J Forensic Sci, March 2008, Vol. 53, No. 2
doi: 10.1111/j.1556-4029.2008.00663.x
Available online at: www.blackwell-synergy.com

Ivo Alberink,[1] *Ph.D. and Arnout Ruifrok,*[1] *Ph.D.*

Repeatability and Reproducibility of Earprint Acquisition*

ABSTRACT: For all forensic disciplines dealing with identification—e.g., of glass, tool marks, fibers, faces, fingers, handwriting, speakers—in which manual (subjective, nonautomated) processes play a role, operator dependencies are relevant. With respect to earprint identification, in the period 2002–2005, the Forensic Ear Identification research project collected a database of 1229 donors, three prints per ear, and laid down a "best practice" for print acquisition. Repeatability and reproducibility aspects of the print acquisition are tested. The study suggests that different operators may acquire prints of differing quality, with equal error rates of the matching system ranging from 9% to 19%. Moreover, it turns out that "matching" earprints are more alike when taken in a consecutive row than when taken on separate occasions. This underlines the importance of (1) studying operator effects, (2) operator training, and (3) not gathering "matching" reference material at the same occasion.

KEYWORDS: forensic science, earprint identification, biometrics, validation, inter-operator effects, (standard) operating procedure

In any forensic science dealing with identification, be it of glass, tool marks, fibers, faces, fingerprints, handwriting, or speakers, in which nonautomated ("manual") processes play a role, results are possibly operator dependent. From a validation point of view (Daubert, Frye) it is important that thorough investigation takes place before performance results are relied upon in judicial practice. The current paper studies operator dependency of a semi-automated identification system for earprints.

To solidify the scientific basis for earprint/earmark identification, in the period 2002–2005 the Forensic Ear Identification (FearID) project was carried out. The project was financed by the European Union and executed by nine institutes, including police academies, universities, the Netherlands Forensic Institute and two commercial partners, over Italy, the Netherlands, and the U.K. In the three countries, a training database was gathered of 1229 donors, who donated three left and three right earprints each. Standard operating procedures were designed for the recovery and lifting of donor earprints, laid down in (1). The document contains directions and instructions guiding the technician in the collection of earprints. An important difference to earlier practice is that earprints were not gathered by making donors apply different amounts of force to special glass or flat surface. Starting from the notion that in practice a perpetrator listens for (absence of) sound, and hence uses a stable "functional force," donors were instructed to listen for a sound supplied behind a glass plate. Earprints were then recovered from the glass plate. (For nonwilling donors, this has the advantage that one may check whether "functional force" was used by inquiring into the content of what was heard.)

An operational software system was developed allowing for scanning, storing, and processing of earprints. An example of this processing is that investigators would add polylines ("skeletons") to the digitized earprint images following the imprint of the ear. From these, connected structures were determined supposedly representing the imprints, referred to as *superstructures.* An illustration of this can be found in Fig. 1. On the basis of these superstructures, further analysis was performed using image processing techniques.

Two methods were used to perform comparisons of earprints, on the basis of the width of the superstructure along its axis and on the basis of its curvature. In (2), the performance of the resulting system for making comparisons is described. For both methods combined, the equal error rate (EER) (see the section on Numerical Evaluation) of the resulting matching system was found to be 8.5% when comparing lab quality prints, and 15% when comparing lab quality prints to "simulated crime scene marks" (made on a surface of choice). Unpublished results on inter-operator effects in the clicking of the polylines suggest that the matching system as such is not dependent on whether the same or different operators add polylines to the prints (per donor).

The standard operating procedure (SOP) laid down in (1) describes in detail what equipment to use, when and how to clean surfaces of the equipment, how to dust and lift earprints, and how to instruct earprint donors. In this way, it should be guaranteed that given different equipment, investigators, locations, etc., these circumstances will not significantly influence the print quality. Two fundamental issues concerning the design of the SOP, and the applicability of resulting systems for matching of prints, are the following:

1. Repeatability: Are earprints from the same ear, taken repeatedly under the same circumstances by the same operator sufficiently similar?
2. Reproducibility: Are earprints from the same ear, taken repeatedly under different circumstances by different operators sufficiently similar?

