well as external factors, such as exposure to ultraviolet radiation or levels of vitamin A in the body. When skin lightness is included in the evaluation, The ground truth of skin lightness should be established by measuring the capture subject's skin for the *L** component of the CIE *L*a*b** colour.[6][11][15] Skin lightness should be measured in a controlled manner using a calibrated instrument such as a colorimeter or a spectrophotometer. Skin lightness shall not be measured or estimated photographically unless the camera is colour-calibrated within the capture environment (see Reference [11] for further details).

In some evaluations, it is not possible to obtain measured skin lightness via the *L** component. In this case, a statement shall be included regarding the specific reasons for use of alternate means of quantification, such as a relative colour analysis, and conclusions should include a statement regarding potential inferential errors this can cause.

5.2.4 Other demographic variables

There are many physical and social characteristics not identified in this document that can be described as demographic variables. Some, but not all, can reasonably be expected to impact biometric performance. For variables not defined in 5.2.2 and 5.2.3, the tester shall prepare a statement that documents how ground truth was established (e.g. self-reported or measured, see 7.3). This also applies to intersectional demographic variables, such as specific intersections of age and gender (e.g. females over 65 years of age).

6 Executing the evaluation

6.1 Generation of mated comparison and identification trials

This clause provides requirements and guidance on establishing ground truth and conducting verification or identification trials where biometric samples from the same data subject and the same biometric characteristics are compared. A mated verification (1:1) trial involves samples from just one data subject and therefore a single demographic group, d_i. In identification applications, the demographic group of the probe sample, d_i, can differ from the demographic group of samples in the demographic reference database. However, in mated identification trials, d_i is always represented in the demographic reference database. The comparison scores from mated verification and identification trials are used in the calculation of false negative differential performance (see 6.4 to 6.6, 6.8) and false negative comparison score differential measures (see 6.9). Demographic differentials based on mated trials can be computed solely based on probe demographics, d_i. In identification trials specifically, the composition of the demographic reference database shall be considered separately from the demographic composition of the probe set and shall also be reported (see 7.3).

6.2 Generation of non-mated comparison and identification trials

6.2.1 General

This subclause provides requirements and guidance on establishing ground truth and conducting verification or identification trials where biometric samples from two different data subjects are compared. A non-mated verification trial can have two different demographic groups: the demographic group of the probe sample, d_i, and the demographic group of the reference, d_j. Biometric performance metrics based on non-mated samples [e.g. false match rate (FMR)] are therefore related to the demographics of both data subjects. Demographic differentials based on non-mated trials (i.e. false positive differential performance and non-mated comparison score differential measures) are therefore also related to the demographics of the multiple data subjects involved.

6.2.2 Verification (1:1)

In 1:1 trials that investigate false positive differential performance, one simplifying approach is to constrain analyses to cohorts where the demographics of non-mated probe and reference samples are matched (e.g. both female or both male). The tester then compares FMR for males and FMR for females. FMR measures can then be constrained to a single demographic group, the demographics of the probe (i.e. FMR_{d_i} instead of FMR_{d_i,d_j} given $d_i = d_j$). When using this simplifying approach, the tester shall compare samples within each group, d_i, as is done in formulae of false positive differential performance (see 6.4 and 6.5).

6.2.3 Identification (1:N)

Measuring demographic differentials for evaluations of identification systems is more complex than measuring demographic differentials for evaluations of verification systems. Unlike mated identification trials, non-mated identification trials are not guaranteed to have samples in the demographic reference database with demographics that match the demographic group of the probe sample, d_i. Consequently, the tester shall specify the demographic composition of the demographic reference database, G, used for identification trials. The demographic reference database can be composed of samples from one or more demographic groups or intersections. To specify G, the tester shall enumerate the total number of references enrolled, N. The tester shall also enumerate the number of enrolled references belonging to each demographic group or intersectional ($N_{d_1}, N_{d_2}, \ldots$). Because G is specified and fixed, false positive identification rate (FPIR) can be parameterized and computed solely as a function of the probe demographic (i.e. $FPIR_{d_i}$ instead of $FPIR_{d_i, G\subseteq d_1, d_2, \ldots}$). This simplifying approach is utilized in formulae of false positive differential performance in identification trials (see 6.6).

