

CESG contract X92A/4009309

Biometric Product Testing Final Report

Issue 1.0
19 March 2001

Tony Mansfield
Gavin Kelly
David Chandler
Jan Kane

Centre for Mathematics and Scientific Computing
National Physical Laboratory
Queen's Road
Teddington
Middlesex
TW11 0LW

Tel: 020 8943 7029
Fax: 020 8977 7091

EXECUTIVE SUMMARY

This is a report of a performance evaluation of seven biometric systems conducted by NPL over the period May to December 2000. The test programme was sponsored by the Communications Electronics Security Group (CESG) as part of their Biometrics Work Programme in support of the “Modernising Government” and other initiatives.

The objectives of the test programme were:

- To show the level of performance attainable by a selection of biometric systems;
- To determine the feasibility of demonstrating satisfactory performance through testing;
- To encourage more testing to be sponsored, and to promote methodologies contributing to the improvement of biometric testing.

Face, Fingerprint, Hand Geometry, Iris, Vein and Voice recognition systems were tested for a scenario of positive identification in a normal office environment, with cooperative non-habituated users. The evaluation was conducted in accordance with the “Best Practices in Testing and Reporting Performance of Biometric Devices” produced by the UK Government Biometrics Working Group, and used 200 volunteers over a three-month period.

Results presented include:

- Failure to Enrol and Failure to Acquire Rates;
- The trade-off between matching errors (False Match Rate vs. False Non Match Rate) and between decision errors (False Acceptance Rate vs False Rejection Rate) over a range of decision criteria;
- Throughput rates of users in the live application, and of the matching algorithm in off-line processing;
- Sensitivity of the systems’ performance to environmental conditions, and the differences in performance over different classes of users.

Biometric system performance is dependent on the application, environment and population. Therefore the performance results presented here should not be expected to hold for all other applications, or in all environmental conditions. In particular caution should be exercised when comparing these results with those of other systems tested under different conditions.

CONTENTS

1	Introduction	4
2	Selection of systems	4
3	Test scenario	5
3.1	Volunteer crew	5
3.2	Environment	6
3.3	Enrolments & verifications	6
4	Test methodology.....	7
4.1	Dealing with enrolment failures.....	7
4.2	Avoiding data collection errors.....	8
5	Results overview	9
5.1	Failure to enrol	9
5.2	Failure to acquire	9
5.3	False match rate (FMR) vs false non-match rate (FNMR)	9
5.4	False acceptance rate (FAR) vs. false rejection rate (FRR).....	10
5.5	Multiple attempt error rates	11
5.6	User throughput.....	12
5.7	Matching algorithm throughput	12
5.8	Performance differences by user & attempt type.....	13
6	Validation of methodology & future enhancements	14
6.1	The requirement for additional system functionality	14
6.2	One attempt may involve a sequence of images	14
6.3	Failure to acquire	15
6.4	Other performance trade-offs.....	15
Appendix A. Test protocol.....		16
A.1	Introduction	16
A.2	Device setup	16
A.3	Volunteer crew	17
A.4	Enrolment.....	17
A.5	Test data collection	18
A.6	Analysis & Reporting.....	18
Appendix B. Consent form & Enrollment data sheet.....		20
Appendix C. Verification data sheet		21
Appendix D. Significance of user & attempt variations.....		22

FIGURES

Figure 1: Age and gender of volunteer crew	5
Figure 2. Environmental conditions during the trials	6
Figure 3. Positioning of systems in test laboratory	7
Figure 4. Detection error trade-off: FMR vs FNMR	10
Figure 5. Detection error trade-off: FAR vs FRR	11
Figure 6. Detection error trade-off: Best of 3 attempts	11

TABLES

Table 1. Brief details of systems tested.....	5
Table 2. Failure to enrol rates.....	9
Table 3. Failure to acquire rates	9
Table 4. User transaction times	12
Table 5. Diagnostic program throughput.....	13
Table 6. Summary of performance differences by user type.....	13

1 INTRODUCTION

1. This is a report of a performance evaluation of seven biometric systems conducted by NPL over the period May to December 2000. The test programme was sponsored by the Communications Electronics Security Group (CESG) as part of their Biometrics Work Programme in support of the “Modernising Government” and other initiatives.
2. The test programme had three main objectives:
 - a. To show the level of performance attainable by a selection of biometric systems;
 - b. To determine the feasibility of demonstrating satisfactory performance through testing;
 - c. To encourage more testing to be sponsored, and to promote methodologies contributing to improvement of biometric testing.
3. The tests provide factual, vendor-independent data on the performance of biometric devices. This will inform CESG on the general capability of biometric technology, and will help in the development of policy on the use of biometrics in Government. It will also assist members of the UK Government Biometrics Working Group (BWG) in the assessment of the applicability of biometric technology to their potential applications.
4. The tests will implement and validate the BWG proposed methodology for biometric testing. The outcome will support the further development of this methodology for use with Common Criteria evaluations of biometric products and systems.
5. It is also hoped that this initial evaluation will, by example:
 - a. Promote the methodology to a wider audience and contribute to the improvement of biometric testing by other organisations; and
 - b. Encourage further testing to be sponsored.To allow wider dissemination of the results (given that open publication of results was not a requirement for vendors participating in the trials), the report has been organised into two parts with different restrictive markings. The intention is that Part I excludes any commercially sensitive information and can be made publicly accessible, while Part II contains full details for CESG and Government Departments.

2 SELECTION OF SYSTEMS

6. The Test Programme was announced on the Biometrics Consortium list server, and some thirty companies responded to the call for submission of devices for testing. Because of overlap in terms of devices proposed, about twenty different systems were considered for inclusion in the test programme.
7. The criteria for selection of systems to test were agreed by CESG and the Biometrics Working Group.
 - a. Fingerprint, hand and iris technologies must be included. Other systems tested should use different technologies, except for fingerprint where two systems might be tested.
 - b. Within a technology, selection should be on the basis of wide availability and commonality of use.
 - c. Systems should be capable of meeting basic CESG performance requirements.
 - d. Systems should be testable under the agreed methodology (and, implicitly, the system performance should not be adversely affected by the proposed test protocol).
 - e. The vendor should be able to support the trials within the required timescales.
8. Using these criteria, seven systems were selected for testing, using face, fingerprint, hand geometry, iris, vein pattern, and voice and recognition. There were two fingerprint systems: one using optical fingerprint capture, the other a chip sensor. Table 1 gives brief details of the tested systems. Systems have been named where vendors are happy for their results to be publicly available. (Full details of all systems are given in Part II of this report, which has a more restricted circulation.)

Short name	Brief description
Face Face (2)	Visionics – FaceIt Verification Demo Alternative enrolment and matching algorithms for this system
FP-chip FP-chip (2)	VeriTouch – vr-3(U) Alternative enrolment and matching algorithms provided by Infineon
FP-optical	<i>Fingerprint recognition system.</i>
Hand	Recognition Systems – HandKey II
Iris	Iridian Technologies – IriScan system 2200
Vein	Neosciences-Biometrics – Veincheck development prototype
Voice	OTG – SecurPBX

Table 1. Brief details of systems tested

9. As there is just one device per technology, it should be noted that the performance results presented are not necessarily fully representative of all systems of the same type. Indeed, even relatively minor modifications to the systems tested can give considerably different performance.

3 TEST SCENARIO

10. The test scenario was one of positive verification in a “normal office environment”, with co-operative non-habituated users. The tests were conducted with 200 volunteers, over a three-month period. The typical separation between enrolment and a verification transaction was one to two months.