The answers to these questions provide information about the applicability of the system, and are the ones the study at hand seeks to answer. That is to say that the report is neither about the classification of prints or performance of the operational system, which can be found in (2), nor about stability of the image processing procedure, but about the stability of print acquisition process of earprints (the dusting and lifting of the prints). Our interest lies in

[1]Department of Digital Technology and Biometry, Netherlands Forensic Institute, PO Box 24044, 2490 AA, Den Haag, The Netherlands.

*This study was carried out within the framework of the FearID project, which is a shared-cost RTD project funded under the 5th Framework Programme of the European Community, within the Competitive and Sustainable Growth Programme, Measurement and Testing Activity, Contract G6RD-CT-2001-00618.

Received 1 June 2007; and in revised form 23 Aug. 2007; accepted 9 Sept. 2007.

© 2008 American Academy of Forensic Sciences

FIG. 1—*Example: original print, clicked polyline, and calculated superstructure.*

stability of the results when different operators collect and dust the prints (reproducibility), or when the same operator goes through the procedure more than once (repeatability).

In the gathering of the FearID Main sample, prints from each donor were all taken in one single run (of six prints), so it is not possible to answer the above two questions using this large database. To be able to investigate questions 1 and 2, a separate experiment was performed.

The structure of the report is as follows: first a description is given of the experiment as it took place in Italy, the Netherlands, and the U.K. After that, a short explanation is given on feature extraction, image processing, and the performance measure used to investigate discriminative value of comparison scores. Next, the results of the experiment with respect to the questions about repeatability and reproducibility are presented. Finally, implications of the results are discussed.

Description of the Experiment

The experiment was executed in the three participating countries, Italy, the Netherlands, and the UK, by two investigators in each country. From here on we refer to the investigators as *operators*.

The SOP consisted of the acquisition, in a consecutive run, of 10 earprints (five left, five right), using the medium Black Gel Lifter. For each of the countries, the experiment consisted of each of two operators executing the SOP twice for 18 donors. That is to say, per country there were 18 donors of earprints, going through the SOP four times. Thus, per donor, 40 earprints were gathered, for 18 donors per country, amounting to $3 \times 18 \times 40 = 2160$ earprints in total. The earprints in the experiment might hence be influenced by the following factors:

1. the *country* in which the earprint was taken,
2. the *donor* of the earprint,
3. the *side* (left or right) of the earprint,
4. the *operator* (per country: investigator *A* or *B*) taking the earprint, and
5. the *number of the run* (1 up to 4) in which the earprint was taken.

For this information in a chart, see Fig. 2.

Per country, the order of the four runs was prescribed. Denoting the two respective operators by *A* and *B*, the first two runs of the SOP on donor 1 were for example taken by operator *A* and the third and fourth run by operator *B*. For every donor, a similar scheme is laid down in Table 1.

In this way, orthogonalization of the operators over the donors was realized to avoid methodological confounding factors such as bias in order, which might, e.g., lead to learning effects. The experimental design allows us to study the relevance of the mentioned factors.

Feature Extraction

On the basis of the earprint scan and the digitally added (clicked) polylines, representing a skeleton for the ear imprint, the area of the print taken up by the imprint of the ear itself is reconstructed (segmented) as in Fig. 1. This area, in the figure on the right, is referred to as superstructure and on the basis of local intensities, a medial axis is attached for it.

Two methods were used for feature extraction, the implementation of which is described in (2). In the first method, pairs of superstructures were compared on the basis of fitting them