There are two options for the tester to configure the demographic reference database used for identification trials. First, the tester can specify a demographic reference database to be representative of a particular use case. Second, a tester can specify a demographic reference database with a constant number of samples per group or intersectional (i.e. $N_{d_1} = N_{d_2} = .. = N_{d_n}$). The tester can then compute $FPIR_{d_i}$ against this database for each probe demographic group, d_i. $FPIR_{d_i}$ can then be used to compute false positive differentials in identification trials (see 6.6). Using this approach, measures of false positive demographic differentials are valid only for the specified demographic reference database.

6.3 Selection of a threshold

Measurements of demographic differential performance shall reflect the differences in error rates, not the differences in success rates. Error rates are threshold-dependent, so measures of differential performance are also threshold-dependent. This is appropriate as most operational biometric systems operate at a fixed threshold. Testers shall select a threshold in an evaluation of demographic differential performance in one of two ways:

— by selecting a threshold reflective of the intended operational use case;

— by selecting a threshold so that a given algorithm achieves a global false match or false positive identification rate on the full demographic reference database (i.e. against the full demographically aggregated population).

Demographic differentials across two or more demographic groups can then be explored.[10]

NOTE Studies that report differentials across cohorts where the threshold is allowed to vary across cohort groups are not representative of how differentials are experienced in real world operations. For example, when performance differs for cohorts A and B, reporting false non-match rate with a constant false match rate masks the fact that to achieve a constant false match rate, the threshold for the two cohorts likely changed.

6.4 Calculating differential performance based on categorical variables for two specific demographic groups

6.4.1 General

For DPMs between two demographic groups, d_i and d_j, based on one or more categorical demographic variables, error rates are measured at a threshold, τ (see 6.3), for the specific groups of interest in the set of demographic groups, D. False positive differential performance, $FPD(\tau)$, and false negative differential performance, $FND(\tau)$, between the two specific demographic groups shall be calculated using either a difference (see 6.4.2) or a ratio (see 6.4.3) or both.

Measures of false positive and false negative differential performance shall not be combined as the demographic differentials at operational thresholds can often vary by orders of magnitude and this often leads to the use of extreme valued weighting parameters (see Annex B).

6.4.2 Differential performance between two groups based on mathematical difference

The mathematical difference in error rates between two groups is calculated as follows:

$$FPD(\tau) = FMR_{d_i}(\tau) - FMR_{d_j}(\tau); d_i, d_j \in D$$

$$FND(\tau)=FNMR_{d_i}(\tau)-FNMR_{d_j}(\tau);d_i,d_j\in D$$

False match rates can vary across several orders of magnitude. Testers should consider differences in order of magnitude before calculating false positive differentials based on a mathematical difference.

6.4.3 Differential performance between two groups based on mathematical ratio

The error rate ratio between two groups, where the larger error rate is in the numerator and the smaller error rate is in the denominator, is calculated as follows:

$$FPD(\tau)=\frac{FMR_{d_i}(\tau)}{FMR_{d_j}(\tau)};d_i,d_j\in D$$

$$FND(\tau)=\frac{FNMR_{d_i}(\tau)}{FNMR_{d_j}(\tau)};d_i,d_j\in D$$

Both false match rates and false non-match rates can, under certain conditions, be small in magnitude, up to and approaching zero. Testers should consider the impact of small numbers and the statistical confidence in those numbers when calculating demographic differentials based on mathematical ratio.

6.5 Calculating differential performance based on categorical variables for more than two groups

6.5.1 General

In some demographic evaluations, testers may quantify DPMs across more than two demographic groups. For differential performance across more than two demographic groups $(d_1,d_2,\ldots,d_n)$ based on one or more categorical demographic variables, error rates are measured at a threshold (τ; see 6.3), for all groups in the set of demographic groups, D. Note that $n=|D|$. False positive differential performance, $FPD(\tau)$, and false negative differential performance, $FND(\tau)$, across all demographic groups shall then be calculated using the largest error rate relative to the geometric mean (see 6.5.2) or a method based on the Gini coefficient (see 6.5.3) or both. To calculate confidence intervals on these DPMs, the techniques outlined in 6.10.1 shall be used.

Measures of false positive and false negative differential performance shall not be combined (see Annex B).

