Forensic Science International 267 (2016) 145–156

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Fact or friction: Examination of the transparency, reliability and sufficiency of the ACE-V method of fingerprint analysis

Sarah V. Stevenage *, Christy Pitfield

Department of Psychology, University of Southampton, UK

ARTICLE INFO

Article history:
Received 27 April 2016
Received in revised form 12 July 2016
Accepted 17 August 2016
Available online 26 August 2016

Keywords:
Fingerprint analysis
ACE-V method
Experts
Fingerprint training

ABSTRACT

Three studies are presented which provide a mixed methods exploration of fingerprint analysis. Using a qualitative approach (Expt 1), expert analysts used a 'think aloud' task to describe their process of analysis. Thematic analysis indicated consistency of practice, and experts' comments underpinned the development of a training tool for subsequent use. Following this, a quantitative approach (Expt 2) assessed expert reliability on a fingerprint matching task. The results suggested that performance was high and often at ceiling, regardless of the length of experience held by the expert. As a final test, the experts' fingerprint analysis method was taught to a set of naïve students, and their performance on the fingerprint matching task was compared both to the expert group and to an untrained novice group (Expt 3). Results confirmed that the trained students performed significantly better than the untrained students. However, performance remained substantially below that of the experts. Several explanations are explored to account for the performance gap between experts and trained novices, and their implications are discussed in terms of the future of fingerprint evidence in court.

1. Introduction

Fingerprints are commonly accepted by lay people as a highly valuable and reliable means of identification. Traders in early China used their fingerprints in clay seals, or on silk or paper to legitimise documents or loans, and by C13th, Eastern doctors noted the use of fingerprints to identify people. However, it was not until Sir Francis Galton published his classification of fingerprint patterns that they began to attract attention in the forensic community [1]. Resting on the principles of persistence and individuality [2], fingerprint matching has been relied upon in court since 1892. However, today, fingerprint evidence is coming under considerable scrutiny, and questions are being raised about its admissibility in court [3]. The purpose of the present paper is to examine the practice of fingerprint analysis, as described by qualified fingerprint experts, and to determine whether that practice is both reliable and sufficient in supporting accurate performance.

1.1. Reliability of fingerprint evidence

Fingerprint analysis relies upon two essential principles. The first is persistence or permanence, which suggests that even though ridge definition may become less distinct through habit or occupation, fingerprints do not fundamentally change with time. The second is individuality or uniqueness, which suggests that no two people as yet have been discovered to have the same pattern of friction ridges. On this basis, fingerprint evidence has assumed a dominant role in court, with bold statements suggesting that fingerprint analysis is 'infallible' [4], and that errors are 'virtually impossible' [5,6]. Consequently, fingerprint evidence has earned a reputation for 'accuracy and objectivity' ([7], cited in [8]). Despite this, concern is now mounting over the accuracy of the human operators who bear the responsibility for courtroom evidence [9,8]. Two high-profile cases illustrate the problem:

The first case relates to Brandon Mayfield who, in May 2004, was arrested in connection with the Madrid bombings. Investigators had retrieved a fingerprint from the bag which held the detonators, and an AFIS search had produced a list of comparison fingerprints. Mayfield's fingerprint was the fourth on the list and a match was confirmed by an FBI analyst, and was subsequently verified by three fingerprint examiners. Worryingly, discrepancies between the latent print and the Mayfield's fingerprint were

* Corresponding author at: Department of Psychology, University of Southampton, Highfield, Southampton, Hampshire, UK. Fax: +44 2380 594597.
E-mail address: svs1@soton.ac.uk (S.V. Stevenage).

http://dx.doi.org/10.1016/j.forsciint.2016.08.026

discounted, and a blind verification procedure was not conducted. Dissatisfied with the FBI's claim of a match, the Spanish police kept the investigation open and ultimately, they arrested Ouhnane Daoud whose fingerprint was matched to the latent print. After two weeks in custody, Brandon Mayfield was released and the FBI admitted an error (see report issued by the Office of the Inspector General into the FBI Mayfield Case; [10]).

The second case relates to Shirley McKie, an investigating officer working on a Scottish murder case. She was identified from a latent print found at the crime scene, and was charged for perjury at denying that the latent print could have belonged to her. Whilst a jury unanimously cleared McKie of the perjury charge, the question of the fingerprint match remained a matter of controversy. Disagreement existed over the number of points of correspondence required to determine an identification, and concern was raised over the lack of a blind verification procedure. In 2011, the Scottish Public Judicial Enquiry into Fingerprinting concluded that a misidentification had occurred (see report issued by [11]).