3.1 Volunteer crew

11. To obtain participants, a call for volunteers was issued by e-mail and in the NPL in-house newsletter. A small payment offered as an incentive for participation (and adherence to the trial “rules”). All those responding were invited to participate, though some withdrew when they could not attend an appointment for enrolment. A limited further call was issued to some staff of the other laboratories on site (NWML and LGC) to achieve slightly over 200 participants. The volunteer crew were thus self-selecting, consisting mostly of staff working on the NPL site. The age and gender profile is shown in Figure 1. This approximates that of the workforce on site.

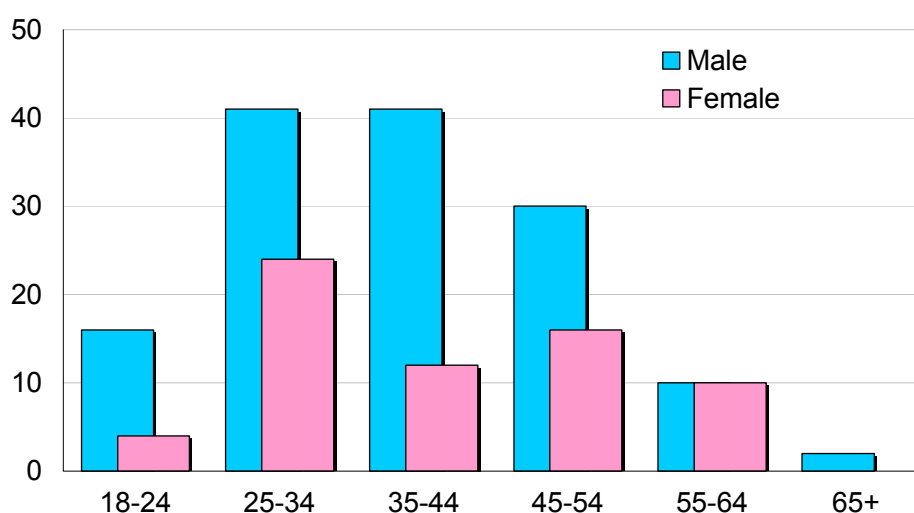


Figure 1: Age and gender of volunteer crew

12. This volunteer crew is not fully representative of the general UK adult population. Women and those older than 45 are under-represented, also the balance between different ethnic

groups is probably incorrect (ethnic origin of volunteers was not recorded). Moreover, as the volunteer crew are used to working in a scientific environment, they are more accepting of technology than the population at large. Potentially this might reduce errors due to the behavioural element in biometric system use.

3.2 Environment.

13. The tests were conducted in a room previously in normal office use.
14. Lighting levels were controlled. The room's fluorescent lighting was always on, and the window blinds kept down to reduce effects of daylight variations. The devices were sited in accordance with recommendations of the product suppliers, and those most sensitive to changes in illumination were positioned away from the window. Similarly one device whose use was sensitive to background noise was located in a quieter area off the main test laboratory. These adjustments are documented with the test results for each device.
15. The temperature and humidity of the test laboratory were not controlled. Figure 2 indicates how outdoor temperature¹ and humidity² varied between the days of the trials

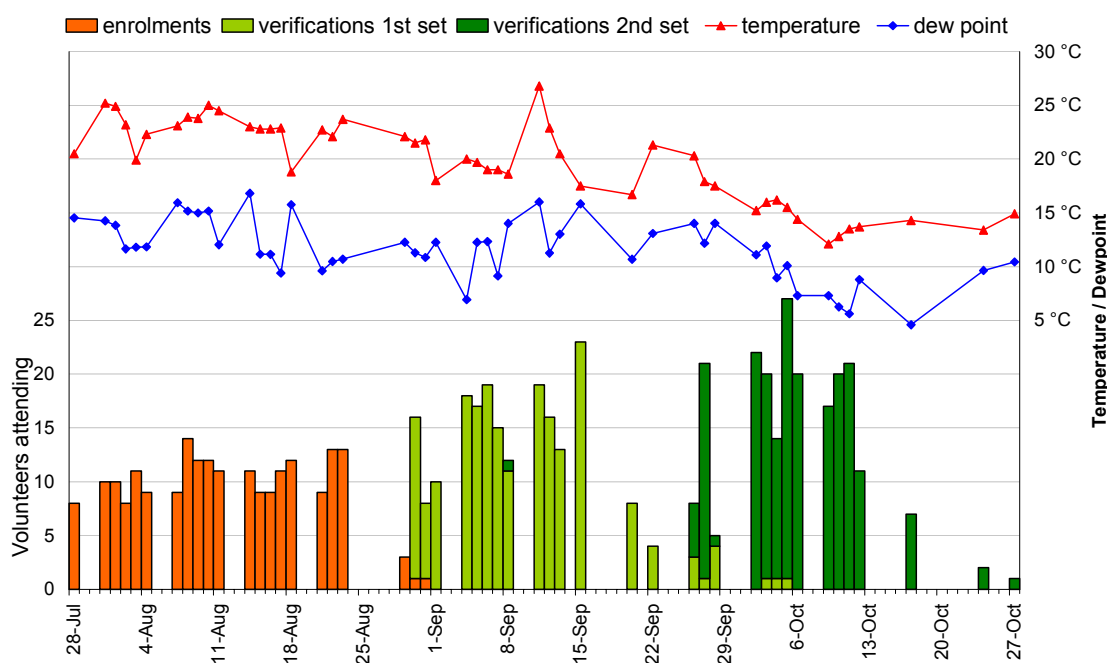


Figure 2. Environmental conditions during the trials

3.3 Enrolments & verifications

16. Figure 2 also shows the daily distribution of enrolment and verification transactions. On average the first set of verifications was made 29 days after enrolment, and the second set of verifications, 55 days after enrolment.

3.3.1 Order effects

17. The order in which the devices were used could potentially affect performance.

¹ Figures based on readings from local weather station.

² Dew point is plotted instead of relative humidity. This removes the strong (inverse) correlation with temperature, and to allows the same °C scale to be used.

- a. On arriving at the test laboratory, volunteers could be out of breath (if they have hurried to make their appointment) or have cold hands/fingers (when cold outside), recovering to a more normal state after a few minutes.
 - b. The illumination for the face recognition system increased the amount of iris visible (i.e. reduces pupil size) with a potential effect on iris recognition when this occurs shortly after.
 - c. Feedback from one fingerprint device might affect user behaviour (e.g. finger pressure) on the other.
18. Other than volunteers attempting speaker verification when out of breath, these order effects did not appear significant. Further order effects may also exist, but are also believed to be insignificant. In view of this, a complex fully randomised sampling plan was not adopted.
- a. Transactions on the Voice system were not conducted until the volunteer had regained their breath.
 - b. The order in which the devices were used alternated between a clockwise order around the room, and anti-clockwise. However, this ordering was often modified to avoid queuing at any system. There were no order correlations between visits.

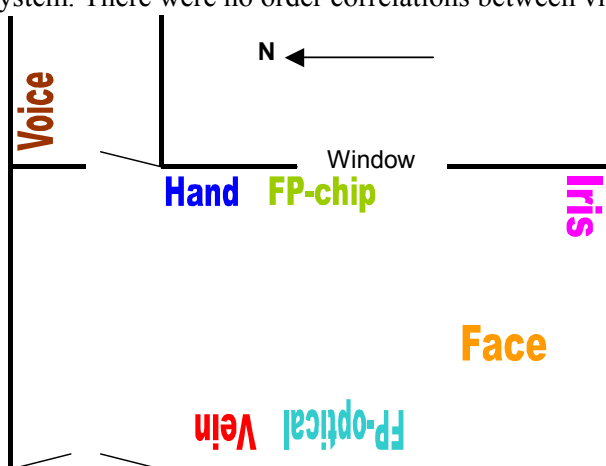


Figure 3. Positioning of systems in test laboratory

4 TEST METHODOLOGY

19. The performance trials were conducted in accordance with *Best Practices in Testing and Reporting Performance of Biometric Devices*³ produced by UK Government Biometrics Working Group. The test protocol followed is described in *A test protocol for the Technical Performance Evaluation of Biometric Devices*. For completeness this Test Protocol is included in Appendix A.
20. Modifications and enhancements to the general test protocol are discussed below.