6.5.2 Differential performance for more than two groups based on the largest error rate relative to the geometric mean

DPMs across more than two groups based on the largest error rate relative to the geometric mean is calculated as follows:[9]

$$FPD(\tau)=\frac{\max\limits_{d_i\in D}\left(FMR_{d_i}(\tau)\right)}{\widehat{FMR(\tau)}}$$

$$FND(\tau)=\frac{\max\limits_{d_i\in D}\left(FNMR_{d_i}(\tau)\right)}{\widehat{FNMR(\tau)}}$$

where $\widehat{FMR(\tau)}$ and $\widehat{FNMR(\tau)}$ are the geometric mean error rates at threshold τ calculated according to $\hat{x}(\tau)=\left(\Pi_{d_i\varepsilon D}x_{d_i}{}^{w_{d_i}}\right)$. The weight of each group, w_{d_i}, is 1 divided by the number of groups: $\frac{1}{n}$.

When using the largest error rate relative to the geometric mean, the selection of the sample size per demographic group is a key consideration. The tester shall select the sample size per demographic group,

N_{d_i} , using the guidance in 6.10.3. When sample size cannot be selected in accordance with 6.10.3, the tester shall use bootstrap techniques as outlined in 6.10.1 to estimate the variance of differential performance measures based on the geometric mean. In either case, the tester shall disclose the sample size per demographic group, N_{d_i} .

The geometric mean of a set of numbers is not defined if any error rate in the set is 0. In such cases, the error rate can be replaced by:

a) the lowest error rate sustained by the number of biometric comparisons or trials;

b) a value, ε, which is a statistical upper bound on the observed error.

ε is determined based on the sample size of the study using a binomial generalization of the statistical technique known as the “rule of 3”. In this model, error rates are replaced with the value, *p*, that is the solution of $I_p(x+1,m-x)=1-\alpha$, for *x* errors in *m* trials, where $\alpha=0.05$ for a 95 % confidence interval. I_p is the incomplete beta function. For *m*, use the number of trials in the group, N_{d_i} .

NOTE When the number of errors (x) equals 0, error rates are replaced with a value $p = 3/N_{d_i}$. In statistical analysis, this is referred to as the “rule of 3”.

6.5.3 Differential performance for more than two groups based on the Gini coefficient

DPMs for more than two groups based on the Gini coefficient is a formula that calculates the spread of error rates using a statistical technique known as the Gini coefficient. This document adopts the approach taken in[13] which normalizes the Gini coefficient by a factor of (n/(n-1)) to correct for downward bias when the number of demographic groups is small. In the context of demographic evaluations of biometric performance, differential performance for more than two groups based on the Gini coefficient is calculated as follows:[13]

$$FPD(\tau)=\left(\frac{n}{n-1}\right)\frac{\sum_{\mathrm{i}}\sum_{\mathrm{j}}\left|FMR_{d_{\mathrm{i}}}(\tau)-FMR_{d_{\mathrm{j}}}(\tau)\right|}{2n^2\,\overline{FMR(\tau)}}\ \forall d_{\mathrm{i}},\ d_{\mathrm{j}}\in D$$

$$FND(\tau)=\left(\frac{n}{n-1}\right)\frac{\sum_{\mathrm{i}}\sum_{\mathrm{j}}\left|FNMR_{d_{\mathrm{i}}}(\tau)-FNMR_{d_{\mathrm{j}}}(\tau)\right|}{2n^2\,\overline{FNMR(\tau)}}\ \forall d_{\mathrm{i}},\ d_{\mathrm{j}}\in D$$

Where $\overline{\mathrm{FMR}(\tau)}$ and $\overline{\mathrm{FNMR}(\tau)}$ are the arithmetic mean error rates at threshold τ calculated according to $\bar{x}(\tau)=\frac{1}{n}\sum_{d_i\varepsilon D}x_{d_i}$.

The Gini coefficient is a measure of the statistical dispersion of a set of numbers.[4] It is robust to the presence of error rates of 0 and is mathematically bounded between 0 and 1.

6.6 Calculating differential performance in identification trials

DPMs in identification trials shall be measured using the FPIR and the false negative identification rate (FNIR). The FPIR and the FNIR are computed against a demographic reference database. The demographic reference database shall be specified by the overall number of references enrolled, *N*, as well as by the number of enrolled references belonging to each demographic group or intersectional, $N_{d_1}, N_{d_2}, \ldots$.