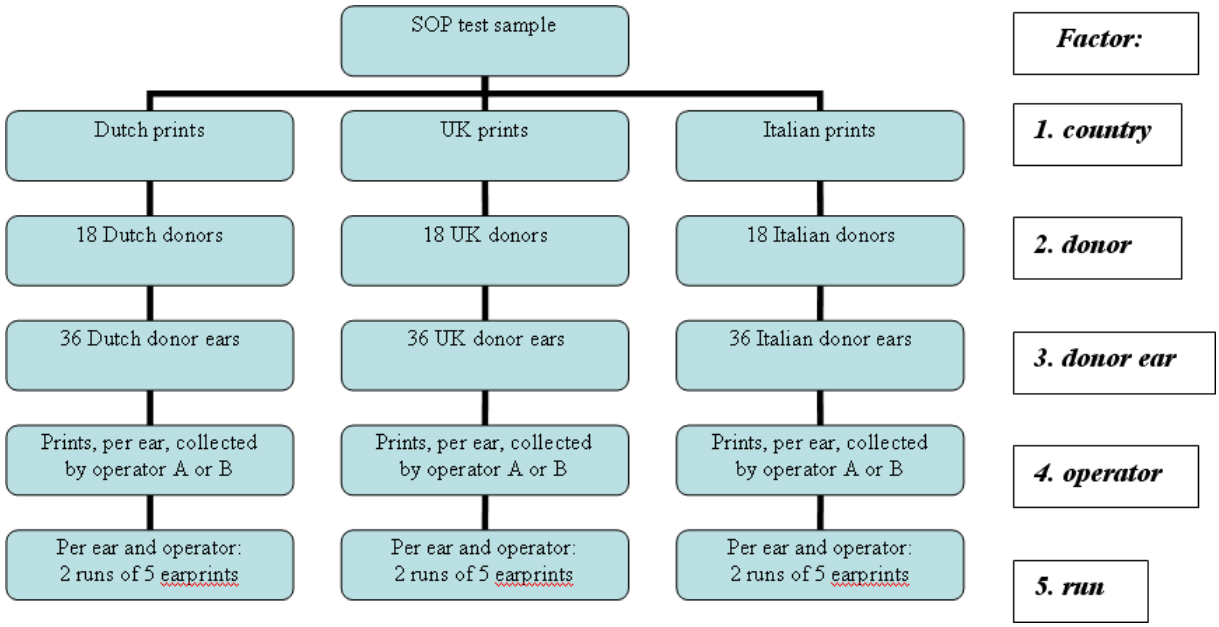

FIG. 2—*Overview of collected data.*

TABLE 1—*Scheme describing the order in which SOP runs were performed.*

Donor	Donor	Donor	Run 1	Run 2	Run 3	Run 4
1	7	13	A	A	B	B
2	8	14	A	B	A	B
3	9	15	A	B	B	A
4	10	16	B	A	A	B
5	11	17	B	A	B	A
6	12	18	B	B	A	A

SOP, standard operating procedure.

optimally on top of each other, the superstructures represented by their widths along the medial axes, weighted by corresponding image intensities. The process is referred to as *weighted width comparison*. The fitting ensures that the method is translation and rotation invariant, which makes the method suitable for comparison of superstructures obtained from prints both from the same ear and from two different ears. The second method consisted in keeping track of the angle of the medial axis with the *x*-axis of the digital image. The resulting signals were fitted on top of each other by optimal translation. Comparison using this method is translation and rotation invariant as well, and referred to as *angular comparison*.

Through training on the FearID Main sample, cf. (2), and on the basis of the statistical technique of binary logistic regression, cf. (3), for any comparison of two earprints the following one-dimensional combination of the above outcomes was determined:

$$D = 1/(1 + \exp(7.2 - 0.68/\text{ang} + 3.6\ln(\text{ww}))).$$

Here ln is the natural logarithm, ww the outcome of the weighted width comparison, and ang the outcome of the angular comparison. The discriminant score D predicts a probability that a pair of prints is matching, 1 pointing at a match, 0 at a nonmatch.

Numerical Evaluation

With respect to performance of biometric systems, a key concept is that of *verification*, or 1 to 1 comparison. The common performance measure EER will be used to test system performance and is described below.

A verification system is a classification system with two classes of outcomes: matching (or positive or acceptance) and nonmatching (or negative or rejection). Given the features in a system, for any comparison of two entities, a single value is constructed optimally summarizing the matching information. Classification takes place according to whether the outcome does or does not exceed some threshold t.

Common performance parameters with this type of system are the probabilities of making a wrong judgment, expressed in the *false rejection rate* (FRR) (cases in which the system declares a nonmatch in case of a match) and *false acceptance rate* (FAR) (cases in which the system declares a match in case of a nonmatch). As the FRR and FAR are threshold-dependent, we rather use the EER, which is by definition the (common) probability of misclassification starting from the threshold t for which $\text{FAR}(t) = \text{FRR}(t)$.