4.1 Dealing with enrolment failures

21. Observations during preliminary testing showed:
- a. Often more than two attempts would be required to obtain an enrolment. This seemed to be particularly the case with the Voice and both Fingerprint systems, where obtaining a good quality “image” is more dependent on user behaviour and familiarity.
 - b. For some systems, the enrolment software did not provide for re-enrolment. In such cases, problem enrolments needed to be deleted, using the underlying operating system, before re-enrolment was possible. For data-integrity reasons, we were reluctant to do this

³ Available at <http://www.cesg.gov.uk/biometrics/>

while under the pressure of processing volunteers, and as a result re-enrolments had to occur on a subsequent visit.

- c. Some systems did not automatically record every enrolment attempt failure.
22. The protocol for dealing with enrolment failures was therefore modified. Where practical, immediate re-enrolment was attempted, (as previously). However, at subsequent visits, whenever a volunteer had failed to enrol on one of the devices, they were asked to try re-enrolling regardless of the number of previous enrolment attempts.

4.2 Avoiding data collection errors

23. Additional procedures were put in place to help avoid data collection errors:
 - a. Errors due to the use of the wrong hand, finger, etc.
 - b. Errors due to attributing the attempt to the wrong identity.

4.2.1 Avoiding use of wrong hand, finger, etc.

24. Users were asked to always use their right index finger, eye or hand as appropriate. Without this consistency, it would be difficult for supervisors to observe and prevent use of the wrong finger, hand or eye at enrolment or verification. The saved images allow further checks that the correct iris, hand or finger was used, though this is easier for iris and hand images than for fingerprint images.

4.2.2 Avoiding attribution of attempt to wrong identity.

25. Each user was allocated a PIN for the trials, which was shown on the named data sheet collected by the user at each session (see e.g. Appendix C). The following possibilities for attributing attempts to the wrong identity must be addressed by checking procedures.
 - a. The user picks up the wrong data sheet⁴.
 - b. The user mistypes their PIN, producing another valid PIN⁵.
 - c. The user forgets to enter their PIN on a system where the PIN is not cleared between attempts. As a result the attempt is made against the previous user's identity⁶.

These were addressed as follows.

26. Feedback on claimed identity

The Voice, Face and Iris systems provided feedback on the claimed identity. This would show the individual and supervisor that failures were due to the wrong PIN being used.

27. Error detecting PINs

The PINs used to claim an identity were chosen to minimise the chance that mistyping would produce another valid identity. This was done using the ISBN error-detection scheme (though avoiding use of "X" as the check digit). The 4-digit PINs $abcd$ have the property that $4a+3b+2c+d$ is exactly divisible by eleven. This detects all single digit errors and transpositions. From the available PINs, the set used was as widely spaced as possible, in the range 1000 – 9999, giving robustness against more complex typing errors.

28. User makes at least 3 attempts per device per session

If a PIN not being entered causes attempts to be recorded against the previous user's identity, these will be the 4th or subsequent attempts. However, these will be ignored as only the first 3 attempts per user per session are analysed.

29. Any incorrect attempts were recorded on the user's data sheet, allowing for annotation of the logged data and exclusion from analysis. Where possible, prior to conducting analyses, the

⁴ This happened twice (of a possible 412 occasions), where the volunteers had very similar names.

⁵ One of the systems recorded when incorrect PINs were entered. Of some 2000 entered PINs, 5 were entered incorrectly. Two single digit errors, one transposition, and two 2-digit errors.

⁶ This could happen on three of the systems tested, occurring twice, once, and no times (of a possible approx 400 occasions).

data saved for verification failures were checked further, to determine if the cause of failure was a mis-acquisition or a mis-labelling.

5 RESULTS OVERVIEW

5.1 Failure to enrol

30. The “failure to enrol” rate measures the proportion of individuals for whom the system is unable to generate repeatable templates. This includes those unable to present the required biometric feature (for example the Iris system failed to enrol the iris of a blind eye), those unable to produce an image of sufficient quality at enrolment, as well as those unable to reproduce their biometric feature consistently. Enrolment failure rates for the systems tested are shown in Table 2. Note that, in cases of difficulty, several attempts were allowed to achieve an enrolment. If necessary, these further enrolment attempts were made at subsequent visits by the volunteer.

System	Failure to enrol rate
Face	0.0%
Fingerprint – Chip	1.0%
Fingerprint – Optical	2.0%
Hand	0.0%
Iris	0.5%
Vein	0.0%
Voice	0.0%

Table 2. Failure to enrol rates

5.2 Failure to acquire

31. The “failure to acquire rate” measures the proportion of attempts for which the system is unable to capture or locate an image of sufficient quality. This includes cases where the user is unable to present the required biometric feature (e.g. having a plaster covering his or her fingerprint); and cases where an image is captured, but does not pass the quality checks. Failure-to-acquire rates for the systems tested are shown in Table 3. The figures exclude cases where the image was not captured due to user error (e.g. the user not positioning themselves correctly) as in these cases the attempt was simply restarted.

System	Failure to acquire rate
Face	0.0%
Fingerprint – Chip	2.8%
FP-chip (2)	0.4% ⁷
Fingerprint – Optical	0.8%
Hand	0.0%
Iris	0.0%
Vein	0.0%
Voice	2.5%

Table 3. Failure to acquire rates

5.3 False match rate (FMR) vs false non-match rate (FNMR)

32. The fundamental operation of a biometric system is the comparison of a captured biometric image against an enrolment template. The false match and false non-match rates measure the

⁷ For verification, minimal quality checks were performed.

accuracy of this matching process. By adjusting the decision criteria there can be a trade-off between false match and false non-match errors; so the performance is best represented by plotting the relationship between these error rates in a detection error trade-off graph.