The concerns raised by both cases are reflected in a number of notable high-profile reports which collectively throw doubt over the reliability of fingerprint evidence. These include the National Academy of Sciences report [12] which called for a committed programme of research to define the sources of variability and bias in human cognitive decision-making. This call is echoed more recently in reports published by the National Institute of Standards and Technology Expert Working Group on Human Factors in Latent Print Analysis [13] and the UK Forensic Science Regulator [14], both of which highlight cognitive bias as a major source of human error (see also the POSTbrief to the UK Houses of Parliament: [15]). With this in mind, researchers have begun to scrutinise both the process and the performance of fingerprint experts in order to address the challenging question of whether fingerprint experts 'possess abilities beyond those of a novice' [15, p. 14].

1.2. Performance of fingerprint analysts

Fingerprint analysis is conducted by trained fingerprint analysts who follow a widely accepted description of good practice known as the ACE-V method. This acronym describes the four stages of *analysis*, *comparison*, *evaluation* and *verification*, and it results in one of four outcomes: The fingerprint pair may be deemed of 'no value' (too poor to analyse), or may be deemed to be 'individualised' (come from the same source), 'excluded' (come from different sources) or 'inconclusive' (insufficient evidence exists to draw a conclusion). The method has been described by Huber [17], Ashbaugh [18] and Champod et al. [19]. Nevertheless, several authors criticise the method for being underspecified and thus untestable in a robust fashion [3,20]. In particular, there is concern over the lack of agreement, both across nations and across time [21], regarding the number of points of correspondence required to determine a 'match'. Whilst standards have varied between 7 (Russia), 12 (Spain) and 16 points of correspondence (UK) [21], current practice leaves the determination of sufficiency of correspondence to the expertise of the analyst. As a consequence, it is difficult to fully specify the inferential process followed by the expert.

Despite this, formal testing of the accuracy of fingerprint experts suggests that performance is good. For instance, Langenburg [22] reviewed a number of early studies suggesting performance levels ranging from zero false positives and only 10/1170 false negatives by 130 participants [23] to only 2/5861 errors across 92 participants, both of which were detected at verification [24]. Moreover, the results of proficiency testing in Australia revealed only 2 transcription errors (which were corrected on checking) and only 7 'inconclusive' decisions out of 782 comparisons over a 6 year period [25]. Whilst impressively high, these data do however reveal that the performance of fingerprint experts is not perfect [26]. Notably, experts outperformed novices, but tended to err on the side of caution with more false negatives (missing a true match) than false positives (wrongly accusing an innocent suspect). In his own tests, Langenburg [22] revealed 3 false negative decisions across 6 experts when they were the sole analyst for 60 comparisons. However, when the experts knew that their decisions were to be verified by a colleague, their performance across just 30 comparisons was weaker, revealing 2 false positive errors (all of which were corrected at an unblind verification stage) but 6 false negative errors across the 6 experts (none of which were corrected). This pattern was replicated by Ulery and colleagues [26,28] with the latter study revealing that these false negative decisions can be repeated by the same analyst upon retest, and can be reproduced by different analysts on blind checking. Evidence also exists to suggest that experts may be biased by the opinion of others. For instance, Langenburg et al. [29] showed that the performance of experts was affected in the direction of opinions presented as if from another examiner or from an internationally recognised expert. Whilst the participants in this study did not necessarily err in terms of making definitive errors, they did err through a greater tendency to reach an 'inconclusive' decision.

A more accurate picture of expert performance may, however, be obtained from studies in which the experts are unaware of being tested, and this has been achieved through inserting test trials into their normal workload. Researchers suggested that this may address the tendency that experts may otherwise have to exercise caution when under scrutiny [30]. When this approach was taken, a number of studies have revealed worrying levels of performance bias. For instance, Dror et al. [31] tested 5 fingerprint experts with what they believed to be the Mayfield and Madrid bomber fingerprints, yielding a strong bias for the experts to conclude that the fingerprints were not a match. In fact, the fingerprints had been previously examined by the experts and confirmed as a match. Nevertheless, with this strong contextual bias, four of the five experts reversed their previous decision, with three experts reporting the fingerprint pair to not match, and one expert reporting that the evidence was insufficient to draw a conclusion. In a similar study, Dror and Charlton [32] asked experts to judge fingerprint pairs given the contextual bias that the suspect had either confessed to the crime (bias to say 'match') or had an alibi (bias to say 'no match'). The results indicated that when the fingerprint pair was similar and thus the task was difficult, the experts were swayed in the direction of the biasing information. Finally, Dror et al. [33] demonstrated that experts can even be swayed by the introduction of technology to the task. When candidate fingerprints were sourced by an automated fingerprint identification system (AFIS), experts made more false positive identifications to those fingerprints towards the top of the AFIS list where the similarity between latent and target fingerprints was presumed to be greater. Moreover, this error arose despite the fact that the true suspect print was present lower in the list. A meta-analysis of key results demonstrated that the effects were sufficient to be of statistical concern, indicating that fingerprint experts are good at their task but they are not totally reliable, and can be biased by context [34]. Indeed, they concluded that the threat of bias is real and unconscious, and may be of particular concern when a procedure is only vaguely specified [9,35].