An illustration of the above is given as Figure 3 in (2). EERs allow to compare performance of the system when varying certain factors, as in the present case: donor country, operator, or run. For this, one could use the so-called receiver operator characteristics (ROC) curve as well. The ROC curve is by definition the collection of points $[\text{FAR}(t), \text{FRR}(t)]$, for all t. Comparison of pairs of ROC curves may be inconclusive though since for some fixed FAR, system 1 may have a lower FRR than system 2, whereas for another value it may be the other way around. This makes it difficult to evaluate the result of the comparison, and use of EERs is preferred.

The EERs have the inconvenient property that it is not straightforward whether an EER of 6% is significantly improving an EER of 8%. Because from a statistical point of view the EER is a complex function of two samples (matching and nonmatching comparisons), its probability distribution is unclear, even under normality assumptions on the samples, and one cannot use parametric models to approximate its behavior. A usual approach is to then use the statistical technique of bootstrapping, cf. (4), so as to obtain confidence intervals for EERs. In practice, this means that based on the two samples, a number of 1000 "re"samples are drawn from the two samples, of the same size as the original samples, to get a simulation of the probability distribution of the (estimator of the) EER. This leads to an approximation of confidence intervals. In the following, rather than reporting point estimates, we will report 95% bootstrap confidence intervals for EERs.

Results

We turn to the results of the experiment. First we explore a possible effect of the factor country on the EER of the system. As a side question, we look whether donor gender leads to significantly different results. Next we look at the effect of different donors on the EER, for the separate countries. After this we study the effect of the factors operator (reproducibility) and run of the SOP (repeatability).

We go about this as follows. First a graphical illustration of differences in outcomes for matching and nonmatching pairs of prints, given fixed states of the relevant factors, is presented by means of box plots. The latter usually provide a good illustration on whether factors have a significant influence on the discriminatory power of the variable D. We quantify the above by the corresponding EER intervals.

To reduce computational complexity, the number of nonmatching pairs of prints was downsampled to circa three times the number of matching pairs. Resulting numbers of comparisons will be reported for each step of the analysis.

Country Effect

We look at the effect of the factor country on the EER of the system. Figure 3 illustrates what happens. From left to right we look at the resulting box plots for the outcomes of D for the different countries, first looking at nonmatching, then at matching pairs of prints.

In the box plots, the boxes denote the interquartile range, their middle lines denoting the median. The whiskers show the distance from the end of the box to the largest and smallest observed values less than 1.5 box lengths from either end of the box. Outliers and extremes are not depicted in the images. Corresponding 95% confidence EER intervals are given in Table 2.

We note considerable differences between the countries. For example, in country 2, the system has a much higher discriminative power than in the other countries. This may result from the following:

1. Per country, different operators gathered the prints.
2. Per country, different donors participated in the experiment.

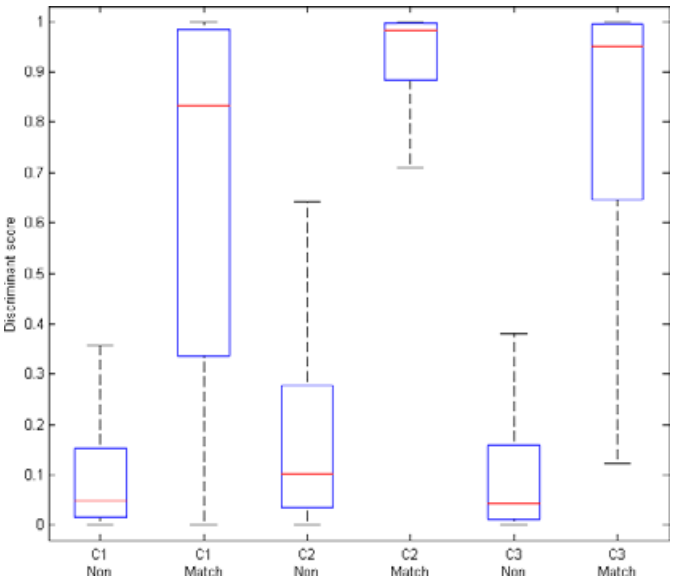

FIG. 3—*Box plots illustrating the effect of the factor "country" on the discriminant scores. From left to right results for the different countries, denoted by C1, C2, and C3, are shown, first for nonmatching comparisons (Non), then for matching ones (Match).*

