NPL Report CMSC 14/02

Best Practices in Testing and Reporting Performance of Biometric Devices

Version 2.01

By
A. J. Mansfield,
National Physical Laboratory
and
J. L. Wayman,
San Jose State University

August 2002

Centre for Mathematics and Scientific Computing
National Physical Laboratory
Queens Road
Teddington
Middlesex
TW11 0LW

Tel 020 8943 7029
Fax 020 8977 7091

ISSN 1471-0005

Authorised by Dr Dave Rayner
Head of Centre for Mathematics and Scientific Computing

National Physical Laboratory
Queens Road, Teddington, Middlesex, TW11 0LW

Best Practices in Testing and Reporting Performance of Biometric Devices

Produced for the Biometrics Working Group
By
A. J. Mansfield, NPL
J. L. Wayman, SJSU

SUMMARY

The purpose of this document is to summarise the current understanding by the biometrics community of the best scientific practices for conducting technical performance testing toward the end of field performance estimation.

Our aims are:

- To provide a framework for developing and fully describing test protocols.
- To help avoid systematic bias due to incorrect data collection or analytic procedures in evaluations.
- To help testers achieve the best possible estimate of field performance while expending the minimum of effort in conducting their evaluation;
- To improve understanding of the limits of applicability of test results and test methods.

This document is a revision of the original version "Best Practices in Testing and Reporting Performance of Biometric Devices – Issue 1", released in February 2000. This new issue is informed by several important developments in the intervening two years:

- We have had extensive comments on the original document from the biometric community. We are greatly indebted to all of those who commented on Issue 1. Even this new issue remains a living document, open to review and correction by the biometrics community. We encourage written comments and criticism on all aspects of the organisation and content.
- We conducted and completed the CESG/NPL Biometric Test Programme of six biometric technologies and seven vendor products, learning much about the shortfalls of the Issue 1 document.
- We have also noted the protocols and results of other recent test programmes.
- We have made some headway in understanding the statistical relationship between test size and confidence intervals.

Issue 1 drew heavily on "The Speaker Verification Test Protocol" and "An Introduction to Testing Biometric Systems", both produced by NIST. This Issue 2 is intended to be fully compatible with those documents, while integrating their salient points more clearly and more seamlessly than in Issue 1.

CONTENTS

1 INTRODUCTION

1. This document is a revision of "Best Practices in Testing and Reporting Performance of Biometric Devices – Issue 1", released in February 2000 [1]. The new issue is informed by several important developments in the intervening two years:
   a. We have had extensive comments on the original document from the biometric community;
   b. We conducted and completed the CESG/NPL biometric test programme [2] of six biometric technologies and seven vendor products, learning much about the shortfalls of the Issue 1 document;
   c. We have noted the protocols and results of other recent test programmes [3-6]; and
   d. We have made some headway in understanding the statistical relationship between of test size and confidence intervals.

2. We are greatly indebted to all of those who commented on Issue 1. Even this new issue remains a "living" document, open to review and correction by the biometrics community. We encourage written comments and criticism on all aspects of the organisation and content.

3. Issue 1 drew heavily and directly from two primary source documents developed by the (U.S.) National Institute of Standards and Technology (NIST): The "Speaker Verification test protocol" [7] and "An Introduction to Testing Biometric Systems" [8]. This Issue 2 is intended to be fully compatible with those documents, while integrating their salient points more clearly and more seamlessly than in Issue 1.

4. This document is concerned solely with scientific "technical performance testing" of biometric systems and devices. In technical performance testing we seek to determine error and throughput rates, with the goal of understanding and predicting real-world error and throughput performance of biometric systems. By error rates we include both false positive and false negative decisions and additionally failure-to-enrol and failure-to-acquire rates across the test population. By throughput rates, we mean the number of users processed per unit time based both on computational speed and human-machine interaction. These measures are generally applicable to all biometric systems and devices. Technical performance tests that are device-specific—for example, fingerprint scanner image quality or the impact of data compression on error rates—are not considered here.

5. We acknowledge that technical performance testing is only one form of biometric testing. Other types of testing not considered in this document, but possibly more important, include:
   a. Reliability, availability and maintainability;
   b. Vulnerability;
   c. Security;
   d. User acceptance;
   e. Human factors;
   f. Cost/benefit;
   g. Privacy regulation compliance.

   Methods and philosophies for these other types of tests are currently being considered internationally by a broad range of groups. This allows us to focus this document quite narrowly.

6. The purpose of this document is to summarise the current understanding by the biometrics community of the best scientific practices for conducting technical performance testing toward the end of field performance estimation. Such a document is necessary because even a short review of the technical literature on biometric device testing over the last two decades or more reveals a wide variety of conflicting and contradictory testing protocols [9-14]. Even single organisations have produced multiple tests, each using a different test method. Test protocols have varied not only because test goals and available data are different from one test to the next, but also because, prior to Issue 1 of this document, guidelines had not existed for protocol creation.

7. Biometric technical performance testing can be of three types: technology, scenario, or operational evaluation. Each type of test requires a different protocol and produces different results. Even for tests of a single type, the wide variety of biometric devices, sensors, vendor instructions, data acquisition methods, target applications and populations makes it impossible to present precise uniform testing protocols. On the other hand, there are some specific philosophies and principles that can be applied over a broad range of test conditions.

8. We recognise that sometimes it will not be possible to follow best practice completely. However, we hope the guidelines highlight the potential pitfalls, making it easier for testers to explain reasons for any deviation and the likely effect on results.

9. This document is organised in the following manner: definitions, test planning, data collection, and data analysis. Each section will discuss each form of test (technology, scenario, operational) independently.

2 DEFINITIONS

2.1 Components of a biometric system

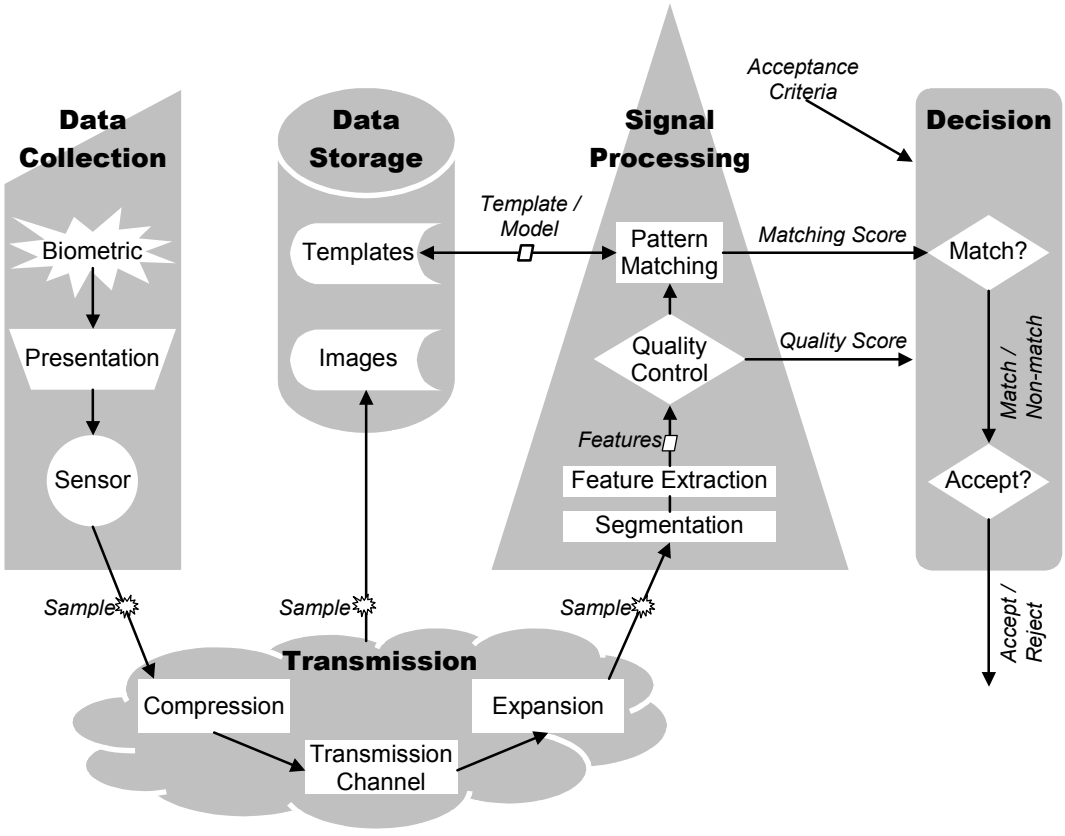

Figure 1 Diagram of general biometric system

10. **Sample:** A biometric measure presented by the user and captured by the data collection subsystem as an image or signal. (E.g. fingerprint, face and iris images are samples.)

11. **Features:** A mathematical representation of the information extracted from the presented sample by the signal processing subsystem that will be used to construct or compare against enrolment templates. (E.g. minutiae coordinates, principal component coefficients and iris-codes are features.)

12. **Template / Model:** A user's stored reference measure based on features extracted from enrolment samples. The reference measure is often a "template" comprising the

biometric features for an ideal sample presented by the user. More generally, the stored reference will be a "model" representing the potential range of biometric features for that user. In these best practices, we shall normally use "template" to include "model".

13. **Matching score:** A measure of the similarity between features derived from a presented sample and a stored template**,** or a measure of how well these features fit a user's reference model. A match / non-match decision may be made according to whether this score exceeds a decision threshold.

14. **Decision:** A determination of the probable validity of a user's claim to identity/non-identity in the system.

15. **Transaction:** An attempt by a user to validate a claim of identity or non-identity by consecutively submitting one or more samples, as allowed by the system decision policy.

16. **Presentation effects:** A broad category of variables impacting the way in which the users' inherent biometric characteristics are displayed to the sensor. For example: in facial recognition, pose angle and illumination; in fingerprinting, finger rotation and skin moisture. In many cases, the distinction between changes in the fundamental biometric characteristic and the presentation effects may not be clear (e.g., facial expression in facial recognition or pitch change in speaker verification systems).

17. **Channel effects:** The changes imposed upon the presented signal in the transduction and transmission process due to the sampling, noise and frequency response characteristics of the sensor and transmission channel.

2.2 Types of evaluation

18. Phillips et al [8] define three basic types of evaluation of biometric systems: (a) technology evaluation; (b) scenario evaluation; and (c) operational evaluation.

19. **Technology evaluation:** The goal of a technology evaluation is to compare competing algorithms from a single technology. Testing of all algorithms is carried out on a standardised database collected by a "universal" sensor. Nonetheless, performance against this database will depend upon both the environment and the population in which it is collected. Consequently, the "Three Bears" rule[1] might be applied, attempting to create a database that is neither too difficult nor too easy, but is "just right" for the algorithms to be tested. Although sample or example data may be distributed for developmental or tuning purposes prior to the test, the actual testing must be done on data that has not previously been seen by algorithm developers. Testing is carried out using offline processing of the data. Because the database is fixed, the results of technology tests are repeatable.

20. **Scenario evaluation:** The goal of scenario testing is to determine the overall system performance in a prototype or simulated application. Testing is carried out on a complete system in an environment that models a real-world target application of interest. Each tested system will have its own acquisition sensor and so will receive slightly different data. Consequently, care will be required that data collection across all tested systems is in the same environment with the same population. Depending upon the data storage capabilities of each device, testing might be a combination of offline and online comparisons. Test results will be repeatable only to the extent that the modelled scenario can be carefully controlled.

[1] "The Three Bears" rule is named after the traditional English children's story [15] in which Goldilocks enters the vacant cottage of the three bears. Inside, she samples everything she finds. The Great Huge Bear's porridge/chair/bed is too hot/hard/high, the Middle Bear's too cold/soft/low, but the Little Small Wee Bear's things are "Just Right!"