To calculate DPMs in identification trials, compute $FNIR_{d_i}$ and $FPIR_{d_i}$ for mated and non-mated transactions from capture subjects belonging to each demographic group, d_i. DPMs in identification trials shall then be computed using the formulae given in 6.4.2, 6.4.3, 6.5.2 and 6.5.3 by substituting $FNIR_{d_i}$ for $FNMR_{d_i}(\tau)$ and $FPIR_{d_i}$ for $FMR_{d_i}(\tau)$. $FNIR_{d_i}$ and $FPIR_{d_i}$ can be computed based on a threshold τ or based on a candidate rank, *R*, and can vary depending on the total number of enrolled references, *N*, and the number of enrolled references belonging to each demographic group or intersectional, $N_{d_1}, N_{d_2}, \ldots$. Measures of demographic differentials are therefore valid only for the specified demographic reference database and threshold setting.

6.7 Calculating demographic differentials for failure-to-enrol rate, failure-to-acquire rate and transaction duration

Evaluations of biometric systems using measures such as failure-to-enrol rates (FtER), failure-to-acquire rates (FtAR), and transaction durations perform trials involving capture subjects. Evaluations that measure differential performance using these measures shall compute the values of the required measures separately based on the demographics of the capture subject, d_i.

The mathematical difference or ratio in biometric DPMs between two groups (d_i, d_j) shall be calculated as follows:

$$DPM = m_{d_i} - m_{d_j}; d_i, d_j \in D$$

$$DPM = \frac{m_{d_i}}{m_{d_j}}; d_i, d_j \in D$$

where m can be FtER, FtAR, or transaction durations and D is the set of all demographic groups.

6.8 Calculating demographic differentials for continuous variables

Demographic variables can be continuous. One approach to assessing continuous demographic variables is to divide their values into demographic groups (e.g. quartiles), converting them to categorical demographic variables. If a continuous demographic variable is converted into a categorical demographic variable for analysis, a statement shall be prepared documenting the specific reasons for determination of group boundaries and regarding potential inferential errors this can potentially cause (see 7.3). Once a continuous demographic variable is converted into a categorical demographic variable, techniques discussed in 6.4 – 6.7 can be used to compute differential performance.

Continuous demographic variables are also appropriate for analysing via multivariate regression modelling. The magnitude of demographic differentials for continuous demographic variables can be measured by calculating regression coefficients. Multivariate regression modelling seeks to establish the relationship between input variables (including continuous demographic variables) and output variables (such as comparison scores) according to:

$$y \sim X\beta + \varepsilon$$

where:

y is the vector of performance observations;

X is the design matrix for demographic variables per subject;

β is a vector of regression coefficients;

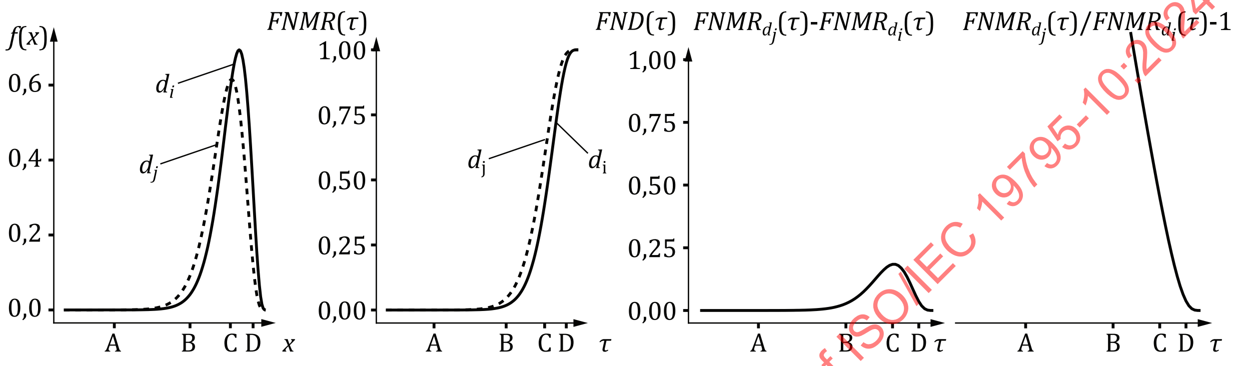
ε is a vector of error terms.

In this model, the design matrix, X, can encompass both linear and non-linear functions of demographic variables. Once regression coefficients have been calculated, model selection or deriving the optimal model from a list of candidates can be performed. Model selection is commonly achieved by calculating and optimizing adjusted “R squared” values (i.e. coefficient of determination or R^2) or the Akaike information criterion. For a demonstration of multivariate regression modelling and model selection for a mix of continuous and categorical variables, see Reference [7].