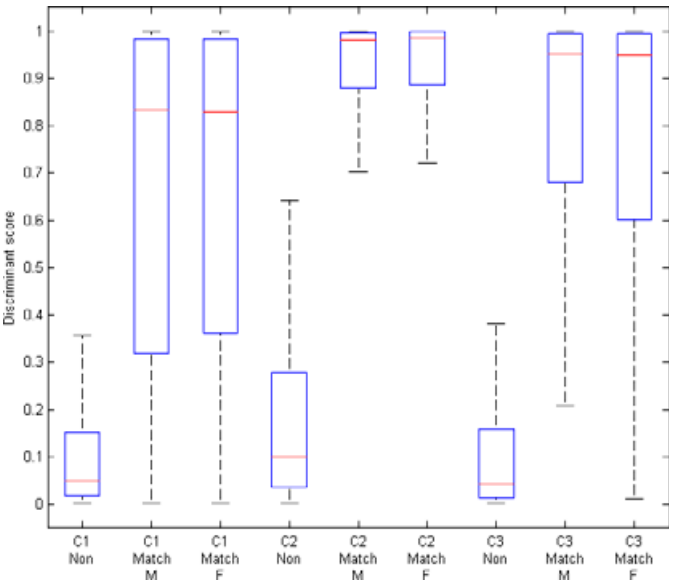

FIG. 4—*Box plots illustrating the effect of the factors "country" and "donor gender" on outcomes of the discriminant score. From left to right results for the different countries, denoted by C1, C2, and C3, are shown, first the outcomes for nonmatching comparisons (Non), then for matching ones (Match), for males (M) and females (F), respectively.*

TABLE 2—*EER 95% confidence intervals, with results divided up per country, together with the numbers of matching and nonmatching outcomes on which these are based.*

Country	Matches	Nonmatches	EER Interval (%)
1	6405	15,169	17–21
2	6802	15,300	8–11
3	6746	15,300	12–15

EER, equal error rate.

TABLE 3—*EER 95% confidence intervals, with results divided up per country and gender, together with the numbers of matching and nonmatching outcomes on which these are based.*

Country	Gender	Matches	Nonmatches	EER Interval (%)
1	M	3325	15,169	18–21
1	F	3080	15,169	17–20
2	M	4921	15,300	8–11
2	F	1881	15,300	9–11
3	M	3364	15,300	11–14
3	F	3382	15,300	13–17

EER, equal error rate.

By construction we cannot distinguish between these two factors. However, it does not make much sense that donors essentially differ per country. One reason why this could have been the case nonetheless is a difference in donor gender. In countries 1 and 3, there were nine male and nine female donors, in country 2 there were 13 male donors and five females, and this might explain the differences in results. In Fig. 4, the possibility of a gender effect is studied. Corresponding EER intervals are given in Table 3.

The differences are not significant and do not explain the differences between the results for different countries.

In country 2, experienced police officers handled the acquisition of the prints, and in countries 1 and 3 experts in biology and anthropology who received only a short training in print acquisition. Hence, the country effect is probably connected to experience of the operators.

Donor Effect

As there was a significant country effect, we furthermore separated the data from the different countries.

Next we look at the effect that individual donors have on the results. We illustrate what happens in Fig. 5, for eight different donors from country 3. The left box plot represents outcomes for nonmatching pairs of prints. Next are box plots corresponding to comparisons of matching prints of the donors 1 up to 8. The difference in discriminative power of the outcomes per donor is