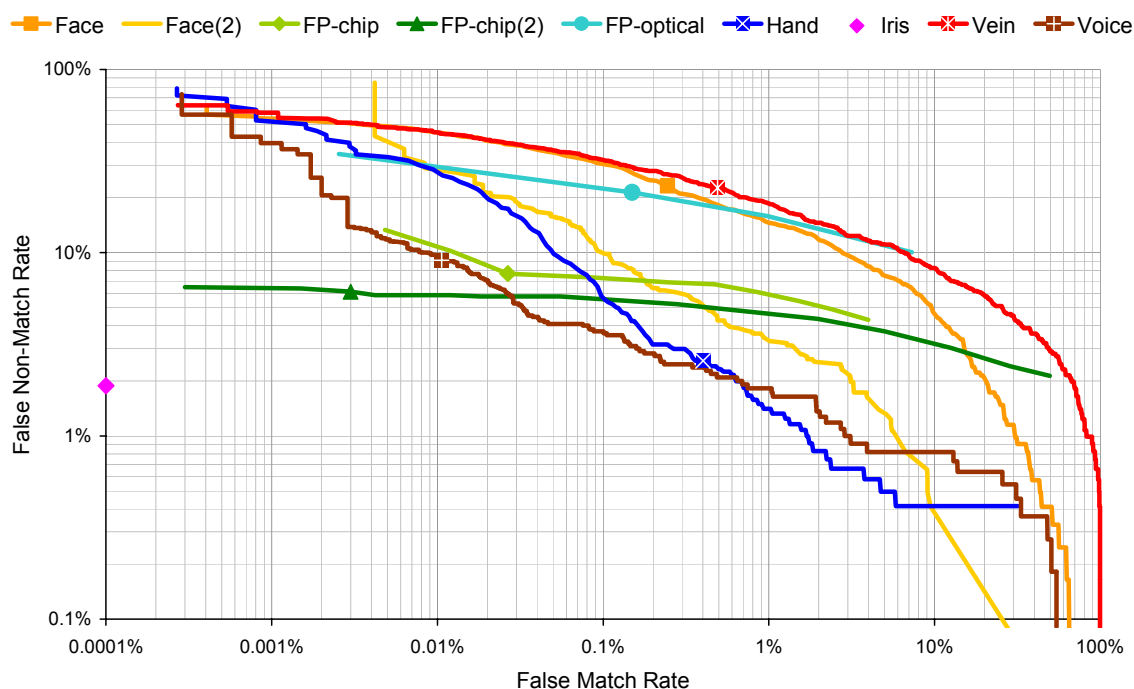


Figure 4. Detection error trade-off: FMR vs FNMR

33. Matching algorithm performance for each system, over a range of decision criteria, is shown in Figure 4. (The lower and further left on the graph, the better the performance). The node on each curve shows performance at the default decision threshold. No curve is shown for the Iris system, which operates with a pre-determined threshold. The iris system had no false matches in over 2 million cross-comparisons. For all the other systems the leftmost point on each curve represents a single false match in the total number of cross-comparisons made.
34. Observing images corresponding to false non-matches showed that some of matching failures were due to poor quality images. Systems vary in how they deal with poor quality images, some will “fail to acquire” such images, while systems will often cope with poor image quality. Therefore the matching error rates should not be considered in isolation from the failure to acquire and failure to enrol rates.

5.4 False acceptance rate (FAR) vs. false rejection rate (FRR)

35. False acceptance and rejection rates measure the decision errors for the whole system. These measures combine matching error rates, and failure to acquire rates in accordance with the system decision policy. When the verification decision is based on a single attempt:

$$FAR(\tau) = (1 - FTA) FMR(\tau)$$

$$FRR(\tau) = (1 - FTA) FNMR(\tau) + FTA$$

where τ is the decision threshold, and FMR, FNMR, FTA, FAR and FRR are the false match rate, false non-match rate, failure to acquire rate, false acceptance rate and false rejection rate respectively.

36. The false acceptance false rejection trade-off curve is shown in Figure 5. The curves for the face, hand geometry, iris and vein systems are unchanged, as these systems had no failures to acquire.

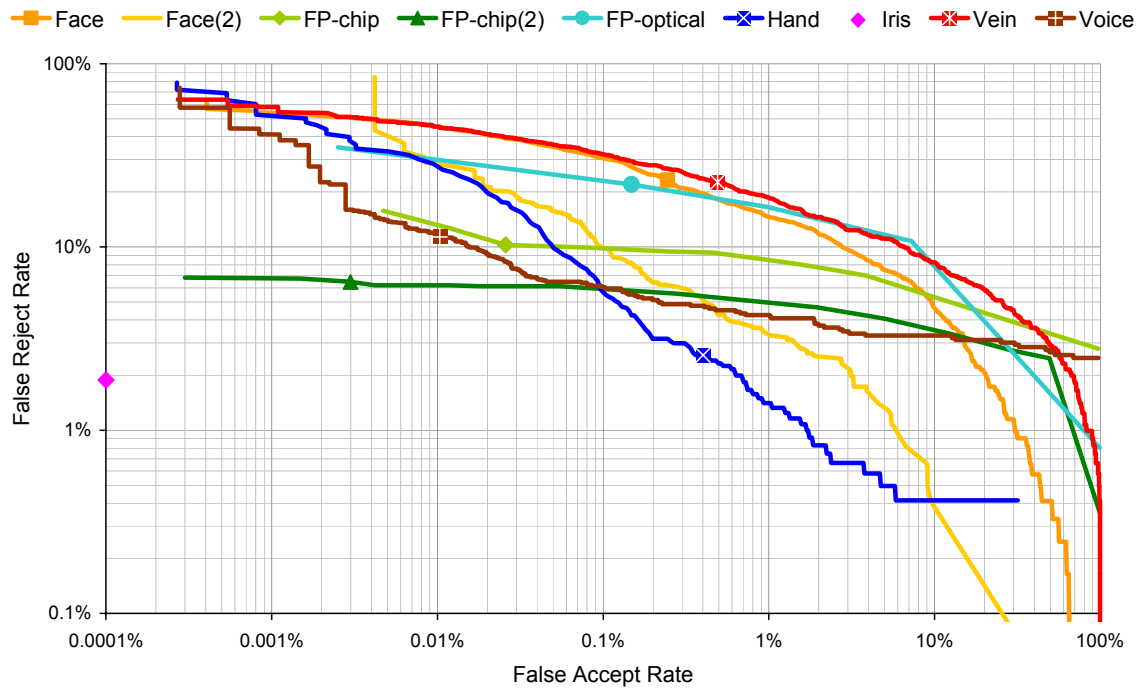


Figure 5. Detection error trade-off: FAR vs FRR

5.5 Multiple attempt error rates

37. Many systems allow multiple attempts, in their normal mode of operation. The effects on error rates of a “best-of-3” decision policy are examined in this section.

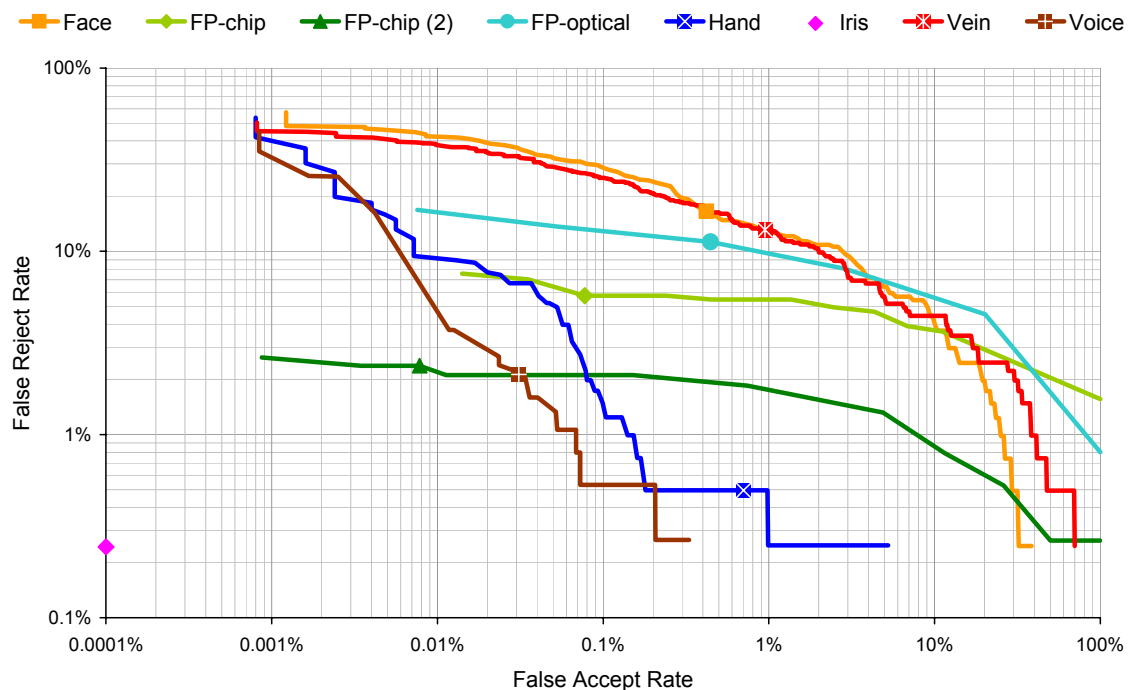


Figure 6. Detection error trade-off: Best of 3 attempts

38. The 3-attempt genuine and impostor scores are the best matching score from the 3 attempts made at the person-visit (scored against the chosen template). The resulting detection error trade-off (DET) curves are shown in Figure 6.