6.9 Comparison score differential measures

In addition to differential performance observed in biometric system outcomes, some demographic studies seek to determine whether demographic groups could, in principle, affect the performance of a biometric system without considering a set threshold. For example, consider the error rate scenarios resulting from different mated score, *x*, and probability density functions, *f*(*x*), in Figure 1, where the choice of the threshold (τ_A, τ_B, τ_C, τ_D) and the differential performance measure (difference vs. ratio) strongly determines the magnitude of observed differential performance. In this case, it is appropriate to compute comparison score differential measures.

NOTE In Figure 1, one is subtracted from the mathematical ratio to place it on the same zero baseline as the mathematical difference.



Key

A Example threshold, τ, for which the mathematical ratio differential is invalid because $FNMR_{d_i}(\tau) = 0$, due to insufficient sample size, for example. The mathematical difference is near zero.

B Example threshold, τ, for which the mathematical ratio is much larger than the mathematical difference $\left[FNMR_{d_j}(\tau)/FNMR_{d_i}(\tau)-1 \gg FNMR_{d_j}(\tau)-FNMR_{d_i}(\tau)\right]$, even as $FNMR_{d_i}(\tau)\to 0$ and $FNMR_{d_j}(\tau)\to 0$.

C Example threshold, τ, for which the mathematical difference is maximal.

D Example threshold, τ, for which the mathematical difference and the mathematical ratio are comparable $\left[FNMR_{d_j}(\tau)/FNMR_{d_i}(\tau)-1 \approx FNMR_{d_j}(\tau)-FNMR_{d_i}(\tau)\right]$

Figure 1 — Example relationship between comparison score differentials and differential performance for two demographic groups, d_i and d_j

Testers can measure comparison score differential measures by comparing the comparison score distribution means for two groups. If x is a continuous comparison score, $f_{d_i}(x)$ is the score probability density function for demographic group d_i, and $f_{d_j}(x)$ is the score probability density function for demographic group d_j, then the difference in the means of the score distributions, $\Delta\mu_{i,j}$, is:

$$\Delta\mu_{i,j} = \int_{x=-\infty}^{\infty} x\left(f_{d_i}(x)-f_{d_j}(x)\right)dx$$

Testers can also measure comparison score differential measures using the Earth Mover's Distance (EMD), which can be interpreted as the overall work required to transform one distribution into another. If f is a probability density function for a continuous score distribution, and F is the corresponding cumulative distribution function, then for demographic groups d_i and d_j, the EMD is:

$$EMD_{i,j} = \int_{x=-\infty}^{\infty} \left|F_{d_i}(x)-F_{d_j}(x)\right| dx$$

When the probability density function *f* is describing a mated comparison score distribution, the mean difference $\Delta\mu$ and the EMD are mated comparison score differential measures. When the probability density function *f* is describing a non-mated score distribution, the mean difference $\Delta\mu$ and the EMD are non-mated comparison score differential measures.

6.10 Calculating uncertainty

6.10.1 Uncertainty in demographic differentials

This subclause provides requirements and guidance for reporting the variance of a differential and calculating a confidence interval. Differentials by nature include a mathematical operation on two values, each with its own error. Error estimates for demographic differentials can be produced using standard methods of error propagation.

There are two ways of statistically comparing two random measures. The first involves testing to establish whether the difference between the two measures is reliably greater than a certain cutoff. The second involves testing if the ratio between the two measures is greater than a certain cutoff. Though the intent of both approaches is to quantify differences in performance, each has distinct statistical implications. The variance of a biometric measure m, $\mathrm{Var}(m)$ can be calculated for two demographic groups, d_i and d_j. The difference between measures for two groups, $\mathrm{Var}\left(m_{d_i} - m_{d_i}\right)$, can be simply calculated as $\mathrm{Var}\left(m_{d_i}\right) + \mathrm{Var}\left(m_{d_j}\right)$, assuming independence. On the other hand, the variance of the ratio of two measures, $\mathrm{Var}\left(m_{d_i} / m_{d_j}\right)$, does not have a closed formula, but given assumptions that *m* is distributed normally and the sample size is large, this variance can be approximated as: $\frac{m_{d_i}^2}{m_{d_j}^2}\left[\frac{\mathrm{Var}\left(m_{d_i}\right)}{m_{d_j}^2} + \frac{\mathrm{Var}\left(m_{d_j}\right)}{m_{d_j}^2}\right]$. Due to the intrinsic volatility of the ratio, larger sample sizes per demographic group ($N_{d_i} \geq 100$) should be used for ratio-based differentials. The 95 % confidence interval around either DPMs based on mathematical differences (see 6.4.2 and 6.7) or DPMs based on error rate ratios (see 6.4.3 and 6.7) can be calculated as $\pm 1.96\sqrt{\mathrm{Var}(DPM)}$.