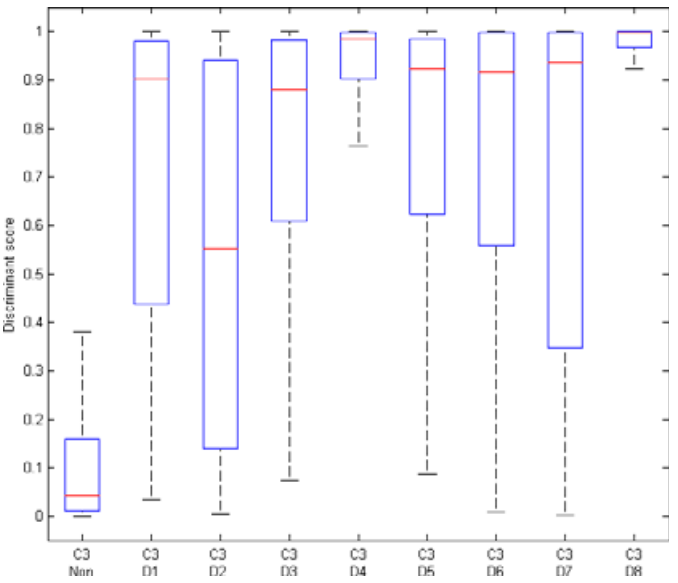

FIG. 5—*Box plots illustrating the effect of individual donors on discriminant scores. Depicted results are from country 3 (C3), first the outcomes for nonmatching comparisons (Non), then for matching ones of the individual donors 1, 2, ..., 8, respectively (D1, D2, ..., D8).*

TABLE 4—*EERs (95% confidence intervals) with results divided up per country and donor.*

Country	Donor	Matches	Nonmatches	EER Interval (%)
3	1	380	15,300	16–20
3	2	380	15,300	24–28
3	3	380	15,300	12–15
3	4	380	15,300	4–6
3	5	324	15,300	11–14
3	6	380	15,300	13–16
3	7	380	15,300	18–22
3	8	361	15,300	8–11

considerable. The corresponding 95% confidence EER intervals are given in Table 4. The above illustrates that one donor's prints will be more informative (smaller EER) than another's.

Operator Effect (Reproducibility)

We turn to the effect of the operators on resulting discriminative power, or reproducibility of the procedure. Dividing the pairs of matching prints with respect to whether they were collected by different or identical operators, in Fig. 6, we look at what happens, for the countries individually. EER intervals are depicted in Table 5.

Following Table 5, there is no significant effect (per country) of the factor operator on discriminative power of the outcomes of discriminant scores.

Run Effect (Repeatability)

We have established that per country, operators obtain comparable EERs, and matching prints are not more alike if taken by the same operator. Next we look at the effect of different runs on resulting discriminative power, or repeatability of the procedure.

TABLE 5—*EERs (95% confidence intervals) when results are divided up with respect to country and operator situation.*

Country	Operator Situation	Matches	Nonmatches	EER Interval (%)
1	1	816	15,169	20–24
1	2	865	15,169	17–21
1	Mixed	3360	15,169	19–22
2	1	895	15,300	8–11
2	2	895	15,300	7–10
2	Mixed	3580	15,300	10–13
3	1	890	15,300	14–17
3	2	885	15,300	13–17
3	Mixed	3551	15,300	13–16

EER, equal error rate.

With respect to "operator situation": by 1 and 2 it means that matching prints were both collected by operator 1 or 2, respectively, by "mixed" that matching prints were gathered by different operators. Pairs of matching prints that were gathered in the same run of the standard operating procedure were left out of the analysis.

Here the idea is that if matching prints are gathered consecutively they might be more alike. Dividing all pairs of matching prints according to whether they were collected in distinct runs or in the same run, in Fig. 7, we study the relevant box plots for the countries individually. Per country, we distinguish between nonmatching prints, matching prints from different runs of the SOP (undifferentiated with respect to "operator situation"), and matching prints from the same run of the SOP, the latter split up according to the operator.

There is a clear tendency of discriminative power being higher when matching prints are from the same run. In EER intervals, see Table 6.

In all of the countries, the system performs significantly better (EERs are smaller) when matching prints have been gathered in consecutive runs of the SOP.