39. This method of obtaining the DET curve is appropriate when all attempts are constrained to use the same finger, face or hand etc. In real life, it may be possible to substitute a different finger, face, hand, etc at the second or third attempt. If so (and assuming the individual impostor attempts are fully independent) the 3-attempt false acceptance rate at any decision threshold is given by $1-(1-\alpha)^3$ where α is the false acceptance rate for a single attempt at the same threshold. Thus, two detection error trade-off curves may be shown:
- Where all three attempts are constrained to use the same finger, hand, face, etc; and
 - Where substitutions are allowed between attempts.
- In the case of the trial systems and data, the two curves follow each other closely⁸, so Figure 6 shows a single curve for each system⁹.

5.6 User throughput

System	Transaction Time (Seconds)			Time includes entry of PIN?
	Mean	Median	Minimum	
Face	15	14	10	Excluded
Fingerprint-Optical	9	8	2	Excluded
Fingerprint-Chip	19	15	9	Excluded
Hand	10	8	4	Included
Iris	12	10	4	Included
Vein	18	16	11	Included
Voice	12	11	10	Excluded

Table 4. User transaction times

40. The time for a user transaction has been calculated using the time differences logged between consecutive transactions (as detailed in Appendix A.6.7). Table 4 shows the mean, median and minimum transaction times to indicate the spread of results. The differences in operation of the trial systems accounts for much of the difference in timings.
- The Face system collected a sequence of images over a 10 second period, saving the best match obtained. The transaction times would be somewhat shorter if the system stopped when the threshold was first exceeded; however, this would not have allowed us to examine performance over a range of decision thresholds.
 - The Iris system would normally work in identification mode, not requiring PIN entry. This would reduce transaction times.
 - The keypad of the Vein system could not cope with rapid entry of the PIN. The time to do this dominates the overall transaction time.
 - The transaction times for the Voice system were dominated by the time taken in giving user prompts and feedback. The prompting and speeds were chosen to be suitable for users unaccustomed to the system, rather than for maximum throughput.

5.7 Matching algorithm throughput

41. The measured throughput of the programs for batch mode running of the matching algorithms is shown in Table 5. These diagnostic programs had significant overheads, for example logging all matching attempts to a file, or handling the Windows interfaces. Therefore, the matching algorithm throughput may be significantly higher than those shown, perhaps by a factor exceeding 100. (In the case of the chip-based fingerprint system, the difference in throughput of the two diagnostic programs illustrates the improvement possible. In an

⁸ The ratio FAR_b/FAR_a of the false acceptance rates derived under the different assumptions varies from 1 to 1.3 for the voice system and fingerprint systems; from 1 to 1.7 for the vein system, and from 1 to 2 for the hand and face systems.

⁹ For the FP-chip, and FP-optical systems, a cross-comparison scoring of all attempts against each template was not available, and the curve shown is derived as detailed in paragraph 39. For FP-chip (2) and all the other systems, the curve was derived using a full set of genuine and impostor scores.

equivalent implementation, the basic FP-chip algorithm would be faster than the more complex alternative FP-chip(2.)

System	Matches per minute	Program interface	System, processor speed, memory, & OS			
Face	800	Windows	Pentium			Win2K
FP-chip	60	Windows	Pentium	133MHz	32Mb	Win98
FP-chip (2)	2,500	Command Line	Pentium	500MHz	64Mb	Win95
FP-optical	50	Windows	Pentium	500MHz	64Mb	Win95
Hand	80,000	Command Line	SunUltra5	270MHz	128Mb	SunOS5.8
Iris	1,500,000	Command Line	SunUltra5	270MHz	128Mb	SunOS5.8
Vein	130	Windows	Pentium	500MHz	64Mb	Win95
Voice	680	Command-Line	Pentium	500MHz	64Mb	Win95

Table 5. Diagnostic program throughput

5.8 Performance differences by user & attempt type

42. Attempts can be categorised by:
- Whether made at enrolment visit or at the second or third visit by the volunteer;
 - The gender of the volunteer;
 - The age of the volunteer;
 - Whether the volunteer was wearing spectacles in the case of Face and Iris systems;
 - The length of the user’s pass-phrase in the case of the Voice system.
- Performance differences between these subsets have been analysed, and are reported for each system in Part II. The general findings are summarised in Table 6.

System	Gender	Age	Visit	Other
	Observations: lowerFRR<higherFRR lowerFRR<higherFRR			Less significant More significant ¹⁰
Face	male<female	younger<older	enrol<later	without<with glasses
FP-chip	male<female	younger<older	enrol<later	
FP-chip(2)	male<female	younger<older	enrol<later	
FP-optical	male<female	younger<older	enrol<later	
Hand	male<female			
Iris				without<with glasses
Vein	male<female	younger<older	enrol<later	
Voice	female<male	younger<older	enrol<later	

Table 6. Summary of performance differences by user type

43. False rejection rates for attempts made immediately following enrolment were generally significantly lower than (less than half) those made at volunteer’s second or third visit.
44. Generally men had a lower false rejection rate than women (the voice system being the only exception), and younger volunteers a lower false rejection rate than their older colleagues. The gender differences appeared the more significant for the Face, Hand and Vein systems, and the age differences the more significant for the Fingerprint systems.
45. As women and over 45’s were under-represented in our volunteer crew, our results may be biased. For a given threshold, with equal numbers of men and women, a slightly higher false non-match rate might be expected. However since false matches are more likely within the same gender class, the equalisation would reduce the false match rate at the same threshold.

¹⁰ The more significant observations have a χ^2 value exceeding 15. (See Appendix D for details.) The probability of such observations being due to the random nature of the sample is in the range 0.01% - 20% dependent on the degree of correlation between different attempts by the same person.

6 VALIDATION OF METHODOLOGY & FUTURE ENHANCEMENTS

46. The evaluation has implemented the BWG proposed methodology for biometric testing, validating many aspects of this methodology. For example:
- Demonstrating the feasibility of the methodology;
 - Showing that the number of volunteers used (200) is sufficient to evaluate performance of biometric systems at their current level of accuracy;
 - The practical significance of issues described in “Best Practices” has been demonstrated:
The need for time separation between enrolments and verification attempts;
The need to minimise the chance of labelling errors;
The modified procedures to simulate unknown impostor attempts when there are dependencies between templates.
- A single evaluation cannot demonstrate repeatability of the results. However, some of the devices evaluated have been tested elsewhere in similar scenarios, and the results are consistent.
47. The evaluation revealed further issues concerning the applicability of the test protocol, and enhancements to best practices. These are noted below.