When a closed form for the variance of a DPM cannot be determined, an estimate of uncertainty can be obtained via the statistical bootstrap procedure (i.e. random sampling with replacement). The basic procedure for computing the bootstrap estimate for the variance of a DPM, $\mathrm{Var}(\mathrm{DPM})$, is as follows:

1) Create a bootstrap sample, *k*, by sampling, with replacement, N_{d_i} trials for each group, d_i where N_{d_i} is the number of trials observed for group d_i.

2) Use the bootstrap sample, *k*, to calculate the estimated biometric system measure m_{k,d_i}^* for each group, d_i.

3) Calculate the estimated differential performance measure DPM_k^* for bootstrap sample, *k*, using the calculated biometric system measures m_{k,d_i} using formulae from 6.5.

4) Repeat steps 1 to 3, for $k = 1, 2, 3\ldots, K$, where *K* is large (typically ≥ 1000).

5) Calculate the variance $\mathrm{Var}\left(DPM^*\right)$ and use as the bootstrap estimate of $\mathrm{Var}(DPM)$.

Additional details for bootstrapping biometric system measures can be found in Reference [14].

6.10.2 Sampling the target population

The target population is generally of primary importance when performing an analysis of demographic differentials in biometric systems. Failure to appropriately sample the population can result in systematic errors in performance estimates for different demographic groups and can mask any true differentials or introduce false differentials.

The rules for sampling the target population help prevent these issues from arising. For tests involving individual demographic variables, these rules shall be followed for each demographic group. For tests involving intersections across demographic variables, the following rules shall be observed for each intersectional group.

— The data subject population recruited for the evaluation or whose samples are used in the evaluation shall be carefully selected to be representative of the demographic groups of interest.

— Any biometric samples utilized for the evaluation shall be collected in similar ways for each demographic group of interest.

— Demographic variables used in the assessment shall be measured the same way for each demographic group of interest.

— The choices and descriptions given for self-reporting demographic categories shall be consistent for each data subject.

— Within each demographic group, other factors known to affect performance (e.g. gender and age) shall be fixed, blocked or counterbalanced where appropriate.

There are different ways to control these factors.[8] These include, in order of least to most external validity:

— fixing these factors at set values (e.g. males between 20 and 30 years of age);

— balancing the groups (i.e. ensuring the same distributions of age and gender within each ethnicity category); or

— mimicking the distribution of data subjects within each group in the broader population of interest.

EXAMPLE 1 If the height of one group is self-reported, then the height of another group will ideally not be measured by the tester. This could introduce a systematic error.

EXAMPLE 2 If intersectional demographic variables across gender and ethnicity are of interest, then other demographic variables within each group (e.g. age and skin lightness) are fixed, balanced or sampled, so as to mimic distributions in the broader population.

EXAMPLE 3 If face samples for one group are captured outdoors using a digital single-lens reflex camera, but samples for another group are captured indoors using a webcam, then any differentials uncovered can be due to differences in acquisition device.

6.10.3 Sample size requirements

Evaluations measuring differential performance or comparison score differential measures of a biometric system should include a representative number of subjects belonging to each identified demographic group. Evaluations that use statistical sampling to make inferences about larger populations should carefully consider the effect magnitude to be measured and the expected performance of the biometric system.

NOTE Evaluations that include the entire population of subjects that are involved with the biometric system do not perform statistical sampling of the subjects. However, the substantial majority of evaluations for performance variation across demographic group do not fall into this category.

EXAMPLE 1 An evaluation to determine whether a biometric system at an airport has false negative differential performance for travellers based on race by conducting 1:1 verification trials for a randomly selected representative set of subjects uses statistical sampling of subjects.

EXAMPLE 2 An evaluation to determine whether a biometric system used by a gym has false negative differentials for current members based on race by conducting 1:1 verification trials for all current members does not use statistical sampling of subjects.