1.3. Transparency of the matching process

The latter work raises particular criticism of the specificity with which the ACE-V method is described. Indeed, this is a concern voiced by Haber and Haber [20] who suggest that the ACE-V method fails to meet the standards of admissibility demanded of the court in terms of both the definition of the method itself, and

the definition of error rates. (See [3] for an excellent review of the extent to which the ACE-V meets the Daubert criteria for admissibility.) Whilst Champod [36] agrees with Haber and Haber's [20] concern, he notes that a lack of transparency need not imply a lack of reliability. Nevertheless, Champod and the Habers agree that greater visibility of the experts' inferential process would be of value both when reaching and verifying a fingerprint decision, and when explaining that decision to a court of lay individuals. This calls for the sort of approach that Ulery et al. [27] referred to as 'white box' research, in which the focus is on the *process* taken by the experts, rather than the 'black box' *outcome* at the end of it.

One step in this direction is provided by Charlton et al. [37] who provided a qualitative analysis of the comments made by fingerprint experts during their task. Their findings revealed a fear of making errors amongst experts, but also revealed the considerable emotional satisfaction experienced when they were able to make a match on a high-profile, serious, or long-running case. Whilst these insights are valuable in terms of the emotional pressures and resultant motivations of the experts, they do not describe the 'white box' process that Ulery et al. envisaged. The purpose of the present study is to address this gap.

A mixed methods approach is taken in the current work with Experiment 1 providing a qualitative view of the fingerprint matching process from the perspective of the expert. With the process made transparent, Experiment 2 then provides an empirical test of the reliability of that process. Finally, and critically, Experiment 3 explores the question of whether the process as described is sufficient to enable a set of trained novices to perform at expert levels, or whether indeed fingerprint experts rely on something more. These questions of transparency, reliability and sufficiency are key in starting to address the concerns that have been levelled regarding reliability and admissibility. As such, the current work aims to provide the forensic science community with greater insight over the inferential process used by fingerprint experts, and greater clarity over training needs to support robust and reliable court decisions.

2. Experiment 1: Method

2.1. Design

A semi-structured interviewed was used to explore the process used by fingerprint experts when analysing a pair of fingerprints. This was followed by a 'think aloud' task to narrate the process with a pair of fingerprints. Thematic analysis was used to extract the common themes in the experts' process, together with any idiosyncrasies reported.

2.2. Participants

A total of 12 expert fingerprint analysts (6 males, 6 females) were recruited from a nearby UK Fingerprint Bureau. Ages ranged from 33 years to 56 years (mean age = 39.6 years, SD = 6.7), and all were fully trained current practitioners. Experience was recorded in terms of the number of years in role, and ranged from 2 years to 24 years (mean experience = 12.16 years, SD = 6.67). All participants took part with their supervisor's permission and were tested at their place of work.

2.3. Materials

A semi-structured interview schedule was used providing a guide for all interviews. In this way, participants were asked about (i) the fingerprint characteristics used for analysis, and (ii) their method of analysis itself.

2.4. Procedure

Participants were informed of the nature of the task, and assured of anonymity of reporting. Following provision of informed consent, participants were interviewed individually within a quiet room at their place of work during a work break. The interview began with a 'think aloud' task in which participants were asked to detail their route from the car park to their desk, and to report the number of doors that they passed through. This enabled the participants to settle into the task of articulating their actions in sequence. Subsequently, participants were asked to provided demographic details before describing the process they used when determining whether a pair of fingerprints represented a match or not. Participants were prompted if necessary to describe the most important characteristics, as well as any misleading characteristics, of the fingerprint. Following this, participants were provided with a pair of fingerprints on paper and were asked to narrate their process with reference to that pair of fingerprints. At no time was any reference made by the experimenter to the ACE-V method of analysis, ensuring that any methodological structure was provided by the experts rather than being superimposed by the interview schedule.