6.1 The requirement for additional system functionality

48. The test protocol required systems to save data for off-line calculation of genuine and impostor matching scores. This capability is often not provided in a vendor’s standard supplied system. This raises the following issues:
- Some systems will be unable to meet this requirement for testing (for example standalone systems which store templates are stored locally, but have insufficient memory to log transaction attempts). This point was raised by some of the vendors who initially expressed an interest in participation in the trials.
 - When the required functionality is achievable with vendor support, it is important that protocols are sufficiently consistent across testing organisations. Otherwise the vendor needs to develop a different customisation for each test, and support costs can be very significant.
 - Sometimes achieving the desired functionality can affect system performance. For example the time taken in logging images may slow the system and affect user behaviour. It is also possible that implementing the required functionality at minimal cost will introduce errors into the system.
49. If all testing, including impostor tests, are conducted “live” these problems are avoided. However, this requires:
- Data collection to be very closely supervised as all results must be logged by the supervisor;
 - Extra attempts to be made to show performance at a variety of decision thresholds; and
 - Extra attempts to be made for live impostor tests.

6.2 One attempt may involve a sequence of images

50. With many biometric systems, a sequence of images is processed in a single verification attempt. For example, with the trial system it appears that:
- The Face system collects images over a period of 10 seconds, and gives the best match obtained;
 - The Chip-based Fingerprint system collects images until a match is obtained, or until timeout;
 - The Optical Fingerprint system scans for fingerprints until an image of sufficient quality is obtained, or the timeout is reached;
 - The Hand Geometry system occasionally requires a second hand placement, when the score is very close to the decision threshold;

e. The Iris system collects images until a match is achieved or until timeout.

51. The current version of “Best Practices” does not explicitly deal with these cases, yet this mode of operation can sometimes bias off-line calculations using the collected data. For example with the face system, in a real impostor attempt the score would be based on the image that best matches the impersonated template. A cross-comparison of stored genuine images uses the image that best matches the genuine template, and therefore may underestimate the false match rate.

52. The questions that must be addressed are:

a. Would the decision be based on a different image if comparison were against a different template?

b. If so, would live impostor attempt scores be higher/lower than off-line scoring with genuine attempt images?

In the case of the tested Optical Fingerprint, Hand Geometry and Iris systems, the image collected does not depend on the template being matched. With the Fingerprint Chip, the collected image might instead be lost before timeout; and, apart from image quality, should be equivalent to the image saved from a genuine attempt.

6.3 Failure to acquire

53. As noted in Section 5.3 (paragraph 34), different systems handle poor quality input in different ways. With some systems this may result in a failure to acquire, and with others a matching failure. In this respect the FAR-FRR trade-off graph provides a better comparison of performance than the FMR-FNMR trade-off graph.

6.4 Other performance trade-offs

54. Systems may have other adjustable parameters affecting performance in addition to (or instead of) an adjustable decision threshold. These allow different performance trade-offs (which, depending on the application, may be more important than the FAR-FRR trade-off). For example, with the Face, Iris, and Chip-Fingerprint systems, which try to match collected images over a fixed time period, there is a trade-off between the time allowed and the false rejection rate.

APPENDIX A. TEST PROTOCOL

A.1 Introduction

This report describes the test protocol planned for the UK Government Biometric Test Programme. The protocol is for “scenario testing” and conforms to the guidelines in “Best Practices in Testing and Reporting Performance of Biometric Devices”. The protocol is intended to be practical in terms of effort and costs, and applicable to many of today’s commercially available biometric devices when operating in their intended environments.

Several systems will be tested at the same time, in a standard indoor (office) environment and using a volunteer crew similar to the general adult UK population. The trials will involve approximately 200 volunteers using each of the systems being tested. Volunteers will attend the trials on three occasions: firstly for enrolment and practice attempts; and later, one and two months after enrolment, to collect “genuine” attempts Detection Error Trade-off (ROC) analysis.

Impostor attempts will be simulated using cross-comparison of genuine attempts against enrolment templates for other enrollees. This will be carried out off-line using vendor-provided software with the collected enrolments and genuine-attempt images and data.

A.1.1 Applicability of this protocol

Biometric limitations — The protocol cannot be used if it takes much longer than a few seconds for the system to extract the required biometric features. For example we could not test a system that uses 10 minutes of typing at a keyboard to make an identity decision. The separation between enrolment and test attempts will be approximately 1 month. If we are interested in the effects of template ageing time over a timespan much greater than this, the protocol may also be inappropriate.

System functionality — We can only test complete systems. These must be able to operate in “verification” mode, matching a single attempt against a single stored template. It is also necessary for the system to log specific information about each attempt, and there must be a capability for off-line generation of matching scores

System Error Rates — We shall not be able to measure error rates to values of 1% or below with any certainty. For example, if 1% of the population have (or lack) some feature causing enrolment failure, there is a 13% chance that no-one in a 200 person sample have that peculiarity. On the other hand to measure error rates exceeding 10% we may be using more volunteers than required, and a smaller test may be more cost effective.

A.1.2 Modelled Scenario

The scenario modelled is that of a verification application in an indoor environment.

Co-operative users — It is hard to replicate the actions and motivations of an uncooperative user.

Overt system — We shall be using volunteers who will be brought to a specific location for testing, and

who will test several devices. This effectively rules out covert testing.

Non-habituated users — Our volunteers will use the system a few times only, with gaps of a few weeks between each use. The level of habituation will therefore be quite low. We shall avoid using volunteers who have extensively used one of the systems under test, so that comparisons are fair. We do not propose replicating a higher level of habituation by allowing practice attempts: this would create additional complexities to be able to separate practice attempts from the real test attempts.

Supervised enrolment, lightly-attended use — Enrolment will be supervised. Subsequent attempts will be lightly attended: there will be someone on hand to sort out problems should these occur. However, it should be noted that, after enrolment, the main role of the supervisor is to ensure the integrity of the data collection process rather than to assist volunteers in their attempts.

Standard environment — The tests will be conducted indoors, in a standard office environment. It is harder, and more costly to conduct the trials in an outdoor environment, and currently relatively few devices will operate satisfactorily in an outdoor environment.

Public users (UK adults) — Volunteer user attitudes are likely to be closer to those of the general public, than that of company employee. Also, volunteers will be local to the testing laboratory, and their biometric features will reflect the UK demographics. Results may be different with other population demographics. We note that our volunteers are probably more scientifically aware (and perhaps better able to follow instruction) than the general public.

Closed system — We shall enrol and test using the same system. Note that if the system would normally use several sensors, where there are considerable variations between sensors, the proposed protocol may not be appropriate.

A.1.3 Performance Measures

The proposed tests will measure the following aspects of performance (where applicable).

- Failure to enrol rate
- Failure to acquire rate
- Detection error trade-off graph (i.e. ROC)
- System false match and false non-match rates
- Penetration rate (where appropriate)
- Binning error rate (where appropriate)
- User throughput
- Matching algorithm throughput (reported with processing system used)
- Sensitivity of performance to (potentially problematic) changes in environment, population, or usage

A.2 Device setup

We allow vendor involvement during device set-up to help ensure that the systems are correctly installed and operating optimally.

A.2.1 Install systems & familiarisation

The complete system will be installed at the test site. Account will be taken of vendor recommendations regarding positioning, illumination, and background noise etc. in so far as these are realistically achievable in a general office/indoor environment. Threshold, image quality and other settings will be set in accordance with vendor advice.

A.2.2 Test sensitivity of performance to environment, population, usage

Some pre-trial tests will be carried out to determine environmental and other factors that may cause problems. This will be a limited investigation, mainly using the testing team. The aim is to determine:

- what potential problems exist,
- if these problems are controlled by the system,
- how significant the problems appear to be,
- whether we need to impose environmental or other controls to minimise the problem during the trials,
- what additional information we need to record to identify difficult subsets of volunteers during subsequent analyses.