21. **Operational evaluation:** The goal of operational testing is to determine the performance of a complete biometric system in a specific application environment with a specific target population. Depending upon data storage capabilities of the tested device, offline testing might not be possible. In general, operational test results will not be repeatable because of unknown and undocumented differences between operational environments. Further, "ground truth" (i.e. who was actually presenting a "good faith" biometric measure) will be difficult to ascertain.

2.3 Identity claims: genuine & impostor, positive & negative, explicit & implicit

22. Biometric authentication has traditionally been described as being for the purpose of either "verification" or "identification".

23. **Verification:** In verification systems, the user makes a "positive" claim to an identity, requiring a "one-to-one" comparison of the submitted "sample" biometric measure to the enrolled template for the claimed identity.

24. **Identification:** In identification systems, the user makes either no claim or an implicit "negative" claim to an enrolled identity, and a "one-to-many" search of the entire enrolled database is required.

25. However, the terms verification and identification do not fully encompass all biometric authentication applications.

26. **Positive claim of identity:** The user claims (either explicitly or implicitly) to be enrolled in or known to the system. An explicit claim might be accompanied by a claimed identity in the form of a name, or personal identification number (PIN). Common access control systems are an example.

27. **Negative claim of identity:** The user claims (either implicitly or explicitly) not to be known to or enrolled in the system. Enrolment in social service systems open only to those not already enrolled is an example.

28. **Explicit claim of identity:** In applications where there is an explicit claim of identity or non-identity, the submitted sample needs to be matched against just the enrolled template for that identity. The accept/reject decision depends on the result of a one-to-one comparison.

29. **Implicit claim of identity:** In applications where there is an implicit claim of identity or non-identity, the submitted sample may need to be matched against many enrolled templates. In this case the accept/reject decision depends on the result of many comparisons, i.e. a one-to-many search.

30. **Genuine claim of identity:** A user making a truthful positive claim about identity in the system. The user truthfully claims to be him/herself, leading to a comparison of a sample with a truly matching template.

31. **Impostor claim of identity:** A user making a false positive claim about identity in the system. The user falsely claims to be someone else, leading to the comparison of a sample with a non-matching template.

2.4 Performance measures

2.4.1 Decision error rates

32. Biometric performance has traditionally been stated in terms of the decision error rates, viz., "false accept rate" and "false reject rate".

33. **False accept rate:** The expected proportion of transactions with wrongful claims of identity (in a positive ID system) or non-identity (in a negative ID system) that are incorrectly confirmed. A transaction may consist of one or more wrongful attempts dependent upon the decision policy. A false acceptance is often referred to in the mathematical literature as a "Type II" error. Note that "acceptance" always refers to the claim of the user.

34. **False reject rate:** The expected proportion of transactions with truthful claims of identity (in a positive ID system) or non-identity (in a negative ID system) that are incorrectly denied. A transaction may consist of one or more truthful attempts dependent upon the decision policy. A false rejection is often referred to in the mathematical literature as a "Type I" error. Note that "rejection" always refers to the claim of the user.

35. Unfortunately, conflicting definitions are implicit in the literature. Literature on large-scale identification systems often refers to a "false rejection" occurring when a submitted sample is incorrectly matched to a template enrolled by another user. In the access control literature, a "false acceptance" is said to have occurred when a submitted sample is incorrectly matched to a template enrolled by another user. The definitions presented here are intended to resolve this conflict.

36. Decision errors are due to matching errors or image acquisition errors (or, with some systems, binning errors). How these fundamental errors combine to form decision errors depends on (a) whether one-to-one or one-to-many matching is required; (b) whether there is a positive or negative claim of identity; and (c) the decision policy, e.g. whether the system allows multiple attempts.

2.4.2 Matching errors

37. To avoid ambiguity with systems allowing multiple attempts or having multiple templates, we define matching algorithm errors to be those for a **single** comparison of a submitted sample against a **single** enrolled template/model.

38. **False match rate (FMR):** The false match rate is the expected[2] probability that a sample will be falsely declared to match a single randomly-selected "non-self" template. (A false match is sometimes called a "false positive" in the literature.)

39. **Non-self:** Genetically different. It has been noted in the literature [16-18] that comparison of genetically identical biometric characteristics (for instance, between a person's left and right eyes or across identical twins) yields different score distributions than comparison of genetically different characteristics. Consequently, such genetically similar comparisons should not be considered in computing the false match rate.

40. **False non-match rate (FNMR):** The false non-match rate is the expected[2] probability that a sample will be falsely declared not to match a template of the same measure from the same user supplying the sample. (A false non-match is sometimes called a "false negative" in the literature.)

2.4.2.1 The Difference between decision errors and matching errors

41. "False match rate" and "false non-match rate" are not generally synonymous with "false accept rate" and "false reject rate". False match/non-match rates are calculated over the number of comparisons, but false accept/reject rates are calculated over transactions and refer to the acceptance or rejection of the stated hypothesis, whether positive or negative. Further, false accept/reject rates include failure-to-acquire rates.

42. In a positive identification system allowing a maximum of three attempts to be matched to an enrolled template, a false rejection will result with any combination of failures-to-acquire and

[2] For both FMR and FNMR the expectations are those for a user selected randomly from the target population.

false non-matches over three attempts. A false acceptance will result if an image is acquired and falsely matched to an enrolled image on any of three attempts.

43. In a negative identification system, a user's claim not to be enrolled in the system will be falsely rejected if an image is acquired and then falsely matched to one or more enrolled templates. Depending upon system policy, a user's claim might be falsely accepted if an image cannot be acquired or if an acquired image is falsely non-matched against the enrolled image.

44. If each user is allowed one enrolment template and makes the same number (and pattern) of verification attempts, the observed error rates will be the best estimates of the true error rates. Note that the error rates are averaged over users as opposed to attempts. Averaging over attempts would weight the error rates towards those of the heavy users of the system and toward those requiring multiple attempts for acceptance.

2.4.3 Image acquisition errors

45. Regardless of the accuracy of the matching algorithm, the performance of a biometric system is compromised if an individual cannot enrol or if they cannot present a satisfactory image at a later attempt.

46. **Failure to enrol rate:** The "failure to enrol" rate is the expected proportion of the population for whom the system is unable to generate repeatable templates. This will include those unable to present the required biometric feature, those unable to produce an image of sufficient quality at enrolment, and those who cannot reliably match their template in attempts to confirm the enrolment is usable. The failure to enrol rate will depend on the enrolment policy. For example in the case of failure, enrolment might be re-attempted at a later date.

47. **Failure to acquire rate:** The "failure to acquire" rate is defined as the expected proportion of transactions for which the system is unable to capture or locate an image or signal of sufficient quality. The failure to acquire rate may depend on adjustable thresholds for image or signal quality.

2.4.4 Binning algorithm performance

48. To improve efficiency in systems requiring a one-to-many search of the enrolled database, some systems may partition template data to separate "bins". An input sample is likewise assigned to a partition and compared only to the portion of the template data contained in the same partition.

49. **Penetration rate:** The penetration rate is defined as the expected proportion of the template data to be searched over all input samples under the rule that the search proceeds through the entire partition regardless of whether a match is found. Lower penetration rates indicate fewer searches and, hence, are desirable.

50. **Binning error rate:** A binning error occurs if the enrolment template and a subsequent sample from the same biometric feature on the same user are placed in different partitions. In general, the more partitioning of the database that occurs the lower the penetration rate, but the greater the probability of a binning error.

2.5 Genuine and Impostor attempts

51. The careful definition of "genuine" and "impostor" transactions forms an important part of our test philosophy and can be used to resolve unusual test situations. These definitions are independent of the type of test being performed.

52. **Genuine attempt:** A "genuine" attempt is a single good faith attempt by a user to match his or her own stored template.

53. **Impostor attempt:** An "impostor" attempt is a single "zero-effort" attempt, by a person "unknown to the system", to match a stored template.

54. **Unknown to the system:** A person is known to the system if (a) the person is enrolled; **and** (b) the enrolment affects the templates of others in the system. An enrolled person can be considered unknown with reference to others in the system only if the other templates are independent and not influenced by this enrolment. Eigenface systems using all enrolled images for creation of the basis-images and cohort-based speaker recognition systems are two examples for which templates are dependent. Such systems cannot treat any enrolled person as unknown with reference to the other templates.

55. **Zero-effort attempts:** An impostor attempt is classed as "zero-effort" if the individual submits their own biometric feature as if they were attempting successful verification against their own template. In the case of dynamic signature verification, an impostor would therefore sign his or her own signature in a zero-effort attempt. In such cases, where impostors may easily imitate aspects of the required biometric, a second impostor measure based on "active impostor attempts" may be required. However, defining the methods or level of skill to be used in active impostor attempts is outside the scope of this document.

56. Stored templates, used in both impostor and genuine transactions, are acquired from users making good faith attempts to enrol properly, as explicitly or implicitly defined by the system management.

2.6 Online and offline generation of matching scores

57. Testing a biometric system will involve the collection of input images or data, which are used for template generation at enrolment, and for calculation of matching scores at later attempts. The images collected can either be used immediately for an online enrolment or identification attempt, or may be stored and used later for offline enrolment and identification.

58. **Online:** Enrolment or calculation of matching scores is said to be "online" when it is done at the time the image or signal is submitted. This has the advantage that the biometric sample can be immediately discarded, saving the need for storage and for the system to operate in a manner different from usual. However, it is recommended that images are collected if possible.

59. **Offline:** Enrolment or calculation of matching scores is said to be "offline" when it is based on images or signals collected earlier. Collecting a database of images for offline enrolment and calculation of matching scores allows greater control over which attempts and template images are to be used in any transaction. Regardless of test type, offline testing can be more appropriate than online testing in several circumstances mentioned later in this best practice document.

60. Technology testing will always involve data storage for later, offline processing. However, with scenario and operational testing, online transactions might be simpler for the tester: the system is operating in its usual manner, and (although recommended) storage of images is not absolutely necessary.

2.7 DET & ROC curves

61. **Receiver operating characteristic (ROC) curves:** Receiver operating characteristic curves are an accepted method for summarising the performance of imperfect diagnostic, detection, and pattern matching systems. An ROC curve plots, parametrically as a function of the decision threshold, the rate of "false positives" (i.e. impostor attempts accepted) on the x-axis, against the corresponding rate of "true positives" (i.e. genuine attempts accepted) on the y-axis. ROC curves are threshold independent,

allowing performance comparison of different systems under similar conditions, or of a single system under differing conditions. Figure 2 shows an example.

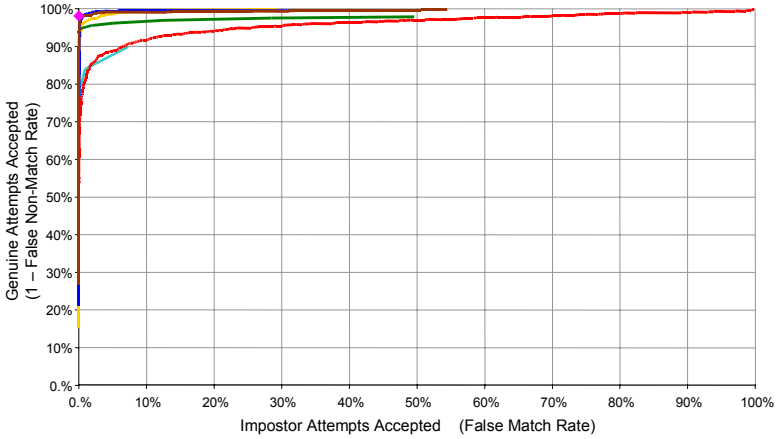

Figure 2 Example ROC curves

62. **Detection error trade-off (DET) curves:** In the case of biometric systems, a modified ROC curve known as a "detection error trade-off" curve [19] is preferred. A DET curve plots error rates on both axes, giving uniform treatment to both types of error. The graph can then be plotted using logarithmic axes. This spreads out the plot and distinguishes different well-performing systems more clearly. For example the DET curve in Figure 3 uses the same data as the ROC curve in Figure 2. DET curves can be used to plot matching error rates (false non-match rate against false match rate) as well as decision error rates (false reject rate against false accept rate).