The interviews were transcribed verbatim, and an inductive thematic coding approach [38] was applied in order to extract common themes (and idiosyncrasies) across the experts. Data immersion enabled familiarisation with the overall dataset and supported the identification of recurring patterns in the data. A coding scheme was generated and applied to the full dataset, and from these codes, common themes were identified, extracted and named. Inter-rater agreement confirmed the appropriateness of the resultant thematic structure. Moreover, after the 12th expert was interviewed, no new themes were emerging. Thus the sample of 12 experts was considered adequate for the purpose at hand.

3. Results and discussion

Overall, the data provided two main themes: *method of analysis*, and *fingerprint characteristics*. The *method of analysis* described the staged approach that analysts took when reaching a decision on a pair of fingerprints. Spontaneously, all experts described the 4 stages of analysis, comparison, evaluation, and validation (ACE-V) and these thus represented the sub-themes. The *fingerprint characteristics* described the details of a fingerprint that the experts looked for, and these were described by 4 sub-themes (first level detail, second level detail, third level detail, scars and creases). The remainder of this section describes and illustrates the experts' process with reference to quotes.

3.1. The ACE-V method of analysis

All participants were asked to describe their methodology when comparing one fingerprint to another. In response, all referenced the ACE-V method, and provided detail on each stage in turn. Particular emphasis was placed on the 'analysis' stage which was often referred to as an information gathering stage at which the latent print (the crime scene print) is assessed both for its quality and for the information that it yields. Participants noted the possibility that the latent print may be too poor to analyse further.

> Essentially you are looking for what information you can find within the [latent print].[1] It could be that you could see a

[1] The individual experts tended to use different terms such as 'fingermark', 'crime scene mark', 'fingerprint' or simply 'print' to refer to the latent print obtained from the crime scene. For consistency and readability, the term 'latent print' is always used, and changes in terminology within the experts' quotations are indicated in square brackets.

pattern. It could be that you could actually see individual sweat pores. It could be that you could see very little – just friction ridges – just the actual lines. So, I am ascertaining what that quality is, and how much of it is there." (Participant 9, Female, 30 years old, 5 years of experience).

Participants also emphasised the importance of assessing the latent print on its own at the analysis stage.

"You take the [latent print] as an individual mark and assess it on its individual merits." (Participant 5, Female, 41 years old, 22 years of experience).

When describing the *comparison* stage, participants again stressed the importance of focussing on the latent print before considering the suspect print. The majority of participants spontaneously and confidently reported that they worked in this manner.

"But you would be analysing the poorer impression, and then... and taking the comparison between that and the better impression." (Participant 9, Female, 30 years old, 5 years of experience).

Comparison of the two fingerprints was described holistically by participants.

"You are looking for dissimilarities as well as similarities. So it is almost a bit like dot-to-dot, or spot-the-difference if you like." (Participant 5, Female, 41 years old, 22 years of experience).

However, the 'think aloud' task enabled participants to be much more descriptive regarding comparison process:

"OK, so yeah, this again... Like we were saying, the pattern is the first thing your eyes are drawn to and these [the two fingerprints used in the think aloud task] are both very similar patterns... and the deltas are both sort of equal distance from the core to the delta. You can see that they are both quite similar. Umm, what I tend to look at – my eye goes straight to the core. Then I look for unique, umm, a few characteristics around the core that my eye is drawn to." (Participant 1, Female, 37 years old, 7 years of experience).

When describing the *evaluation* stage of the process, participants considered this to be a reflection on the previous comparison stage enabling them to reach a decision as to which of three outcomes they would endorse: The pair of fingerprints is either declared to be a match; is declared not a match; or the data are deemed to be inconclusive.

"I'll then move to my decision-making – the final stage – which is the evaluation. So I am kind of considering 'have I got agreement', or 'have I got disagreement'. If there is agreement, is it in sufficient quantity to be able to step over a line and say 'yes' I have identified a [latent print of a finger or palm]' or do I have to say 'no, I can't identify them – it's not identified'? Or perhaps I can't say for definite – it's inconclusive." (Participant 9, Female, 30 years old, 5 years of experience).

Finally within this theme, participants all described the *verification* stage during which a second expert completes the assessment, comparison and evaluation stages to determine whether the same conclusion is reached. As this stage falls outside the activities of the first expert, this stage is not reflected on any further.

3.2. Fingerprint characteristics

Having discussed the process in general terms, the participants then described the fingerprint characteristics that they relied upon for the *analysis* and *comparison* stages. In this respect, participants described first level detail, second level detail, third level detail, and scars and creases.