The effect magnitude to be measured and the expected performance of the biometric system shall be considered early in the design process as they can determine whether the evaluation can reasonably meet its objectives. The effect magnitude expresses the size of a differential that the evaluation aims to detect. The expected performance is an a priori estimate of the system's performance measures of interest in the evaluation. The chosen effect magnitude, expected performance and associated rationale shall be stated explicitly in the experimental design documentation (see 7.1). The sample size of the evaluation shall be set based on the effect magnitude, the target reliability of effect magnitude detection, and the expected performance of the system, taking into consideration both type I and type II errors (see Annex A). Measures of differential performance and comparison score differential measures shall not be computed for demographic groups with insufficient samples.

Failure to set an appropriate effect magnitude can result in an underpowered or an overpowered evaluation. In an underpowered evaluation, the sample size can be too small to detect a differential of the desired effect magnitude. In an overpowered evaluation, the sample size can be large enough to find statistically significant demographic differentials related to uncontrolled factors outside the scope of the evaluation. This can cause errors in inference.

Sample size can drive the cost and labour associated with the evaluation, which can be major considerations during planning. If cost is a significant factor, then statistical calculations should be performed to identify reliably detectable effect magnitudes given the sample size that the evaluation budget allows. These effect magnitudes should be stated explicitly in the test plan and reported (see 7.1).

Sample size constraints can require testers to make trade-offs with respect to their choice of demographic groups. Some demographic groups can be difficult to recruit or may have insufficient samples represented in available test datasets. Sample sizes are more likely to be small for intersections. Testers should be aware that the sample sizes for intersections can be too small to reliably detect the desired effect magnitude. Testers shall choose groups with adequate sample sizes to detect the desired effect magnitude and target reliability. For example, when appropriate, testers can combine demographic groups to generate groups with larger sample size (e.g. combining age groups of “20-30” and “31-40” into a single “20-40” age group).

7 Reporting the evaluation results

7.1 Reporting the experimental design

This subclause establishes reporting requirements for a biometric evaluation assessing performance variation across demographic groups. The test plan shall be reported (see 5.1). Any deviations from the test plan encountered during evaluation execution shall also be documented.

7.2 Reporting the target application

The target application in an evaluation on demographic differentials shall be reported. Target application consists of two parts: the target use case and the target system component under evaluation.

Biometric systems often have different target applications. Systems can be designed to enrol individuals, to verify a claimed identity or to identify individuals on an allow list or a deny list. The target application of the system can determine the kind of demographic differential that is of primary concern. For example, in an access control scenario, higher rates of false negative results mean that a particular demographic group is more likely to be denied access to a resource or facility. However, in a law enforcement context, these higher rates of false negative results mean that a particular demographic group is less likely to be identified if searched against a database. This can result in criminal activity continuing without a subject being identified. Studies of demographic differential performance in biometric systems shall clearly state the differential performance measures under evaluation and the consequences of these differentials in the context of the target application.

7.3 Reporting the test population

The test population in an evaluation of demographic differentials shall be reported, including the overall number of subjects in the evaluation.

For all evaluations, the following shall be reported for the probe or capture subject set. In verification evaluations, the following shall also be reported for the reference set used in comparison trials. In identification evaluations, the following shall also be reported for the demographic reference database used in identification trials:

- demographic variables and groups of interest;

- demographic group size;
- any statements required for demographic variables listed in 5.2.2 and 5.2.3;
- how the demographic groups of each subject were established (e.g. self-report or measured);
- number of subjects for whom demographic data changed across longitudinal collections.

Additionally, in identification evaluations the rationale for the selection of the size and demographic composition of the demographic reference database used in identification trials shall be reported (see 6.6).

7.4 Reporting differential performance

7.4.1 Reporting differential performance on previously collected datasets.

This subclause establishes reporting requirements for new meta-analyses and analyses using previously collected samples and demographic data. There can be significant value in studying performance variation based on previously collected datasets and previously conducted biometric performance tests. However, such datasets can be annotated in a fashion that does not conform to the requirements outlined in Clause 5. Evaluations based on historical data collections that deviate from the requirements outlined in this document shall detail all areas of non-conformity. Such evaluations shall also describe potential inferential errors that can result from these deviations.