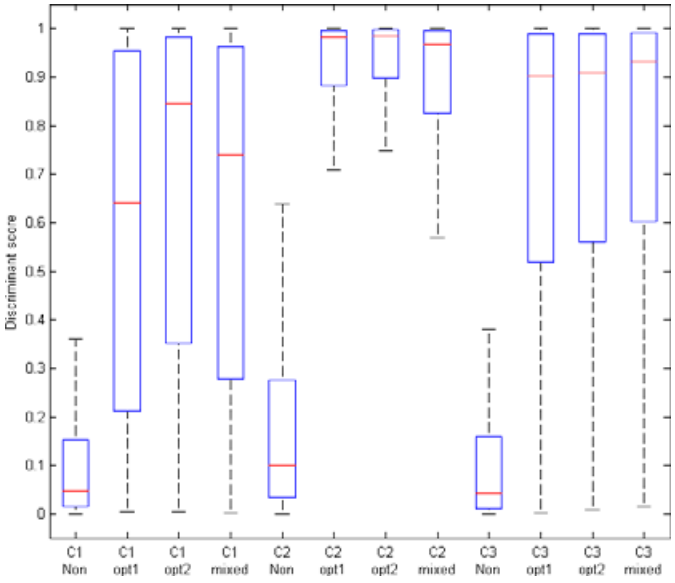

FIG. 6—*Box plots illustrating the effect of individual operators on discriminant scores, for the separate countries individually. From left to right results are shown for the different countries, denoted by C1, C2, and C3. Per country, first the outcomes are given for nonmatching comparisons (Non), then for matching ones, gathered by operator 1 only (opt1), by operator 2 only (opt2), and by both operators (mixed), respectively. None of the matching prints was gathered in the same run of the standard operating procedure (SOP).*

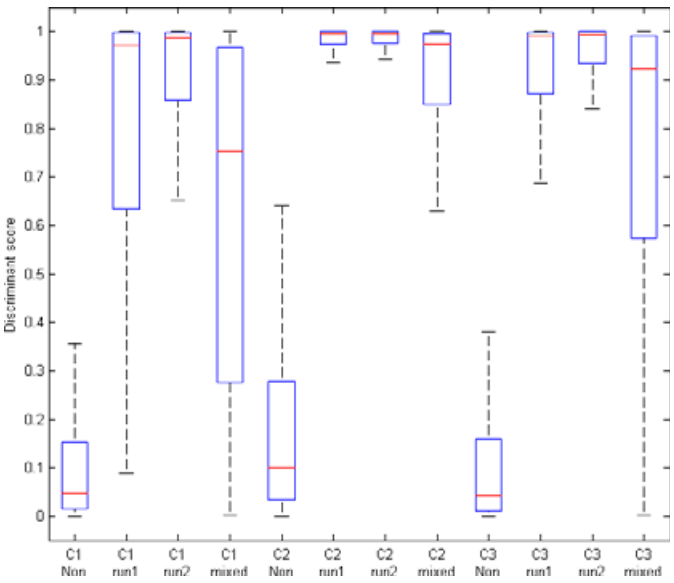

FIG. 7—*Box plots illustrating the effect of taking consecutive runs on discriminant scores, for the separate countries individually. From left to right results are shown for the different countries, denoted by C1, C2, and C3. Per country, first the outcomes are given for nonmatching comparisons (Non), then for matching ones, gathered by operator 1 in the same identical run (run 1), then the same for operator 2 (run 2), and then for matching prints from different runs (either from the same or separate operators) (mixed) respectively.*

TABLE 6—*EERs (95% confidence intervals) when results are divided up with respect to country and "run situation."*

Country	Run Situation	Matches	Nonmatches	EER Interval (%)
1	1	672	15,169	11–13
1	2	692	15,169	11–14
1	Mixed runs	5041	15,169	19–22
2	1	716	15,300	5–7
2	2	716	15,300	4–7
2	Mixed runs	5370	15,300	9–12
3	1	712	15,300	8–10
3	2	708	15,300	9–11
3	Mixed runs	5326	15,300	13–16

EER, equal error rate.

Run situations are denoted by 1 or 2 if the matching prints were gathered in the same run of the standard operating procedure, by operator 1 or 2, respectively. "Mixed runs" means that matching pairs of prints were gathered in different runs.

Conclusions and Discussion

The objective of the experiment was to study repeatability and reproducibility of the acquisition process of earprints. To this end, the possibility of country, donor, operator, and run effects was considered on the basis of an image processing based approach. We found a significant effect of the operator country on the EER performance of the comparison scores, probably due to experience of the operators. For the countries individually, no operator effects were found in the sense that matching prints from different "runs" of the SOP, taken by the same operator, do not produce results more alike than those taken by different operators. On the other hand, matching prints from the same run of the SOP lead to results that are significantly more alike than those taken in distinct runs (run effect). Different donors have largely differing behavior of comparison scores for matching prints: one person may leave consistently "better" earprints than another.