Some of the potential sensitivities to test, and what may be done to analyse or control any problems are shown in the following table:

Tech-nology	Effect to test	If effects seem significant
All	age, gender, template-ageing	Compare of error rates for different subsets of volunteers/attempts
All	lighting level & direction	Control lighting levels during trial
All	dirt/smears on sensor	Set policy for cleaning devices
All	movement during attempt	Provide appropriate instructions for volunteers
All	positioning	Provide appropriate instructions for volunteers
Finger-print	Dry / cold / cracked / damp / wet fingers	Advise volunteers on improving fingerprint quality. Record temperature & humidity
Hand geo-metry	rings, plasters, etc.	Log attempts made with rings etc. Provide separate error rates for these cases
Iris, Face	Glasses	Record those who wear glasses/contact lenses Provide separate error rates for these cases

A.2.3 Set enrolment & transaction attempt policies

The enrolment policy will be set to deal with the problems identified, with the aim of achieving the greatest number of good enrolments.

The supervisors who will conduct enrolment will be trained and familiar with each system and its common problems.

A.2.4 Produce system information for volunteers.

For each system, a short description of how the system operates, and how it should be used will be prepared in consultation with the system vendor. This is to reduce the burden of describing full details of the systems at enrolment, and before later transaction attempts.

A.3 Volunteer crew

A call for volunteers will be issued. To encourage participation a small reward will be offered. If more than 200 people volunteer, participants will be selected at random from the volunteers.

Before enrolment participants will be informed of the purpose of the trials, what is required of them, and what information will be collected and stored. They will be asked to sign to give their consent to the collection of biometric images and information, and to confirm that they have not previously used any of the devices being tested. Age category and gender of participants will be recorded, together with any information found useful in identifying problem cases in the preliminary trials.

A.4 Enrolment

Each participant will attempt to enrol on each system under test. The order of enrolment on the devices being tested will be randomised. Only one set of equipment will be used for each system to avoid “channel” effects. Enrolment will be conducted using the enrolment functions of the supplied systems, and will supervised by a member of staff who had been trained for this purpose. Enrolment images will be collected by the system. *(We use the word image to refer to the actual input signal; this may not strictly be an image in the case of non-optical devices. If the system is unable to record actual enrolment images, it may be possible to conduct the required analyses using the image templates.)*

Immediately after enrolment, several attempts will be made to check that the participant can be reliably verified. Advice to help users achieve successful verifications will be given if necessary. If they cannot be reliably verified this shall count as an enrolment failure.

If enrolment fails, one re-enrolment will generally be attempted. *(In some cases it may be clear that subsequent attempts must fail, for example if the volunteer does not have the required biometric feature. In such cases no re-enrolment attempt would be made. In other cases the enrolment failure may due to a clearly identifiable error which can easily be overcome, for example failures due to not following the proper enrolment process. In such cases more than two enrolment attempts might be made.)*

Some systems allow an “override” to register a poor quality image as an enrolment template in cases of difficulty; such features will not be used. Any problems with enrolment will be noted by the enrolment supervisor.

Cases where the enrolment template cannot be generated, or where all practice attempts fail, are

considered to be failed enrolments. In these cases, subsequent verification attempts are not required of the participant on the device in question. Data from failed enrolments will be removed from the enrolment database and will not be used in analysing false match or false non-match error rates.

A.5 Test data collection

Volunteers will make two sets of transactions, at approximately one and two months after enrolment. On each occasion they should make (at least) three attempts. This will allow direct calculation of “best of three attempt” rejection rates, and can also reveal whether some users are much more error prone than others.

Attempts will be largely unsupervised, but there will be a supervisor on hand to help in case of difficulty. Users may observe attempts made by others, but will not be allowed to make practice attempts (apart from those they made as part of enrolment). This is to ensure that only the genuine transactions are recorded. It is also the case that practice attempts could artificially lower the failure to acquire rate. Additional attempts (i.e. after the required 3 attempts) may be made. It is important to ensure that no attempt is made against the identity of another participant. If a volunteer is keen to see a rejection, it is permitted that they may make an attempt against a non-participating identity. Again, such attempts should not take place immediately prior to their “genuine” attempts.

The order of using the devices will be random across users, and not correlated with the order of use on other occasions. Users will be asked to try to make these attempts successful, and to refrain from making bogus attempts (e.g. using the wrong finger on fingerprint devices, or pulling faces on face recognition devices). As an incentive to obey these instructions, payment for participation is linked to making the required number of good attempts.

Attempt images will be collected by the system, and user details, date and time logged. To avoid data entry errors, user identity will be entered using a swipe card or smart card if possible.

The supervisor will note any problems that arise during the test data collection, so that non-genuine attempts are not included in the analyses. Details of such attempts should be reported.

A.6 Analysis & Reporting

A.6.1 Data collected

Collected by system

- event logs as collected automatically by each system
- images of all test attempts
- enrolment database
- enrolment images

Collected by supervisor:

- log of failed enrolments
- log of (non-genuine) attempts to be excluded
- user details, e.g. age, sex (*The relevant user information to collect will depend on the sensitivities identified in preliminary tests.*)

A.6.2 Failure to enrol rate

The proportion of volunteers failing to obtain an enrolment (of sufficient quality) will be reported along with the enrolment policy and any quality threshold settings.

A.6.3 Failure to acquire rate

The proportion of attempts resulting in a failure to acquire error, averaged across all enrollees, will be reported together with any quality settings.

A.6.4 Detection Error Trade-off plot

The following enrolments and attempts will be excluded when deriving false match and false non-match rates:

- enrolment templates associated with any failed enrolment,
- attempts made on the day of enrolment,
- attempts made by non-enrollees, non participants in the trials, or by participants not completing the trials,
- attempts noted as a non-genuine in the supervisor log book,
- attempts resulting in failure to acquire errors
- extra attempts (4th or later attempt) made by any user on any day. (This is to ensure there is no imbalance due to some users making many more attempts than others).

Distance scores for genuine transactions may have been generated “live” during data collection. Otherwise we use vendor provided software for generating these distance scores off-line from the collected images.

Some systems do not generate distance scores, but can operate at various security settings. In such cases the attempts will be analysed using off-line software at different security settings. In such cases we consider the distance measure to be the strictest security setting at which the attempt results in a match.

We use the supplied software to generate impostor attempt distance scores, by comparing each attempt against the templates for all other enrollees. In the case of non-independent templates it will be necessary to re-enrol all enrollees apart from the one who made the attempt.

The Detection Error Trade-off curve plots the proportion of genuine transaction scores exceeding the matching threshold (*we assume that low scores imply a good match and high scores a poor match*) against the proportion of impostor transaction scores below that threshold, as the threshold varies.

A.6.5 System false accept & false reject rates

In cases where the usual decision policy of the system is not based on a single attempt-template comparison, we give the false accept rate and false reject rate using the actual decision policy, at the system settings used.

A.6.6 Penetration rate & binning error rate.

If a binning algorithm is used, we need to know the “bin” for each template and each genuine attempt.

The penetration rate is the average proportion of the database that would need to be searched if the system were operating in identification mode, where the average is taken over all genuine attempts. This can be estimated if we know the number of attempts in each bin, and which bins are compared against each other. A bin error occurs when an attempt is placed in a bin which is not compared with the correct bin for the biometric entity used, and hence will fail to match.

A.6.7 User throughput & matching algorithm throughput.

User throughput measures the elapsed time of a single transaction. All attempts are to be timed at a consistent point during the transaction (e.g. the start time). The difference in times between the first and second, or second and third attempts, by an individual on one day approximates the total transaction time. This assumes that the 2nd and 3rd attempts immediately follow the first attempt.

We can time the off-line calculation of impostor distance scores and compute the number of

template-attempt matches performed to obtain the matching algorithm throughput. As the time is hardware dependent, the system used should be specified with the resulting throughput rate.