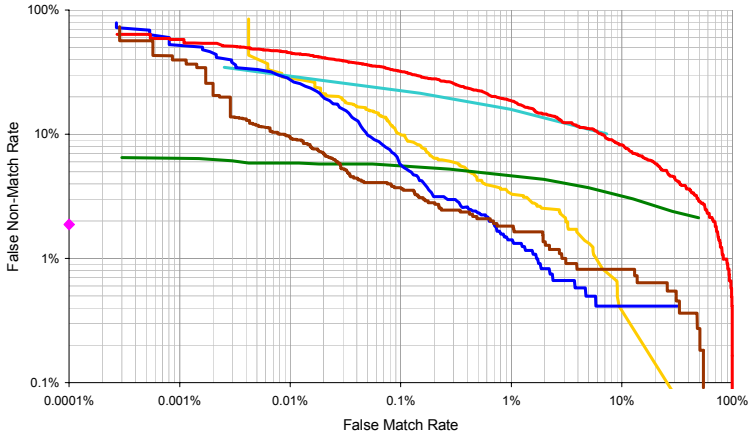

Figure 3 Example DET curves

2.8 Statistical terms

63. **Variance:** The variance is a measure of the spread of a statistical distribution. If μ is the mean of the distribution for a random variable X, then $V(X) = E[(X-\mu)^2]$. If known, it can show how close an estimated result is likely to be to its true value.

64. **Confidence interval:** A (95%) confidence interval for parameter x consists of a lower estimate L, and an upper estimate U, such that the probability of the true value being within the interval estimate is the stated value (e.g. Prob(x$\in$[L,U]) = 95%). The smaller the test size, the wider the confidence interval will be.

65. **Type I error:** Rejecting a true hypothesis.

66. **Type II error:** Accepting a false hypothesis.

3 PLANNING THE EVALUATION

67. The first steps in planning an evaluation will be to decide:

a. What exactly is the evaluation trying to determine?

b. Which is the appropriate evaluation type: technology, scenario, or operational?

These two decisions will form the basis for developing an appropriate test protocol, specifying appropriate environmental controls, volunteer selection, test size, etc. Choice of evaluation type might be determined by, for example, the ready availability of databases of test samples for a technology evaluation, or of an installed system for an operational evaluation. There might also be circumstances in which all three types of testing would be carried out sequentially, perhaps gradually narrowing down technology options and systems under consideration for the eventual deployment of a biometric identification system.

3.1 Determine information about the system etc.

68. Information about the system(s) to be tested is needed to determine the appropriate data collection procedures etc.

a. Does the system log transaction information? If not, then this information will need to be recorded manually by the volunteer or supervisor.

b. Does the system save sample images or features for each transaction? This will be necessary if matching scores are to be generated offline.

c. Does the system return matching scores or just accept/reject decisions? In the latter case we may have to collect data at a variety of security settings to generate a DET curve.

d. Does the system generate independent templates? The correct procedures for collecting or generating impostor transactions are different if templates are dependent.

e. Is the vendor's "Software developer's kit" (SDK) available? Offline generation of genuine and impostor matching scores will require use of software modules from the SDK: (i) for generation of enrolment templates from enrolment images; (ii) for extracting sample features from the test images; and (iii) for generating the matching scores between sample features and templates. Matching scores produced by the offline codes should be equal to those produced by the live system. This may involve adjustment of parameters.

f. What are the recommended image quality and matching decision thresholds for the target application? These settings will affect the quality of presented samples.

g. Do we approximately know what error rates are expected? This information can help in determining whether the test size is appropriate.

h. What are the factors that will influence performance for this type of system?

69. **Scenario & operational evaluation**: In scenario and operational testing any adjustments to the devices and their environment for optimal performance (including quality and decision thresholds) will need to take place prior to data collection. This should be done in consultation with the vendor. For example, stricter quality control can result in fewer false matches and false non-matches, but a higher failure to acquire rate. The vendor is probably best placed to decide the optimal trade-off between these figures. The decision threshold also needs to be set appropriately if matching results are presented to the user: positive or negative feedback will affect user behaviour.

3.2 Controlling factors that influence performance

70. When planning a biometric performance evaluation, factors influencing the measured performance are explicitly or implicitly divided into one of four classes:

a. Factors incorporated into the structure of the experiment (as independent variables) so that we can observe the effect the factors may have;

b. Factors controlled to become part of the experimental conditions (unchanging throughout the evaluation);

c. Factors "randomised out" of the experiment;

d. Factors judged to have negligible effect, which will be ignored. Without this final category the experiment would become unnecessarily complex.

71. Performance figures can be very application, environment and population dependent, and these aspects should therefore be decided in advance. This may involve some preliminary testing of systems to determine which factors are most significant and which may be safely ignored. Appendix A provides a list of many user and environmental factors that have been found to affect performance of one or more types of biometric system.

72. In determining which factors to control, there may be a conflict between the needs for "internal validity" (i.e. that differences in performance are due only to the independent variable(s) recorded in the study) and "external validity" (i.e. that the results truly represent performance on the target application).

73. **Technology evaluation:** For technology testing, a "generic" application and population might be envisioned, applying the "Three-Bears" rule, so that the tests are neither too hard nor too easy for the systems being evaluated.

74. **Scenario evaluation:** For scenario testing, a real-world application and population might be imagined and modelled in order that the biometric device can be tested on representative users, in a realistic environment.

75. **Operational evaluation:** In operational testing, the environment and the population are determined *in situ* with little control over them by the experimenter.

3.3 Volunteer selection

76. Both the enrolment and transaction functions require input signals or images. These samples must come originally from a test population, or "crew". We do not accept as best practice the generation of artificial images[3] (or the generation of new images by changing data from real images).

77. **Scenario evaluation:** For scenario evaluations, this crew should be demographically similar to that of the target application for which performance will be predicted from test results. This will be the case if the test population can be randomly selected from the potential users for the target application. In other cases we must rely on volunteers.

78. **Operational evaluation:** In the case of operational testing, the experimenter may have no control over the users of the system.

79. Enrolment and testing will be carried out in different sessions, separated by days, weeks, months or years, depending upon the target application. A test crew with stable membership over time is difficult to find, and it should be expected for some volunteers to "drop-out" between enrolment and testing.

80. Recruiting the crew from volunteers may bias the tests. People with unusual features, the regularly employed, or the physically challenged, for instance, may be under-represented in the sample population. Those with the strongest objections to the use of the biometric technology are unlikely to volunteer. It may be necessary to select unevenly from volunteers in order that the volunteer crew is as representative as possible, and does not under-represent known problem cases. Our understanding of the demographic factors affecting biometric

[3] The use of artificially generated images would improve the "internal validity" of technology evaluations, as all the independent variables affecting performance are controlled. However, "external validity" is likely to be reduced. Moreover, the database is likely to be biased in respect of systems that model the biometric images in a similar way to that used in their generation.

system performance is so poor that target population approximation will always be a major problem limiting the predictive value of our tests.

81. The volunteer crew must be fully informed as to the required data collection procedure, must be aware of how the raw data will be used and disseminated, and must be told how many sessions will be required and the durations of those sessions. Regardless of the use of the data, the identities of the crew should never be released. A consent form acknowledging that each volunteer understands these issues must be signed, then maintained in confidence by the researchers. A sample consent form is included as Appendix C.

82. **Technology & scenario evaluation**: Volunteers in technology and scenario evaluations should be appropriately motivated so that their behaviour follows that of the target application. If volunteers become bored with routine testing, they may be tempted to experiment, or be less careful. Such possibilities must be avoided.

3.4 Test size

83. The size of an evaluation, in terms of the number of volunteers and the number of attempts made (and, if applicable, the number of fingers/hands/eyes used per person) will affect how accurately we can measure error rates. The larger the test, the more accurate the results are likely to be.

84. Rules such as the "Rule of 3" and "Rule of 30", detailed below, give lower bounds to the number of attempts needed for a given level of accuracy. However, these rules are over-optimistic, as they assume that error rates are due to a "single source of variability", which is not generally the case with biometrics. Ten enrolment-test sample pairs from each of a hundred people is not statistically equivalent to a single enrolment-test sample pair from each of a thousand people, and will not deliver the same level of certainty in the results[4].

3.4.1 Rule of 3

85. The "Rule of 3" [21-24] addresses the question "What is the lowest error rate that can be statistically established with a given number N of (independent identically distributed[5]) comparisons?" This value is the error rate p for which the probability of zero errors in N trials, purely by chance, is (say) 5%. This gives:

$$p \approx 3/N \text{ for a 95\% confidence level.}^{6}$$

So, for example, a test of 300 independent samples returning no errors can be said with 95% confidence to have an error rate of 1% or less.

3.4.2 Rule of 30

86. Doddington [7] proposes the "Rule of 30" for helping determine the test size:

> To be 90% confident that the true error rate is within ±30% of the observed error rate, there must be at least 30 errors.[7]

---

[4] As the test size increases, the variance of estimates decrease, but the scaling factor depends on the source of variability. For example, volunteers may have differing error rates [20], giving a component of variance that scales as 1/(the number of volunteers) instead of 1/(number of attempts). This effect is discussed in more detail in Appendix B.

[5] The assumption of independent identically distributed (i.i.d.) attempts may be achieved if each genuine attempt uses a different volunteer, and if no two impostor attempts involve the same volunteer. With n volunteers, we would have n genuine attempts and $n/2$ impostor attempts. However, cross-comparisons between all submitted samples and enrolled templates generates many more impostor attempts and, according to [25], achieves smaller uncertainty despite dependencies between the attempts. Thus, except perhaps in the case of operational testing, there is little merit in restricting data to a single attempt per volunteer to achieve the i.i.d. assumption.

[6] $p \approx 2/N$ for a 90% confidence level.

So, for example, if we have 30 false non-match errors in 3,000 independent genuine trials, we can say with 90% confidence that the true error rate is between 0.7% and 1.3%

87. The rule comes directly from the binomial distribution assuming independent trials, and may be applied by considering the performance expectations for the evaluation. For example, suppose the performance goals are a 1% false non-match rate, and a 0.1% false match rate. This rule implies 3,000 genuine attempt trials, and 30,000 impostor attempt trials. Note however the key assumption that these trials are independent. Strictly speaking this would require 3,000 enrolees, and 30,000 impostors. The alternative is to compromise on independence by re-using a smaller set of volunteers, and to be prepared for a loss of statistical significance.

3.4.3 Collecting multiple transactions per person

88. If we did not have to consider the cost and effort in obtaining and enrolling volunteers, the ideal test would have many volunteers, each making a single transaction. However in real life, it is significantly easier to get existing enrolees to return than to find and enrol new volunteers. Moreover, whenever an attempt is made, with marginally more effort we can collect several additional attempts at the same time. For example, an evaluation might use 200 volunteers, each enrolling and making three genuine transactions on two further occasions, giving 1200 genuine (though not fully independent) attempts. Instead, with the same effort, the evaluation might have used 240 volunteers, each making a single genuine transaction, giving 240 independent genuine attempts. In this case using multiple transactions allows a six-fold increase in genuine attempts, and a four-fold increase in impostor (cross-comparison) attempts.