7.4.2 Reporting differential performance for two or more groups

The following shall be stated when reporting the performance of demographic differentials between two or more demographic groups:

- the selected demographic groups of interest (see 5.2);
- the sample sizes for the demographic groups of interest and rationale for deviations from the target sample size set in the evaluation;
- the reliably detectable effect magnitude given the sample sizes for the demographic groups of interest;
- in evaluations of verification or identification differential performance, the threshold selected in the evaluation and the rationale for its selection (see 6.3);
- in evaluations of identification trials, the demographic reference database used. This shall be specified by reporting the overall number of references enrolled, N , as well as by the number of enrolled references belonging to each demographic group, $N_{d_1}, N_{d_2}, \ldots$ (see 6.6)];
- the values of the performance measures of interest by demographic group;
- the resulting values for the selected differential performance measure (see 6.4 to 6.9);
- the process, rationale and a description of the impact of large disparities in sample size across demographic groups if replacing error rates with epsilon (see 6.5.2);
- the confidence interval of the differential performance measure and the procedure used in its calculation (see 6.10.1);
- whether the demographic differentials are statistically significant or operationally significant in a given target application. Reporting on this requirement will use the effect magnitude and confidence interval calculations;
- if the analyses and inductions supported or refuted the original model or hypothesis of the evaluation (see 5.1).

7.4.3 Reporting differential performance against a benchmark

Measures of demographic differentials in 6.4 to 6.8 are designed to compare biometric performance across two or more demographic groups. In some evaluations, testers seek to compare the performance of a single demographic group to a specific benchmark. In this case, the denominators in 6.4.3 and 6.7 (e.g. $FMR_{d_j}(\tau)$ and $FNMR_{d_j}(\tau)$) can be replaced with a benchmark performance level, B. Statistical tests can then be run to calculate if all differential performance measures (e.g. $FPD(\tau), FND(\tau)$) are statistically significant. In this type of evaluation, the tester shall report all criteria in 7.4.2 as well as the benchmark performance level B and the statistical test used.

7.4.4 Reporting error trade-off metrics

Error trade-off metrics are common in studies of overall (i.e. non-demographic) biometric recognition error rate performance. For example, detection error trade-off (DET) curves are often used to compare algorithm performance across different thresholds. However, because error trade-off metrics are plotted parametrically on threshold, they mask whether false negative differentials, false positive differentials, or both differentials are varying across demographics. DET curves may be used for illustrative purposes but shall not be used to report measured differentials in the context of demographic differential performance evaluations. The tester can tabulate or plot multiple error rates and false positive and false negative differentials as a function of the operating threshold.

EXAMPLE The tester tabulates error rates and differentials at selected threshold values and visualizes these values on a DET curve.

7.4.5 Reporting threshold management policy

Most biometric systems are configured with a fixed threshold that is not adjusted to account for environmental, demographic or other conditions. Some systems can vary their threshold setting depending on biographic data presented by the capture subject (e.g. in a passport) or with measured environmental conditions. In evaluations of verification or identification differential performance for systems varying thresholds across demographic groups, the evaluation report shall state:

— how demographic group membership is established by the system;

— how the threshold is changed per group;

— false positive and false negative performance values at each group's respective threshold;

— false positive and false negative performance differentials at each group's respective threshold;

— demographic groups of the subjects.

The evaluation report shall also report false positive and false negative differentials for each demographic group at the groups' respective thresholds.

Setting thresholds per demographic group creates two issues. First, this approach constitutes differential treatment. The evaluation report shall state that the system exercises a differential treatment policy. Second, this approach introduces the potential for demographics determined by the system to not match demographics determined by the tester. The evaluation report shall describe any mismatches in demographic group between the system and those determined according to methods in this document.

NOTE In cases where a system modifies the threshold per user-presented biographic documents, an informed attacker can potentially effect adverse changes in security levels, primarily by reducing FMR.

7.5 Reporting comparison score differential measures

Demographic studies that seek to report on differences in score distributions (i.e. comparison score differential measures) should do so using methods described in 6.9. Additionally, there are other standard statistical methods that can be applied to test for differences in biometric comparison score distributions such as:

- the Z-test, where these distributions can be shown to approximate the normal distribution;
- the Komogorov-Smirnov test, where these distributions deviate from the normal, as is often the case when distributions have long or skewed tails.

7.6 Reporting exception handling

Reporting exception handling is specific to scenario and operational evaluations of biometric systems. These evaluations should explicitly track and report on any exception handling protocols and the rate at which these protocols are required across demographic groups. For instance, females may opt-out of using a system at different rates than males in an operational test.