A.6.8 Sensitivity to population & environment

Where there appear to be differences in performance due to population, environment or usage changes (see section A.2.2), in some cases we will be able to assess the affects on performance by analysing subsets of the attempts. For example we can compare the error rates for different age categories, for people with glasses against those without glasses etc. We can also compare the error rates for attempts one month after enrolment with those two months after enrolment (and with error rates immediately after enrolment) to see the effects of template ageing. Comparing the error rates for the first attempt with those for the second and third attempt made on any occasion may show possible improvement in performance due to habituation.

APPENDIX B. CONSENT FORM & ENROLLMENT DATA SHEET

<p>Name</p> <p>Laboratory</p> <p>Phone</p> <p>Email</p>	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td colspan="3" style="text-align: center;">TRIAL ID</td> </tr> <tr> <td><input type="checkbox"/> Male</td> <td colspan="2"><input type="checkbox"/> Female</td> </tr> <tr> <td colspan="3">Age:</td> </tr> <tr> <td><input type="checkbox"/> 18-24</td> <td><input type="checkbox"/> 25-34</td> <td><input type="checkbox"/> 35-44</td> </tr> <tr> <td><input type="checkbox"/> 45-54</td> <td><input type="checkbox"/> 55-64</td> <td><input type="checkbox"/> 65+</td> </tr> <tr> <td colspan="3">Other</td> </tr> <tr> <td colspan="2"></td> <td><input type="checkbox"/> Glasses</td> </tr> <tr> <td colspan="2"></td> <td><input type="checkbox"/> Contact Lenses</td> </tr> </table>	TRIAL ID			<input type="checkbox"/> Male	<input type="checkbox"/> Female		Age:			<input type="checkbox"/> 18-24	<input type="checkbox"/> 25-34	<input type="checkbox"/> 35-44	<input type="checkbox"/> 45-54	<input type="checkbox"/> 55-64	<input type="checkbox"/> 65+	Other					<input type="checkbox"/> Glasses			<input type="checkbox"/> Contact Lenses
TRIAL ID																									
<input type="checkbox"/> Male	<input type="checkbox"/> Female																								
Age:																									
<input type="checkbox"/> 18-24	<input type="checkbox"/> 25-34	<input type="checkbox"/> 35-44																							
<input type="checkbox"/> 45-54	<input type="checkbox"/> 55-64	<input type="checkbox"/> 65+																							
Other																									
		<input type="checkbox"/> Glasses																							
		<input type="checkbox"/> Contact Lenses																							

I am happy to participate in these trials. I consent to my biometric data being collected during the trial and stored electronically.
 I permit use of this data for the purposes of evaluating performance of biometric devices, by the National Physical Laboratory, the Government Biometrics Working Group, and by the manufacturers of the devices under test. *[Data made available outside NPL will consist of only the collected biometric data, and the personal details in the box above.]*

Signed: _____

System	Enrolled OK	Problems / Notes
Face		
Iris		
Vein		
Hand Geometry		
Voice		
Fingerprint Optical Reader		
Fingerprint Chip Reader		

Return for recognition attempts on: _____

APPENDIX C. VERIFICATION DATA SHEET

«FirstName» «LastName»	TRIAL ID	«PIN»
Please make 3 attempts on each system Try your best to be correctly recognised - Do NOT try and trick the systems		
System	& Brief Instructions	Comments
VEIN		
1. Place RIGHT hand on pad	<input type="checkbox"/>	
2. Click button under your fingers to take image	<input type="checkbox"/>	
3. Enter «PIN» on keypad, check on screen, then press *	<input type="checkbox"/>	
FINGERPRINT – OPTICAL SENSOR		
Enter «PIN» in ID box – Check this before proceeding	<input type="checkbox"/>	
1. Press VERIFY to make a verification	<input type="checkbox"/>	
2. Use RIGHT INDEX finger	<input type="checkbox"/>	
FACE		
Enter «PIN» and check your image displayed	<input type="checkbox"/>	
1. Press START VERIFICATION	<input type="checkbox"/>	
2. Stand on marked spot and face camera	<input type="checkbox"/>	
IRIS		
1. If needed click START or to show ID entry box	<input type="checkbox"/>	
2. Enter «PIN» and click OK	<input type="checkbox"/>	
3. Use RIGHT eye	<input type="checkbox"/>	
FINGERPRINT – CHIP SENSOR		
Enter «PIN» in ID box – Check this before proceeding	<input type="checkbox"/>	
1. Press START to commence verification	<input type="checkbox"/>	
2. Use RIGHT INDEX finger	<input type="checkbox"/>	
HAND GEOMETRY		
1. Enter «PIN» and press “#YES” key	<input type="checkbox"/>	
2. Use RIGHT hand	<input type="checkbox"/>	
VOICE		
Dial 6901 and follow instructions	<input type="checkbox"/>	
	<input type="checkbox"/>	
	<input type="checkbox"/>	

For impersonation attempts use ID «PIN-impostor»	<input type="checkbox"/>	
	<input type="checkbox"/>	
	<input type="checkbox"/>	
Options for payment		
<input type="checkbox"/> (NPLML Staff)	Please make payment with my November salary My staff number is:	
<input type="checkbox"/> (non NPLML staff)	Please send a cheque to:	
<input type="checkbox"/>	Please donate my payment to the NPL Sports Club Pavilion Rebuild Fund	
<input type="checkbox"/>	Please donate my payment to Save the Children	
<input type="checkbox"/> I wish to waive payment		Signed:

APPENDIX D. SIGNIFICANCE OF USER & ATTEMPT VARIATIONS

55. Attempts can be categorised by:
- Whether made at the enrolment visit or at the second or third visit by a volunteer;
 - The gender of the volunteer;
 - The age of the volunteer;
 - Whether the volunteer was wearing spectacles in the case of Face and Iris systems;
 - The length of the user's pass-phrase in the case of the Voice system.
- Performance differences between these subsets have been analysed, and are reported for each system in Part II.

56. To determine the statistical significance of any observed differences (i.e. the probability of the difference being attributable to sampling error) a simple χ^2 test was used.

- a. The number of correct and failed verifications at the default threshold were counted for each class. E.g.

Observed	Under 35	Over 34	Combined
FRR	3.9%	11.5%	8.3%
Rejected	29	116	145
Verified	710	893	1603
Total	739	1009	1748

- b. If there were no difference between classes the combined error rate would apply to both classes.

Expected	Under 35	Over 34	Combined
FRR	8.3%	8.3%	8.3%
Rejected	61.3	83.7	145
Verified	677.7	925.3	1603
Total	739	1009	1748

Observed-Expected	
	-32.3 32.3
	32.3 -32.3

- c. The test statistic used is

$$\sum \frac{(Obs. - Exp.)^2}{Exp.} = (32.3 - 1/2)^2 \left(\frac{1}{61.3} + \frac{1}{83.7} + \frac{1}{677.7} + \frac{1}{925.3} \right) = 31.17$$

(The subtraction of 1/2 represents the correction for continuity; and is used because the observed values can only take integer values.)

- If all attempt results are statistically independent, the test statistic would follow a χ^2 distribution (with 1 degree of freedom). In the example case χ^2 exceeds 31.17 with probability less than 0.01%. However, this overstates the significance since there are dependencies between each attempt made by the same user.
- If all N attempts by any user had the same result (the maximum correlation possible), while attempts by different users are independent, then the test statistic divided by N follows a χ^2 distribution (with 1 degree of freedom). In the example case, if there are 9 attempts per user, the probability of χ^2 exceeding $\frac{31.17}{9} = 3.46$ is 6.28%. This understates the significance, since user attempts are not correlated to such an extent.
- Both results are shown, the true significance lies between these values.