89. User behaviour may vary with each successive attempt due to increased familiarity with the device or feedback of their authentication results. For example, the first attempt a user makes might have a higher failure rate than any following attempts. As a result the observed false non-match rate will depend on the pattern of attempts per user, as defined by the test protocol. Generally we will be trying to measure error rates averaged not only over the target population, but also over the types of attempt a user might reasonably make. Averaging over multiple attempts can help in this case. However, we should be alert to the possibility that altering the number and pattern of attempts per user might significantly affect the measured error rates.

3.4.4 Recommendations on test size

90. The number of people tested is more significant than the total number of attempts in determining test accuracy. Our general recommendation therefore is:

a. Firstly, the crew should be as large as practicable. The measure of practicality is likely to be the expense of crew recruitment and tracking.

b. Then, collect sufficient samples per volunteer so that the total number of attempts exceeds that required by the Rule of 3 or Rule of 30 as appropriate. If it is possible to collect these multiple samples on different days, or from different fingers, eyes or hands (and the additional samples are still representative of normal use[8]), doing so can help reduce the dependencies between samples by the same person.

---

[7] The rule generalises to different proportional error bands, for example:
To be 90% confident that the true error rate is within ±10% of the observed value, we need at least 260 errors; and
To be 90% confident that the true error rate is within ±50% of the observed value, we need at least 11 errors.

[8] For example, use of the little finger is probably not representative of normal use of a fingerprint system, and the resulting error rates will be different [26]. Similarly an inverted left hand would not be representative in a right-handed hand geometry system.

c. Finally, once data has been collected and analysed, it may be possible to estimate the uncertainty in the observed error rates, and determine whether the test was indeed large enough.

91. We should also note that "the law of diminishing returns" applies. A point will be reached where errors due to bias in the environment used, or in volunteer selection, will exceed those due to size of the crew and number of tests.

3.5 Multiple tests

92. The cost of data collection is so high that we are tempted to create technical evaluation protocols so that multiple tests can be conducted with one data collection effort.

93. **Technology evaluation:** In the case of biometric devices for which image standards exist (fingerprint [27][9], face [28], voice [29]), it is possible to collect a single corpus for offline testing of pattern matching algorithms from multiple vendors.

94. In effect, we are attempting to decouple the data collection and signal processing subsystems. This is not problem-free however, as these subsystems are usually not completely independent. For instance the quality control module, which may require the data collection subsystem to reacquire the image, is part of the signal processing subsystem. Further, even if image standards exist, image quality is affected by the vendor-specific user interfaces that guide the data collection process. Consequently, offline technical evaluation of algorithms using a standardised corpus may not give a good indication of total system performance, and also can be biased in favour of some systems and against others.

95. **Scenario evaluation:** Multiple scenario evaluations can be conducted simultaneously by having a volunteer crew use several different devices or scenarios in each session. This approach will require some care. One possible problem is that the volunteers will become habituated as they move from device to device. To equalise this effect over all devices, the order of their presentation to each volunteer must be randomised.

96. A further potential problem occurs where ideal behaviour for one device conflicts with that for another. For example, some devices work best with a moving image, while others require a stationary image. Such conflicts may result in lower quality test images for one or more of the devices under test.

97. **Operational evaluations** do not generally allow for multiple testing from the same collected data set.

4 DATA COLLECTION

4.1 Avoidance of data collection errors

98. Collected biometric image samples or features are properly referred to as a "corpus". The information about those images and the volunteers who produced them is referred to as the "database". Both the corpus and the database can be corrupted by human error during the collection process. In fact, error rates in the collection process may easily exceed those of the biometric device. For this reason, extreme care must be taken during data collection to avoid both corpus (mis-acquired image) and database (mis-labelled image) errors.

99. Some typical corpus errors are:

[9] FBI/NIST "Appendix G: Image Quality Standard for Scanners" [27], although originally written for document scanners used to produce digitised images from inked fingerprint cards, is held as a specification for fingerprint sensor image quality. The dual use of this standard is problematic, particularly for the non-optical fingerprint sensors.

a. Volunteers using the system incorrectly (and outside the limits allowed by the experimental controls), e.g., mistakenly using a fingerprint scanner upside down;
b. Cases where a blank or corrupt image is acquired if the user enters a PIN but moves on before a proper image is captured.

Example database errors are:

c. Volunteers being issued with the wrong PIN;
d. Typing errors in PIN entry;
e. Using the wrong body part, e.g., using a middle finger when the index finger is required.

100. Data collection software minimizing the amount of data requiring keyboard entry, multiple collection personnel to double-check entered data, and built-in data redundancy are required. Supervisors must be familiar with the correct operation of the system, and the possible errors to guard against. To avoid a variable interpretation of what constitutes a mis-acquired image, objective criteria must be set in advance. Any unusual circumstance surrounding the collection effort, and the transactions affected, must be documented by the collection personnel.

101. Even with precautions, some data collection errors are likely to be made, which can add uncertainty to the measured test results. "After-the-fact" database correction will be based upon whatever redundancies are built into the collection system. In this respect, systems that can save sample images and/or transaction logs offer more scope for error correction than systems where all the details have to be recorded manually.

4.2 Data and details collected

102. The data that may be collected will depend on the biometric system implementation. For the purposes of evaluation, the ideal situation is for the systems to save sample images or features, and to automatically log enrolments and transactions including details of claimed identity and matching and quality scores. This brings the following advantages:

a. Enrolment templates and matching scores can be generated offline provided that the vendor SDK is available. This will allow for a full cross-comparison of samples and templates, giving a higher number of impostor scores.
b. The collected images can be re-used, to evaluate algorithm improvements, or (provided the images are in a suitable format) to evaluate other algorithms in a technology evaluation.
c. The ability to check for potential corpus or database errors by visually inspecting the images, or through examining the transaction log.
d. A reduction in the amount of data that needs to be recorded by hand, and the consequent transcription errors.

However, many biometric systems do not provide this ideal functionality in their normal mode of operation. With vendor co-operation it may be possible to incorporate this functionality into an otherwise standard system, but care must be taken that system performance is not affected. For example, the time taken in logging images may slow the system and affect user behaviour. If sample images or features cannot be saved, enrolment, genuine and impostor transactions will have to be conducted online, and results recorded manually if necessary. This will require closer supervision to ensure that all results are logged correctly.

103. Some systems do not return matching scores, just a match / non-match decision at the current security setting.

a. To plot a detection-error trade-off (DET) graph in such cases, genuine and impostor transaction data must be collected or generated at a number of security settings. The vendor should be able to advise on the appropriate range of security settings. The selected values for the security setting (perhaps "low", "medium" and "high") will parameterise the DET curve in place of the decision threshold.

b. In the case of online testing, correct estimation of error rates will require each volunteer to make transactions at each chosen security setting. It is incorrect to assume that a user accepted at a strict setting will always be accepted at a more lenient setting, or that a user rejected at a lenient setting can never be accepted at stricter settings.

104. **Technology evaluation:** In technology evaluations, sample images will be collected, and analysed offline using vendor SDKs. The sample data does not have to be generated through volunteers interacting with a biometric system in its normal operating mode.

105. **Scenario evaluation:** In scenario evaluations, volunteers will use the biometric system(s) being tested in a controlled but realistic environment. Enrolment, and scoring of genuine and impostor transactions may be either online (as the volunteers use the system) or offline (using the saved sample images and the vendor SDK). It is strongly recommended that sample images or features are collected.

106. **Operational evaluation:** In operation evaluations, volunteers will use the operational system being tested in their normal way. It is unlikely that an operational system will save transaction samples, so genuine and impostor scores must be recorded online. If the system provides a transaction log, there may be no need to note by hand the outcome of every attempt. In some cases, the enrolment database can be used to generate impostor scores through inter-template comparison (the circumstances when this is appropriate are discussed in section 4.5.2.2).

4.3 Enrolment and test transactions

107. Each volunteer may enrol only once (though an enrolment may generate more than one template, and multiple attempts at enrolment may be allowed to achieve one good enrolment). Care must be taken to prevent accidental multiple enrolments.

108. **Scenario & operational evaluation**: In scenario and operational evaluations, images may be recorded as a corpus for offline testing or may be input directly into the biometric system for online enrolment. In the latter case we recommend that the raw images used for the enrolment be recorded.

109. In all evaluations, it is acceptable to perform practice tests at the time of enrolment to ensure that the enrolment images are of sufficient quality to produce a later match. Scores resulting from such practice tests must not be recorded as part of the "genuine" comparison record. However, the practice samples can be used in offline generation of impostor transactions.

110. **Technology evaluation**: In technology evaluations, every enrolment must be carried out under the same general conditions. Many data collection efforts have been ruined because of changes in the protocols or equipment during the extended course of collection[10]. The goal should be to control presentation and transmission channel effects so that such effects are either (a) uniform across all enrolees, or (b) randomly varying across enrolees.

111. **Scenario evaluation:** In scenario evaluations, enrolment must model the target application enrolment. The taxonomy [24] of the enrolment environment will determine the applicability of the test results. Obviously, vendor recommendations should be followed and the details of the environment should be completely noted. The "noise" environment requires special care. Noise can be acoustic, in the case of speaker verification, or optical, in the case of eye, face, finger or hand imaging systems.

[10]A famous example is the "Great Divide" in the KING speech corpus [30]. About halfway through the collection, for a reason nobody remembers, the recording equipment had to be temporarily disassembled. It was later re-assembled according to the original wiring diagram; nonetheless the frequency response characteristics were slightly altered, creating a divide in the data and complicating the scientific analysis of algorithms based on the data.

Lighting "noise" is of concern in all systems using optical imaging, particularly any light falling directly on the sensor and uncontrolled reflections from the body part being imaged. Lighting conditions should reflect the proposed system environment as closely as possible. It is especially important to note that test results in one noise environment will not be translatable to other environments.

112. **Operational evaluation:** In operational evaluations the experimenter may have no control over the enrolment conditions. Indeed, the enrolments could have been performed before the period of evaluation commenced.

113. Regardless of evaluation type, the quality control module may prevent acceptance of some enrolment attempts. Quality control modules for some systems requiring multiple images for enrolment will not accept images that vary highly between presentations; other quality control modules will reject single poor quality images. If these modules allow for tuning of the acceptance criteria, we recommend that vendor advice be followed. Multiple enrolment attempts should be allowed, with a pre-determined maximum number of attempts or maximum elapsed time. All quality scores and enrolment images should be recorded. Advice or remedial action to be taken with volunteers who fail an enrolment attempt should be predetermined as part of the test plan. The percentage of volunteers failing to enrol at the chosen criteria must be reported.

114. All quality control may not be automatic. Intervention by the experimenter may be required if the enrolment measure presented was inappropriate according to some pre-determined criteria. For instance, enrolling volunteers may present the wrong finger, hand or eye, recite the wrong enrolment phrase or sign the wrong name. This data must be removed, but a record of such occurrences should be kept.

Technology & scenario evaluation: In technology and scenario evaluations, enrolment data should not be removed simply because the enrolled template is an outlier.

Operational evaluation: In operational evaluations, no information regarding appropriate presentation may be available.

Data editing to remove inappropriate biometric presentations may have to be based on removal of outliers, but the effect of this on resulting performance measures should be fully noted.

115. As the tests progress, an enrolment supervisor may gain additional working knowledge of the system, which could affect the way later enrolments are carried out. To guard against this, the enrolment process and criteria for supervisor intervention should be determined in advance, and adequate supervisor training provided.