Besides via image processing routines, comparison of prints was implemented on the basis of anatomical annotation, by anthropological experts. Results of the combined approaches are described in (2). Inter-operator effects in the anatomical annotation of earprints are studied in (5). Here, it turns out that the resulting classification system is unstable in cases where operators are varied, in the sense that smaller EERs are achieved when identical operators annotate matching earprints than when different operators do this. Because of this, in this paper anatomical annotations were not used in the classification process. This does not mean that the authors are of the opinion that earprint identification on the basis of anatomical annotation by experts does not work. It does mean that the FearID project was unsuccessful in objectifying the expert knowledge through pattern recognition on the basis of annotated points, with predefined labels attached. It makes sense to test expert knowledge through proficiency tests. The reader interested in the biological or anthropological side of earprint comparison is referred to (6) and (7).

We compare the EERs found to the ones encountered in the larger main FearID database. As stated in the introduction, this consists of 1229 donors donating three left and three right earprints each. Performance was determined on a validation subset of 458 donors. The database was gathered in the same countries by the same operators as the current experiment. With respect to matching pairs of prints, prints were always gathered in a single run by a single operator. As a result, we compare the results from the Main sample to the ones in the current study for matches from the same run. The results are shown in Table 7.

TABLE 7—*Comparison of EER intervals per country for both the sample of the current SOP experiment and the main FearID validation sample.*

EER intervals	Country 1	Country 2	Country 3
Current experiment	11–13%	5–7%	8–10%
FearID validation sample	14–17%	8–11%	11–14%

SOP, standard operating procedure; EER, equal error rate; FearID, Forensic Ear Identification.

Table 7 suggests a continuing country effect, with worse system performance on the larger validation sample than on the sample of the current experiment. Performance on the validation sample is probably more relevant (operators having gathered experience and the situation seems to resemble reality more). Next to this, the current study suggests run effects which will have influenced EERs for the Main sample as well. The EERs reported in (2) are therefore probably better (smaller) than the ones obtainable in an operational environment. On the other hand, the threshold values for the classification acquired on the basis of the training sample will be too conservative.

In the current setting of earprint comparison, with respect to questions about repeatability and reproducibility, we find definite country, run, and donor effects. This underlines the importance of studying operator effects in the forensic field. Moreover, it shows that it is worthwhile that operators are well trained and experienced in print acquisition. Finally, it shows that when gathering "matching" traces for a database, it is worthwhile to not gather them all at the same time.

Acknowledgments

Authors would like to thank Cor van der Lugt and Ruud van Basten (LSOP-ICR, The Netherlands), Marta Giacon and Francesca De Conti (University of Padova, Italy), Zale Johnson (National Training Centre for Scientific Support to Crime Investigation, Durham, U.K.) and Sarah Sholl (University of Glasgow, U.K.) for gathering and annotation of the sample. Authors would also like to thank one of the anonymous referees for useful comments on content and layout of the paper, and Gerda Edelman for reviewing the paper.

References

1. Johnson Z. Standard operating procedure for the taking of earprints. Internal FearID report, 2003; http://forensic.to/fearid/Procedure.doc (accessed Feb. 14, 2008).
2. Alberink I, Ruifrok A. Performance of the FearID earprint identification system. Forensic Sci Int 2007;166:145–54.
3. Pampel FC. Logistic regression: a primer. Sage university pagers series on quantitative applications in the social sciences, series no. 07-132. Thousand Oaks, CA: Sage Publications, 2000.
4. Efron B, Tibshirani RJ An introduction to the bootstrap. Monographs on statistics and applied probability 57. New York: Chapman & Hall, 1993.
5. Alberink IB, Ruifrok ACC, Kieckhoefer H. Inter-operator test for anatomical annotation of earprints. J Forensic Sci 2006;51:1246–54.
6. Meijerman L, Sholl S, De Conti F, Giacon M, Van der Lugt C, Drusini A, et al. Exploratory study on classification and individualization of earprints. Forensic Sci Int 2004;140:91–9.
7. Meijerman L. Inter- and intra-individual variation in earprints [dissertation]. Leiden: Barge's Anthropologica, 2006.

Additional information and reprint requests:
Ivo Alberink, Ph.D.
Netherlands Forensic Institute
Department of Digital Technology and Biometry
PO Box 24044
2490 AA Den Haag
The Netherlands
E-mail: i.alberink@nfi.minjus.nl