4.4 Genuine transactions

116. **Technology evaluation:** For technology evaluations, test data should be collected in an environment that anticipates the capabilities of the algorithms to be tested. Test data should be neither too hard nor too easy to match to the enrolment templates.

117. For technology evaluations, the time interval between the enrolment and the test data will be determined by the desired difficulty of the test. Longer time intervals generally make it more difficult to match samples to templates due to the phenomenon known as "template ageing". Template ageing refers to the increase in error rates caused by time related changes in the biometric pattern, its presentation, and the sensor.

118. **Scenario evaluation:** For scenario evaluations, test data must be collected in an environment, including noise, that closely approximates the target application. For all types of tests, the test environment must be consistent throughout the collection process. Great care must be taken to prevent data entry errors and to document any unusual circumstances surrounding the collection. It is always advisable to minimise keystroke entry on the part of both volunteers and experimenters.

119. For scenario evaluations, test data must be separated in time from enrolment by an interval commensurate with "template ageing" of the target system. For most systems, this interval may not be known. In such cases, a rule of thumb would be to separate the samples at least by the general time of healing of that body part. For instance, for fingerprints, 2 to 3 weeks should be sufficient. Perhaps, eye structures heal faster, allowing image separation of only a few days. Considering a hair cut to be an injury to a body structure, facial images should perhaps be separated by one or two months. In the ideal case, between enrolment and the collection of test data, volunteers would use the system with the same frequency as the target application. However, this may not be a cost effective use of volunteers. It may be better to forego any interim use, but allow re-familiarisation attempts immediately prior to test data collection.

120. **Operational evaluation:** It may be necessary to "balance" the frequency with which volunteers use the system so that results are not dominated by a small group of excessively frequent users.

121. In operational evaluations, it may not be possible to detect data collection errors. Data can be corrupted by impostors or genuine users who intentionally misuse the system. Although every effort must be made by the researcher to discourage these activities, data should not be removed from the corpus unless external validation of the misuse of the system is available.

122. Operational evaluations may allow for the determination of the effects of "template ageing" from the acquired data if the collected data carries a time stamp.

123. With all types of evaluation, test data should be added to the corpus regardless of whether or not it matches an enrolled template. Some vendor software will not record a measure from an enrolled user unless it matches the enrolled template. Data collection under such conditions would be severely biased in the direction of underestimating false non-match error rates. If this is the case, non-match errors must be recorded by hand. Data should be excluded only for predetermined causes independent of comparison scores.

124. **Technology & scenario evaluation:** In both technology and scenario evaluations, the collection must ensure that presentation and channel effects are either: (a) uniform across all volunteers; or (b) randomly varying across volunteers. If the effects are held uniform across volunteers, then the same presentation and channel controls in place during enrolment must be in place for the collection of the test data. Systematic variation of presentation and channel effects between enrolment and test data will obviously lead to results distorted by these factors. If the presentation and channel effects are allowed to vary randomly across test volunteers, there must be no correlation in these effects between enrolment and test sessions across all volunteers.

125. **Operational evaluation:** In operational evaluations the channel and presentation effects are not controlled – they are those of the target application.

126. Specific testing designed to test either user habituation (improving matching scores) or template ageing (degrading scores) will require multiple samples over time. Unless template ageing and habituation occur on different known time scales, there will be no way to de-convolve their counteracting effects.

127. **Scenario & operational evaluation:** The operational thresholds chosen during scenario or operational data collection may influence the performance of volunteers: stricter thresholds encouraging more careful presentation of the biometric pattern, looser thresholds allowing more sloppiness. Therefore the database itself may not be as threshold independent as generally assumed.

128. Sometimes, volunteers may be unable to give a usable sample to the system as determined by either the experimenter or the quality control module. The failure-to-acquire rate measures the

proportion of such transactions, and is quality threshold dependent. As with enrolment, quality thresholds should be set in accordance with vendor advice.

Scenario & operational evaluation: The experimenter should attempt to have the system operators record information on failure-to-acquire transactions where these would otherwise not be logged.

129. All attempts, including failures to acquire, should be recorded. In addition to recording the raw image data, details should be kept of the quality measures for each sample if available and, in the case of online testing, the matching score(s).

4.5 Impostor transactions

130. Impostor transactions can be generated online or offline in several ways.

a. Online impostor transactions involve volunteers submitting samples to be compared against other people's enrolments;

b. Offline impostor transactions are computed by comparing samples already collected as genuine transactions against enrolment templates. Offline computation allows a full cross-comparison approach in which every sample is compared against every non-self template.

131. **Technology evaluation:** In technology testing, impostor transactions are always analysed offline. However, occasionally there may be a database of impostor attempts, to be analysed instead of or in addition to the set of cross-comparison impostor transactions.

132. **Scenario evaluation:** In scenario testing, which method is the most appropriate will probably depend on whether the system is able to save samples from genuine transactions. If so, cross-comparison will generate many more impostor attempts than could be achieved through online use of volunteers.

133. **Operational evaluation**: In operational testing, development of impostor scores may not be straightforward. If the operational system saves sample images or extracted features, impostor scores can be computed offline. If, as is likely, this data is not saved, impostor scores can be obtained through online testing. Because of the non-stationary statistical nature of the data across users, it is preferable to use many volunteer impostors, each challenging one randomly chosen non-self template, than to use a few volunteers challenging many non-self templates. In some cases the use of inter-template comparisons for impostor transactions may be appropriate.

4.5.1 Online collection of impostor transactions

134. Online impostor transactions can be collected by having each user make zero-effort impostor attempts against each of a pre-determined number of non-self templates randomly selected from all previously enrolments. The random selection should be independent between users.

135. In cases where impostor transactions are being collected before all volunteers have enrolled, the first enrolled templates will have a higher probability of being selected for an impostor comparison. However, this will not bias the calculation of impostor error rates if, as is usually the case, volunteers are enrolled in an order that has no regard to the quality of their biometric measures.

136. In a system that uses partitioning, samples should be compared only to templates in the same bin. If the sample and template are in different bins, the transaction will fail without a matching attempt, and there will be no contribution to false match rate estimation, wasting volunteer and data collection effort. Further, within-bin comparisons may have higher false match rates than random comparisons due to the similarities that caused the patterns to be binned together [26]. Consequently, the testing of large-scale fingerprint systems should make impostor comparisons among only those prints of the same "arch/loop/whorl" type. Testers of

speaker recognition systems may wish to make impostor comparisons only among speakers of the same gender.

137. If the volunteer is aware that an impostor comparison is being made, changes in presentation behaviour may result in unrepresentative results. Therefore, to avoid even subconscious changes in presentation, ideally volunteers should not be told whether the current comparison is a genuine or impostor transaction.

138. Impostor attempts must be made under the same conditions as the genuine attempts. The use of so-called "background databases" of biometric features acquired from different (possibly unknown) environments and populations cannot be considered best practice.

139. Resulting impostor scores are recorded, together with the true and impostor identities of the volunteer. As it is likely that these impostor transactions are taking place alongside genuine transactions, care must be taken that results are attributed to the correct set of scores.

4.5.1.1 Dependent templates

140. For systems that have dependent templates, volunteers making impostor attempts must not be enrolled in the database when the attempt is made. This might involve selecting a subset of volunteers, who will not be enrolled in the system and so can be used as impostors.

4.5.2 Offline generation of impostor transactions

141. Offline impostor comparisons can be made in the same basic way as online comparisons: either randomly selecting with replacement both samples and templates for the non-self comparisons, or randomly selecting some number of non-self templates from all those enrolled for comparison with each sample. The random selection of templates should be independent for each sample.

142. Offline computation can also use the "full cross-comparison" approach, in which each sample is compared with every non-self template.

143. Offline development of matching scores must be carried out with software modules of the type available from the vendors in software developer's kits (SDK). One module will create templates from enrolment images. A second module will create sample features from test samples. These modules will sometimes be the same piece of code. A third module will return a matching score for any assignment of a sample feature to a template. If processing time is not a problem, all features can be compared to all templates. If there are T templates and N features (from the same volunteer crew), $N(T-1)$ comparisons against non-self templates can be performed. These impostor comparisons will not be statistically independent, but this approach is statistically unbiased and represents a more efficient estimation technique than the use of randomly chosen impostor comparisons [25].

144. Many biometric systems collect and process a sequence of samples in a single attempt, for example:

a. Collecting samples over some fixed period, and scoring the best matching sample;
b. Collecting samples until either a match is obtained or the system "times out";
c. Collecting samples until one of sufficient quality is obtained, or the system "times out";
d. Collecting a second sample when the score from the first sample is very close to the decision threshold.

In such cases, a single sample from a genuine attempt might not be suitable as an impostor sample. In case (a), the sample saved will be the one that best matches the genuine template; however, an impostor attempt would be based on the sample best matching the impersonated template. To determine whether it is appropriate to base cross-comparison on a single genuine sample, two questions must be addressed:

e. Does the saved sample depend on the template being compared?
f. If so, does this materially affect the matching scores generated?

If the answers to both these questions are yes, then either the whole sample sequence should be saved and used in offline analysis, or impostor scores should be generated online.

4.5.2.1 Dependent templates

145. For systems with dependent templates, generation of unbiased impostor scores may require a "jack-knife" approach to create the enrolment templates. The jack-knife approach is to enrol the entire crew with a single volunteer omitted. This omitted volunteer is then used as an unknown impostor, comparing his/her sample(s) to all enrolled templates. This enrolment process is repeated for each volunteer, and a full set of impostor scores can be generated.

146. A simpler technique is to randomly partition volunteers into impostors and enrolees. Offline enrolment ignores the data from volunteers labelled impostors, while offline scoring ignores data from volunteers labelled as enrolees. This is a less efficient use of the data than the jack-knife approach.

4.5.2.2 Inter-template comparisons

147. Sometimes cross-comparison of enrolment templates can provide impostor scores. This can be useful, for example, in operational evaluations where samples or features of transactions are not saved. In the case that only single samples are given for enrolment, and enrolment and test quality control are equivalent, N test (or enrolment) templates can be compared to the remaining $(N - 1)$ test (or enrolment) templates. However, if more than a single image is used to create the enrolment template, this is likely to results in biased estimation of impostor scores [24]. This is true whether the enrolment template is averaged or selected from the best enrolment image. No methods currently exist for correcting this bias.

4.5.3 Intra-individual comparisons

148. With some biometric system, the user may be able to present different biometric entities, e.g., any of up to ten fingers, left or right eye or hand, etc. To improve the independence of different samples from a single volunteer, it is possible that an evaluation allows enrolment of more than one finger/hand/eye/etc. as different (sub-) identities. However, within-individual comparisons are *not* equivalent to between-individual comparisons, and must not be included in the set of impostor transactions. For example genetically identical fingers will have a similar number of fingerprint ridges, and are more likely to match than fingerprints from different people.

5 ANALYSIS

5.1 Failure to enrol rate; failure to acquire rate

149. The failure to enrol rate is estimated as the proportion of volunteers who could not be enrolled under the pre-determined enrolment policy.

150. The failure to acquire rate is estimated as the proportion of recorded transactions (both genuine transactions and any online impostor transactions) that could not be completed due failures at presentation (no image captured), feature extraction, or quality control.

151. **Technology evaluation:** In technology evaluations, analysis is based on a previously collected database and there will be no problem in obtaining a sample image. Even so there may be enrolment or acquisition failures, for example, when the image sample is of too low a quality for features to be extracted.

5.2 Detection error trade-off curves

152. The DET measures will be developed using the genuine and impostor matching scores from comparisons between single test samples and single enrolment templates. These scores will be

highly dependent upon the details of the test, and the quality control criteria in place for judging the acceptability of an acquired image. Stricter quality control will increase the failure to acquire rate, but decrease the false match and non-match error rates.

153. Each transaction will result in a recorded matching score. Scores developed for genuine transactions will be ordered. Impostor scores will be handled similarly. Outliers will require investigation to determine if labelling errors are indicated. Removal of any scores from the test must be fully documented and will lead to external criticism of the test results.

154. Histograms for both genuine and impostor scores can be instructive but will not be used in the development of the DET. Consequently, we make no recommendations regarding the creation of the histograms from the transaction data, although this is a very important area of continuing research interest. The resulting histograms will be taken directly as the best estimates for the genuine and impostor distributions. Under no circumstances should models be substituted for either histogram as an estimate of the underlying distribution.

155. Detection error trade-off curves are established through the accumulation of the ordered genuine and impostor scores. As the score varies over all possible values, the DET curve is plotted parametrically, each point (x, y) on the DET representing the false match and false non-match rates using that score as the decision threshold. The false match rate is the proportion of impostor scores at or below the current value of the score parameter, and the false non-match rate is the proportion of genuine scores exceeding the score parameter[11]. The curves should be plotted on "log-log" scales, with false match rate on the abscissa (x-axis) and false non-match rate on the ordinate (y-axis).

156. DET curves can also be used to plot the relationship between the false accept rate (FAR) and false reject rate (FRR) in a similar manner. FAR and FRR will depend on the false match rate (FMR), false non-match rate (FNMR), failure to acquire rate (FTA) and if applicable the binning error rate (BER) and penetration rate (PR), in a manner that will depend on the decision policy. If acceptance depends on a single successful match then

$$\mathrm{FAR} = \mathrm{PR} \times \mathrm{FMR} \times (1 - \mathrm{FTA})$$
$$\mathrm{FRR} = \mathrm{FTA} + (1 - \mathrm{FTA}) \times \mathrm{BER} + (1 - \mathrm{FTA}) \times (1 - \mathrm{BER}) \times \mathrm{FNMR}$$

5.3 Binning error versus penetration rate curve

157. Full testing of negative identification systems requires the evaluation of any binning algorithms in use. The purpose of these algorithms is to partition the template data into subspaces. An input sample is likewise partitioned and compared only to the portion of the template data that is in the same partition.

158. The process of partitioning the template data, however, can lead to partitioning errors. An error occurs if the enrolment template and a subsequent sample from the same biometric feature on the same user are placed in different partitions. In general, the more partitioning of the database that occurs the lower the penetration rate, but the greater the probability of a partitioning error. These competing design factors can be graphed as a binning error versus penetration rate curve.

159. Fortunately, the testing corpus collected for offline testing can be used in a second test to establish both penetration and bin error rates. Both enrolment templates and test samples are binned using the offered algorithm. Binning errors are assessed by counting the number of matching template-sample pairs that were placed in different bins and reporting this as a fraction of the number of pairs assessed. The penetration rate is assessed by counting the number of comparisons required under the binning scheme for each sample against the template database. The average number over all input samples, divided by the size of the database, represents the penetration rate. These results can be represented as a point on a two-dimensional graph.

[11] We assume scores increase as samples and templates become less similar.

160. Frequently, the partitioning algorithm will have tuneable parameters. When this occurs, the experimenter might graph a series of points (a curve or a surface) expressing the penetration and error rate tradeoffs over the range of each parameter.

6 UNCERTAINTY OF ESTIMATES

161. Performance estimates will be affected by both systematic errors and random errors. Random errors are those due to the natural variation in volunteers, samples etc. Systematic errors are those due to bias in the test procedures, etc. For example, if certain types of individual are under-represented in the volunteer crew, this might bias the results.

162. The uncertainty arising from random effects will reduce as the size of the test increases, and can be estimated from the collected data as shown in the following section.

163. It may also be possible to determine the effects of some of the systematic errors. For example we might check whether the error rates for an under-represented category of individuals are consistent with the overall error rates. This would show whether a properly balanced volunteer crew would give different error rates. We could repeat some of the performance trial in different environmental conditions to check that the measured error rates are not unduly sensitive to small environmental changes.

6.1 Estimates for variance of performance measures

164. In this section we give formulae and methods for estimating the variance of performance measures. The variance is a statistical measure of uncertainty, and can be used in estimating confidence intervals etc.

165. The formulae estimating the variance of performance measures depend on some assumptions about the distribution of matching errors listed below:

a. **The volunteer crew is representative of the target population:** This will be the case if, for example, the volunteer crew is drawn at random from the target population.

b. **Attempts by different subjects are independent:** This will not always be completely true. Volunteers' behaviour will be influenced by what they see others do. However the correlations between volunteers are likely to be minor in comparison to the correlations within a set of attempts by one volunteer.

c. **Attempts are independent of threshold:** Otherwise the estimates for the error rates may be biased except at the threshold used for data collection.

d. **Error rates vary across the population:** Different subjects may have different individual false non-match rates, and different subject pairs may have different individual false match rates. I.e., we explicitly allow for "goats", "wolves", and "lambs" [20].

e. **The number of observed errors is not too small:** In cases with no observed errors, the formulae would give a zero variance, but the "Rule of 3" would apply.

6.2 Variance of observed false non-match rate

166. We provide formulae for estimating the variance of the false non-match rate for cases when there is a single sample per volunteer, and when there are multiple attempts per volunteer. These formulae will also be appropriate for estimating variances of failure to acquire rates and failure to enrol rate. We shall use the following notation:

n number of enrolled volunteers;

m_i number of samples from the i^{th} volunteer;

$m = \frac{1}{n}\sum_i m_i$ the average number of samples per volunteer;

a_i number of false non-matches for the i^{th} volunteer;

$p_i = \frac{a_i}{m_i}$ proportion of unmatched samples for the i^{th} volunteer;

$\hat{p}$ observed false non-match rate over all volunteers;

$\hat{V}(\hat{p})$ estimated variance of observed false non-match rate.

6.2.1 Single attempt per volunteer

167. In the case where each of n volunteers makes a single attempt:

$$\hat{p} = \frac{1}{n}\sum a_i \tag{1}$$

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1} \tag{2}$$

A derivation of this estimate may be found in many statistical textbooks, e.g., [31].

168. This formula has sometimes been misapplied to cases where subjects make several attempts. The replacement of the number of subjects n, by the number of attempts nm, is generally not valid.

6.2.2 Multiple attempts per volunteer

169. In the case where each volunteer makes the same number of attempts, the appropriate estimates are given by [31, Section 17.5, page 514]:

$$\hat{p} = \frac{1}{n}\sum p_i = \frac{1}{mn}\sum a_i \tag{3}$$

$$\hat{V}(\hat{p}) = \frac{\sum (p_i - \hat{p})^2}{n(n-1)} = \frac{1}{(n-1)}\left(\frac{\sum a_i^2}{m^2 n} - \hat{p}^2\right) \tag{4}$$

Note that when $m = 1$, the estimates are the same as those in formulae (1) and (2).

170. There will be occasions when the number of attempts per subject varies. Some subjects might not complete the desired number of attempts. Failures to acquire may also cause attempts to be missing from the false non-match rate calculations. In this case we may use:

$$\hat{p} = \frac{\sum a_i}{\sum m_i} \tag{5}$$

$$\hat{V}(\hat{p}) = \frac{\sum (m_i(p_i - \hat{p}))^2}{m^2 n(n-1)} = \frac{\sum a_i^2 - 2\hat{p}\sum a_i m_i + \hat{p}^2 \sum m_i^2}{m^2 n(n-1)} \tag{6}$$

This formula for the variance, provided by [31, Section 17.5] is an approximation, to give an expression in a usable form. Note that, when all m_i are equal, we obtain the same estimates as in formulae (3) and (4).

6.3 Variance of observed false match rate

171. We provide an estimate for the variance of the false match rate for the case where a full set of cross comparisons is made. We shall use the notations:

n number of volunteers (and of enrolment templates);

m number of samples per volunteer;

b_{ij} number of samples from the i^{th} volunteer falsely matching the j^{th} template (and b_{ii}=0);

$c_j = \sum_i b_{ij}$ number of false matches against the j^{th} template;

$d_i = \sum_j b_{ij}$ number of false matches in total by the i^{th} volunteer.

Then the observed false match rate is

$$\hat{q} = \frac{\sum_{ij} b_{ij}}{mn(n-1)} \tag{7}$$

and the variance of the observed false match rate may be estimated as:

$$\begin{aligned} \hat{V}(\hat{q}) &= \frac{\sum_i (c_i + d_i)^2 - \sum_i \sum_{j \neq i} (b_{ij}^2 + b_{ij} b_{ji})}{m^2 n(n-1)(n-2)(n-3)} - \frac{(4n-6)}{(n-2)(n-3)} \hat{q}^2 \\ &\approx \frac{1}{m^2 n^2 (n-1)^2} \sum (c_i + d_i)^2 - \frac{4}{n} \hat{q}^2 \end{aligned} \tag{8}$$

The second line of this estimate (in the case m =1) is the formula given by Bickel [24, equation 17.48], which has been experimentally verified in [25].

6.4 Estimating confidence intervals.

172. With sufficiently large samples, the central limit theorem [31] implies that the observed error rates should follow an approximately normal distribution. However, because we are dealing with proportions near to 0%, and the variance in the measures is not uniform over the population, some skewness is likely to remain until the number of volunteers is quite large.

173. Under the assumption of normality, $100(1 - \alpha)\%$ confidence bounds on the observed error rates are given by $\hat{p} \pm z(1-\alpha/2)\sqrt{\hat{V}(\hat{p})}$, where $z(\ )$ is the inverse of the standard normal cumulative distribution. I.e. the area under the standard normal curve with mean 0, variance 1 from $-\infty$ to $z(1 - \alpha/2)$ is $(1 - \alpha/2)$. For 95% confidence limits the value $z(0.975)$ is 1.96.

174. Often when this formula is applied, the confidence interval reaches into negative values for the observed error rate – but negative error rates are impossible. This is due to non-normality of the distribution of observed error rates. Non-parametric methods, such as the bootstrap [32-34] can be used to obtain confidence intervals in such cases.

6.4.1 Bootstrap estimates of the variance, confidence intervals etc

175. Bootstrap estimation reduces the need to make assumptions about the underlying distribution of the observed error rates and the dependencies between attempts. The distributions and dependencies are inferred from the sample itself. By sampling *with replacement* from the original sample, we can create a bootstrap sample. Then, with a large number of such bootstrap samples, we obtain the empirical distribution of our estimators, which can be used to construct confidence intervals, etc.

176. To illustrate the process, suppose we are estimating the false match rate using a full set of cross comparison with n volunteers, each providing m attempts to be compared against all $(n - 1)$ non-self templates. Let $X(v, a, t)$ be the result of the cross-comparison of the a^{th} attempt by volunteer v against template t, and $\mathbf{X} = \{X(v, a, t) \mid t \neq v \in \{1, \ldots, n\}, a \in \{1, \ldots, m\}\}$ the full set of $mn(n - 1)$ cross-comparison results. Each bootstrap sample is constructed from $\mathbf{X}$ in a way that replicates the original sample structure and dependencies:

a. Sample n volunteers with replacement: $v(1), \ldots, v(n)$. (Sampling with replacement means the list is likely to contain more than one occurrence of the same item).
b. For each $v(i)$ sample with replacement $(n - 1)$ non-self templates: $t(i, 1), \ldots, t(i, n–1)$
c. For each $v(i)$ sample with replacement m attempts made by that volunteer: $a(i, 1), \ldots, a(i, m)$
d. The bootstrap sample is
$\mathbf{Y} = \{X(v(i), t(i, j), a(i, k)) \mid i \in \{1, \ldots, n\}, j \in \{1, \ldots, n–1\}\ a \in \{1, \ldots, m\}\}$.

Many bootstrap samples are generated, and a false match rate obtained for each. The distribution of the bootstrap values for the false match rate is used to approximate that of the observed false match rate.

177. The bootstrap values allow a direct approach for constructing $100(1 - \alpha)\%$ confidence limits: choosing L (lower limit) and U (upper limit) such that only a fraction $\alpha/2$ of bootstrap values are lower than L, and $\alpha/2$ bootstrap values are higher than U. We recommend using at least 1000 bootstrap samples for 95% limits, and at least 5000 bootstrap samples for 99% limits.

7 REPORTING PERFORMANCE RESULTS

178. Performance measures such as the DET curve, failure to enrol and failure to acquire rates, and binning penetration and error rates are dependent on test type, application and population. So that these measures can be interpreted correctly, the following additional information should be given:

a. Details of the system tested. This should include more than just the biometric component, as factors such as the user interface will influence performance too;

b. The type of evaluation:

   Technology evaluation: Details of the sample databases used;
   Scenario evaluation: Details of the test scenario;
   Operational evaluation: Details of the operational application;

c. Size of evaluation:

   Number of volunteers;
   (Number of fingers/hands/eyes/etc. enrolled by each volunteer);
   Number of visits made by volunteer;
   Number of transactions per volunteer (or volunteer finger, etc.) at each visit;

d. Demographics of the volunteer crew;

e. Details of the test environment;

f. Time separation between enrolments and test transactions;

g. Quality and decision thresholds used during data collection;

h. Details of which factors potentially affecting performance were controlled, and how these were controlled (see Appendix A);

i. Details of the test procedure, e.g., policies for determining enrolment failures, etc.;

j. Details of abnormal cases, and data excluded from analysis;

k. Estimated uncertainties (if calculated);

l. Deviations from these Best Practices should also be explained. Sometimes it will be necessary to compromise one aspect, in order to achieve another. (E.g. randomising the order of using fingers on a fingerprint device might lead to user confusion and a higher number of labelling errors.)

179. For comparing performance of different systems, the decision error DET (false reject rate vs false accept rate), which shows the combined effect of matching errors, image acquisition errors and binning errors, will be more helpful than graphs showing the fundamental error rates.

8 CONCLUSIONS

180. We recognise that the recommendations in this document are extremely general in nature and that it may not be possible to follow best practice completely in any test. Often, compromises will need to be made. For example, controlling conditions (for internal validity of results) may be at odds with obtaining a truly representative environment (for external validity). In such situations the experimenter must decide what is the best compromise that will achieve the evaluation objectives, but should also report what was done to enable a correct interpretation of the results.

181. However, we hope that the concepts presented will serve as a framework for the development of scientifically sound test protocols for a variety of devices in a range of environments.

182. As with the previous issue, we welcome comments, criticism and suggestions on this version of Biometric Testing Best Practices, which will be taken into account as the Best Practices are revised further. (All paragraphs are numbered or labelled to facilitate commenting.)

ACKNOWLEDGEMENTS

183. This work owes directly to the support and encouragement of Philip Statham, Tony Mason, and Geoff Lister of CESG. This issue and Issue 1.0 are based originally on the work of Drs. Alvin Martin and George Doddington of NIST. We are most grateful to the careful review and encouragement received from the U.K. Biometrics Working Group and the Office of the e-Envoy, as well as from the U.S. Biometrics Consortium.

REFERENCES

[1] *Best practices in testing and reporting performance of biometric devices, Issue 1*. Report for CESG and Biometrics Working Group, February 2000.

[2] MANSFIELD, A.J., KELLY, G.P., CHANDLER, D.J., and KANE, J. *Biometric product testing final report*. Report for CESG and Biometrics Working Group, March 2001. http://www.cesg.gov.uk/technology/biometrics/media/Biometric%20Test%20Report%20pt1.pdf

[3] BLACKBURN, D., BONE, M., and PHILLIPS, J. *Facial recognition vendor test 2000*. February 2001. http://www.dodcounterdrug.com/facialrecognition/FRVT2000/documents.htm

[4] *BIOIS: Comparative study of biometric identification systems: Public final report*. Study by Fraunhofer-IGD for BSI (German Information Security Agency) and BKA (German Federal Police Agency), May 2000.

[5] *IBG's comparative biometric testing*. http://www.biometricgroup.com/e/performance_data.htm

[6] MAIO, D., MALTONI, D., CAPPELLI, R., WAYMAN, J.L., and JAIN, A.K. Fvc2000: Fingerprint verification competition. *IEEE Trans. PAMI*, 2002, **24**(3), 402-412. Details also online http://bias.csr.unibo.it/fvc2000/

[7] DODDINGTON, G.R., PRZYBOCKI, M.A., MARTIN, A.F., and REYNOLDS, D.A. The NIST speaker recognition evaluation: Overview methodology, systems, results, perspective. *Speech Communication*, 2000, **31**(2-3), 225-254.

[8] PHILLIPS, P.J., MARTIN, A., WILSON, C.L., and PRZYBOCKI, M. An introduction to evaluating biometric systems. *Computer*, (Feb 2000), 56-63.

[9] FEJFAR, A. and MYERS, J.W. The testing of three automatic identity verification techniques. *Proceedings of the International Conference on Crime Countermeasures*, Oxford, July 1977.

[10] DAVIES, D.W. and PRICE, W.L. *Security for computer networks*, Wiley, 1984. (Sections 7.10 & 7.11 review several performance evaluations)

[11] HOLMES, J.P., WRIGHT, L.J., and MAXWELL, R.L. *A performance evaluation of biometric identification devices*. Sandia report SAND91-0276, June 1991.

[12] BOUCHIER, F., AHRENS, J.S., and WELLS, G. *Laboratory evaluation of the iriscan prototype biometric identifier*. Sandia report SAND96-1033, April 1996.

[13] RAUSS, P., PHILLIPS, P.J., HAMILTON, M.K., and DEPERSIA, A.T. Feret (face recognition technology) recognition algorithms. *Proceedings of Automatic Target Recognizer System and Technology Conference*, July 1996.

[14] ROETHENBAUGH, G. ICSA biometric certification. *Biometric industry product buyer's guide*, ICSA, 1998, 27-31.

[15] SOUTHEY, R. The story of the three bears. *The doctor: Volume iv*, Longman, 1837.

[16] NEWMAN, H.H., FREEMAN, F.N., and HOLZINGER, J.K. *Twins*, Chicago University Press, 1937. Results on fingerprint similarity between identical twins cited in [31, table 10.20.1].

[17] DAUGMAN, J. and DOWNING, C. Epigenetic randomness, complexity, and singularity of human iris patterns. *Proceeding of the Royal Society Biological Sciences*, 2001, **268**, 1737-1740. Details also online: http://www.cl.cam.ac.uk/users/jgd1000/genetics.html

[18] JAIN, A.K., PRABHAKAR, S., and PANKANTI, S. On the similarity of identical twin fingerprints. *Pattern Recognition*, 2002, **35**(11), 2653-2663. Also online http://www.cse.msu.edu/cgi-user/web/tech/document?NUM=00-23

[19] MARTIN, A., DODDINGTON, G.R., KAMM, T., ORDOWSKI, M., and PRZYBOCKI, M.A. The DET curve in assessment of detection task performance. *Proceedings of Eurospeech*, Rhodes, Greece, 1997, 1895-1898.

[20] DODDINGTON, G., LIGGETT, W., MARTIN, A., PRZYBOCKI, M., and REYNOLDS, D. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *ICSLP*, November 1998.

[21] LOUIS, T.A. Confidence intervals for a binomial parameter after observing no successes. *The American Statistician*, 1981, **35**(3), 154.

[22] HANLEY, J.A. and LIPPMAN-HAND, A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *Journal of the American Medical Association*, 1983, **249**(13), 1743-1745.

[23] JOVANOVIC, B.D. and LEVY, P.S. A look at the rule of three. *The American Statistician*, 1997, **51**(2), 137-139.

[24] WAYMAN, J.L. Technical testing and evaluation of biometric identification devices. *Biometrics: Personal identification in networked society*, edited by A.K. Jain*, et al.*, Kluwer, 2000, 345-368.

[25] WAYMAN, J.L. Confidence interval and test size estimation for biometric data. *Proceedings of the IEEE AutoID Conference*, 1999. Available online at http://www.engr.sjsu.edu/biometrics/nbtccw.pdf

[26] WAYMAN, J.L. *Multi-finger penetration rate and ROC variability for automatic fingerprint identification systems*. National Biometric Test Center, May 1999. Available on-line at http://www.engr.sjsu.edu/biometrics/nbtccw.pdf

[27] Appendix G: Interim IAFIS image quality specifications for scanners. *Electronic fingerprint transmission specification*, Criminal Justice Information Services, CJIS-RS-0010 (V4), 1995. Available online at http://www.engr.sjsu.edu/biometrics/publications_appendixg.html

[28] NIST. *Best practice recommendations for capturing mugshots and facial images. Version 2*. NIST Image Group, September 1997. http://www.itl.nist.gov/iad/894.03/face/bpr_mug3.html

[29] COX, R. Three new speech coders from the ITU cover a range of applications. *IEEE Communications Magazine*, 1997, **35**(9 - special issue on standardisation and characaterisation of G729), 40-47.

[30] GODFREY, J., GRAFF, D., and MARTIN, A. Public databases for speaker recognition and verification. *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1994.

[31] SNEDECOR, G.W. and COCHRAN, W.G. *Statistical methods*, Iowa State University Press, 1967 (Sixth edition).

[32] EFRON, B. and TIBSHIRANI, R.J. *An introduction to the bootstrap*, Chapman and Hall, 1997.

[33] DIEGERT, K.V. Estimating performance characteristics of biometric identifiers. *Proceedings of Biometrics Consortium Conference*, San Jose, CA, June 1996.

[34] BOLLE, R.M., RATHA, N., K., and PANKANTI, S. Confidence interval measurement in performance analysis of biometric systems using the bootstrap. *Proceedings of Workshop on Empirical Evaluation Methods in Computer Vision*, Hawaii, December 2001.

APPENDIX A – FACTORS INFLUENCING PERFORMANCE

184. This appendix lists some of the user and environmental factors that have been found to affect performance. These factors may need to be controlled or recorded during the data collection phases of an evaluation.

185. During the planning stages of an evaluation, for each potential influencing factor, one might consider:

a. What controls (if any) will be needed to minimise (or ascertain) the effects on performance? This might involve setting conditions to be constant over all attempts, or may involve randomisation so that the effects are distributed evenly over users, attempts etc.

b. What assumptions or reasons render it unnecessary to control a particular factor? For example, a factor might affect the test scenario in the same way it would the target application. In other cases, preliminary investigations might show that the effect of particular factors is minimal for the device(s) concerned.

c. What information should be recorded during the evaluation either (i) to help determine the significance (or show the insignificance) of any factor, or (ii) to identify exceptional cases that might otherwise unduly bias results? If a problem is related to an identifiable subset of participants, it may be possible to compare the error rate figures for that subset against the remaining participants.

Such a checklist can be included with the reported results.

186. The factors listed will generally cause problems with only a subset of biometric technologies. For example, illumination changes affect only optical-based systems (e.g., those based on Face, Fingerprint, Retina, Iris or Vein imaging), while acoustic noise would affect sound-based systems (e.g., Speaker verification). Moreover, some biometric devices will operate in a way to control the effects of any problems. Equally, problems may be observed that are not included in our lists.

187. When problems are caused, generally the effect is to reduce the sample quality, thereby increasing the failure to enrol, failure to acquire or the false non-match rate. However, there are also some cases in which noisy or problem images will allow spurious matches increasing the false match rate.

A.1 List of factors

188. **Population demographics**

AGE: Children (who change more rapidly) and older people (where perhaps minor damage to the measured biometric takes longer to heal) tend to have more false non-matches and failures to acquire than average.

ETHNIC ORIGIN: The quality of a person's biometric (for a particular biometric system) may depend on their ethnic origin, gender and occupation. A biometric system tuned to a specific target population may perform less well if used with a different ethnic or gender mix.

GENDER;

OCCUPATION.

189. **Application**

TIME ELAPSED BETWEEN ENROLMENT AND VERIFICATION: Template ageing, i.e., changes in the users biometric pattern and method of presentation, will vary in accordance with the delay between creation of the enrolment template, and the verification or identification attempt. Generally, performance a short time after enrolment, when the user appearance and behaviour has changed very little, is far better than that obtained weeks or months later.

TIME OF DAY: Behaviour and physiology can change during the day.

USER FAMILIARITY: As users become familiar with the system, they are more likely to position themselves correctly, and to know the appropriate action to compensate for many of the verification problems that might arise.

USER MOTIVATION: Users will act differently according to the importance of the biometric transaction.

190. **User physiology**

BEARDS & MOUSTACHES: Can affect *face* systems.

BALDNESS;

DISABILITY, DISEASE or ILLNESS: for example:

- AMPUTATION: unable to use *hand* or *finger* based systems;
- ARTHRITIS: difficult to use *hand* or *finger* based systems;
- BLIND: unable to use *iris* or *retina* based systems, and also affects user positioning for other systems;
- BRUISES: temporary effect on *face* or *hand* images;
- COLDS, LARYNGITIS: temporary effect on *voice*;
- CRUTCHES: may make it difficult to stand steadily;
- SWELLING: temporary effect on *face* or *hand* images;
- WHEELCHAIRS: system may be at wrong height for those in wheelchairs;
- CHANGES IN MEDICAL CONDITION: can be faster than normal ageing effects.

EYELASHES: long eyelashes can make less *iris* visible.

FINGERNAIL GROWTH: affects *hand* and *finger* positioning.

FINGERPRINT CONDITION:

- DEPTH AND SPACING OF RIDGES;
- DRY, CRACKED or DAMP.

HEIGHT: The very tall or very short (or those in wheelchairs) may have difficulty in positioning themselves correctly.

IRIS COLOUR INTENSITY;

SKIN TONE: Can affect ability of system in correctly locating *faces* or *irises.*

191. **User behaviour**

DIALECT, ACCENT, and NATIVE LANGUAGE: will influence *voice* systems. E.g., a system optimised for US English speakers may perform less well on UK English speakers, or with other languages.

EXPRESSION, INTONATION, and VOLUME: affect *Voice* systems.

FACIAL EXPRESSIONS;

LANGUAGE (ALPHABET): influences *handwritten signature* systems.

MISPOKEN OR MISREAD PHRASES: will affect *Voice* systems.

MOVEMENT: Some systems require the subject to remain still, while others work better with some movement.

POSE, POSITIONING:

- FACING CAMERA, PROFILE, ANGLED;
- HEAD TILT: affects *face* and *iris* systems.
- OFFSETS & ROTATIONS: affect *fingerprint* and *hand* systems.
- DISTANCE TO CAMERA;
- TOO HIGH, TOO LOW, TOO FAR LEFT, or TOO FAR RIGHT.

PRIOR ACTIVITY:

- OUT OF BREATH: will affect *Voice* systems.
- SWEATINESS will affect *fingerprint* systems.
- SWIMMING: shrivelling of fingers will affect *fingerprint* systems.

STRESS, TENSION, MOOD, DISTRACTIONS.

192. **User appearance**

BANDAGES/BANDAID: can alter or mask part of a *hand*, *face* or *fingerprint.*

CLOTHING:

HATS, EARRINGS, SCARVES: can affect *face*-based systems.
SLEEVES: can hinder *hand*-based systems.
HEEL HEIGHT: changes apparent user height.
TROUSERS / SKIRTS / SHOES: influence on *gait* recognition.
CONTACT LENSES: coloured or patterned contact lenses affect *iris* recognition.
COSMETICS: will temporarily alter *face* appearance.
GLASSES, SUNGLASSES: can partly obscure the *face* or *iris*.
FALSE FINGERNAILS: can alter positioning for *hand* or *finger* based systems.
HAIR STYLE / COLOUR: will temporarily alter *face* appearance.
RINGS;
TATTOOS.

193. **Environmental Influences**

BACKGROUND:
COLOUR, CLUTTER, CONTAINING FACES or SHADOWS: can affect performance of *face*-finding systems.
NOISES & OTHER VOICES: can alter the recorded signal with *voice*-based systems and also affect the ability of the user to hear the instructions.
LIGHTING LEVEL, DIRECTION, REFLECTIONS: may affect camera-based systems.
WEATHER:
TEMPERATURE, HUMIDITY: influence *fingerprint* dryness/dampness, visibility of *veins*, thermal images, etc.
RAIN, SNOW: Wet hair will affect *face* appearance.

194. **Sensor and hardware**

DIRT / SMEARS / RESIDUAL PRINTS:
CAMERA LENS;
PLATEN.
FOCUS;
SENSOR QUALITY: Microphone quality (*Voice*) Camera quality (*Imaging systems*).
SENSOR VARIATIONS:
BETWEEN SENSORS: Different instances of the same sensors may perform slightly differently. Differences will be greater with different versions or different types.
SENSOR WEAR;
SENSOR REPLACEMENT.
TRANSMISSION CHANNEL: The transmission channel can add noise to the signal. Moreover, it can vary between attempts. For example, the route and networks used for phone calls may vary, and quality may be load dependent.

195. **User Interface**

FEEDBACK: Performance can depend on the feedback users receive. E.g., do they see their submitted fingerprint, enabling them to alter their presentation to achieve a better quality biometric sample?
INSTRUCTION;
SUPERVISION: There may be differences in enrolments, user training, and user attempts due to the differences and changes in supervisors.

A.2 Examples for reporting

196. **Finger position**

Observation: The guides on the scanner seemed to position fingers within the tolerances for the algorithms.
Control: None
Record: N/A

197. **Illumination**
Observation: Changes in illumination due to variations in daylight etc. cause enrolment and verification problems.
Control: Trials to take place in a room with natural daylight excluded, and constant lighting levels.
Record: N/A

198. **Illumination**
Observation: Stray illumination caused reflections on the iris.
Control: Unit modified to shield sensor from extraneous light sources.
Record: N/A

199. **Glasses**
Observation: We found it almost impossible to enrol people with glasses on face system X.
Control: People with glasses were asked to remove them to use this device.
Record: Number of people wearing glasses, so that the figure can be included in failure to enrol rates,

200. **Dirt on platen**
Observation: Accumulating oils on platen caused degradation in fingerprint system performance.
Control: System to be cleaned regularly (stating cleaning schedule).
Record: When system cleaned.

201. **Weather**
Observation: Sweaty fingers caused enrolment/verification problems.
Control: None, weather conditions assumed to be typical.
Record: Temperature, humidity during the trial.

APPENDIX B – VARIANCE OF PERFORMANCE MEASURES AS A FUNCTION OF TEST SIZE

202. As the test size increases, the variance of estimates will decrease, but the scaling factor depends on the source of variability. For example, volunteers may have differing error rates, giving a component of variance that scales as 1/(the number of volunteers).

203. To illustrate the relationship between test size and the variance of error rates, we consider a typical test, where volunteers each make multiple genuine attempts, and impostor attempts are generated by cross-comparison of the genuine samples against the set of enrolled templates. (For the purposes of this example, we shall ignore possible additional levels of sampling, such as the use of multiple fingers, eyes or hands per volunteer. This would lead to additional components of variance, with intermediate scaling factors.)

204. The variance of the observed false non match rate ($FNMR_{OBS}$) has components scaling as
1/(number of volunteers) due to variability of volunteers; and
1/(number of genuine attempts) due to the (residual) variability of attempts.
If n volunteers each make m independent genuine attempts the variance is given by:

$$V(FNMR_{OBS}) = \frac{\sigma^2}{n} + \frac{\rho(1-\rho) - \sigma^2}{mn}$$

where ρ and σ^2 are the true mean and variance, respectively, of FNMR for different people.

205. Similarly the variance of observed false match rate (FMR_{OBS}) has components that scale as
1/(number of impostor volunteers);
1/(number of impersonated templates);
1/(number of (impostor volunteer, impersonated template) pairs);
1/(number of genuine samples);
1/(number of impostor attempts).

If n volunteers each make m independent genuine attempts, and cross comparisons are made against t templates the variance is given by

$$V(FMR_{OBS}) = \frac{\sigma_V^2}{n} + \frac{\sigma_T^2}{t} + \frac{\sigma_X^2 - \sigma_V^2 - \sigma_T^2}{nt} + \frac{\sigma_S^2 - \sigma_V^2}{mn} + \frac{\alpha(1-\alpha) - \sigma_S^2 - \sigma_X^2 + \sigma_V^2}{mnt}$$

where: α and σ_V^2 are the mean and variance of FMR for different impersonators;
α and σ_T^2 are the mean and variance of FMR for different impersonated templates;
α and σ_X^2 are the mean and variance of FMR for different impersonator-impersonated template pairs; and
α and σ_S^2 are the mean and variance of FMR for different samples.

206. In these formulae, σ^2 will be in the range 0 to $\rho(1-\rho)$, while σ_V^2, σ_T^2, σ_X^2 and σ_S^2 will lie between 0 and $\alpha(1-\alpha)$. In the most optimistic case:

$V(FNMR_{OBS}) = \rho(1-\rho)/(mn)$ and $V(FMR_{OBS}) = \alpha(1-\alpha)/(mnt)$,

while in the most pessimistic case:

$V(FNMR_{OBS}) = \rho(1-\rho)/n$ and $V(FMR_{OBS}) = \alpha(1-\alpha)/\min(n, t)$.

In real cases, the truth is somewhere between these extremes. Section 6 provides methods for estimating the actual variance.

207. Doddington et al [20] show that biometric systems can have "goats", "lambs" and "wolves". "Goats" have a personal false non-match rate significantly higher than that for the overall population, "lambs" are those whose templates incur a disproportionate share of false matches, while "wolves" are those whose samples are particular successful at giving false matches. This would imply that, for FNMR, the component of variance for volunteers is non-zero, and for FMR, the components for volunteers, and for templates are non-zero.

APPENDIX C – EXAMPLE VOLUNTEER CONSENT FORM

Consent form for Biometric Performance Trial	
Name	<*name*>
Contact details	<*details*>
Identifier(s) used in test corpus	<*identifier*s>
I willingly participate in these trials. I consent to <*images/recordings*> of my <*finger/ face/ iris/ hand/ …*> and my questionnaire responses[12] being collected during the trial and stored electronically. I agree to the use of this data by <*testing organisation*> and <*list other companies that may use the data*> for the purposes of evaluating performance of biometric systems and identifying problems and improvements. I understand that my name[13]/identity will not be stored or shown in any released database[14]. or report.	
Signature	

Figure 4 Sample volunteer consent form

[12] It can be useful to record other information about the volunteer crew, e.g., age occupation etc.

[13] May need to be changed when testing signature systems.

[14] When the corpus contains images from two types of biometrics, e.g., signatures and face images, it should not be possible to align the different types of images e.g., associating a face with a signature.