First-level detail was emphasised by all participants. It was described as the overall fingerprint pattern, and participants described the defining characteristics of the whorl, loop and arch.

"The classification is the formal term for the type of pattern you can see. So it is the whorl – which goes round in a 360, often misquoted as 'swirls', but the whorls are the circular ones. The arches look like the underside of a bridge. And the loops have a direction to them – either the left or the right hand side. These are probably the 3 easily recognised and very different types of pattern." (Participant 4, Male, 33 years old, 7 years of experience).

Participants also described the importance of the directionality of the pattern, emphasising the fact that loops may start either from the left or the right, and whorls may spiral either clockwise or anticlockwise.

"Loops that have a direction to them – to either the left or the right hand side." (Participant 5, Female, 41 years old, 22 years of experience)

"... the configuration of the ridges in the core of the pattern – whether they have got like an anticlockwise spiral or a clockwise spiral..." (Participant 3, Male, 41 years old, 10 years of experience)

The classification of a latent print by its pattern was described as fundamental to the subsequent *comparison* stage. During comparison, the concordance of pattern between latent print and suspect fingerprint was emphasised, and this determined whether the participants needed to proceed any further.

"Usually, if the patterns don't match, it's pretty straightforward to make that call, in which case that pair can be discarded, or you can move onto the next fingerprint or the next person depending on how you are going. So, if you've got a loop, and the chap you're looking at has got all arches, you can say 'No' pretty quickly. If the patterns match, then you need to go on to the next level of detail." (Participant 11, Male, 34 years old, 2 years of experience).

Participants also described what they referred to as the 'core' and the 'delta' within the overall fingerprint pattern. All participants noted the importance of locating the delta, and comparing the distance between the delta and the core across the two fingerprints under consideration at the *comparison* stage.

"you don't get a delta on an arch because of the nature of the pattern, but when you have a loop or a whorl type pattern, the way the ridges are formed and the directions that they take, they can form what's called a delta, which is where three ridges will form to a point – and that's known as the delta. And on a loop pattern, you'd have one delta. And on a whorl pattern, you'd have two deltas. And where those deltas are in relation to what's called the core of the pattern, which is the tightest innermost re-curve of the loop of the ridges, then you can use that to help with your comparison and make a decision whether they're identified or not by counting [ridges] from the delta to the core. So in relation to where the delta is to other parts of the pattern, they can be used for identification." (Participant 12, Male, 56 years old, 24 years of experience)

Second level detail: All participants described the second level of detail that they relied upon when examining a latent print and

when making a comparison. These second level details included bifurcations (splits in individual ridges) and ridge endings.

> "Once you've established that the pattern is the same, and the distance between the core and deltas is the same, you can look at second level detail which is the ridge endings and bifurcations – so basically, just where the ridges stop and split – and what you're looking to see is that they're in the same relative position to each other." (Participant 4; Male; 33 years old; 7 years of experience)

Third level detail: All participants mentioned third level details in their fingerprint analysis. This refers to features such as sweat pores. Whilst a match (or otherwise) may be determined without this level of detail, participants were keen to describe both the value and the limitations associated with this third level of detail.

> "Then you could move onto something called third level detail which is looking at individual ridge units perhaps or looking at the relative positions of pores, which would be poreoscopy and/or the shape of the edges of the ridge, which is something called edgeoscopy, which we don't do very much because usually we can make an identification based on the pattern and the ridge endings and bifurcations." (Participant 8; Female; 44 years old; 14 years of experience)

Creases and scars: Finally, all participants reflected on the value (or otherwise) of features such as creases and scars. Their comments centred on the issue of availability of the features given variation in the methods of obtaining fingerprints, and variation in the resultant fingerprint quality.

> "You could have creases available. Obviously you have creases in your hands where there's flexion, so on the inside of the knuckles and also across the palm you end up with a set of creases which run in certain directions. Also due to age and use of the hands you'll end up with additional creases going in, which if they're replicated they can be used; again it's not something you can rely on totally, but it is something else that can be used to back up your conclusions. Creases don't always transfer because you're looking at a control method, which is the fingerprint form; and the [latent print] itself, which is a chance impression – it's not taken with any kind of control; so it's possible that creases aren't necessarily replicated in both." (Participant 10; Female; 39 years old; 10 years of experience)

In addition, the fundamental issue of permanence was raised given that such features are acquired following age or insult during an individual's lifetime, and can heal to differing degrees, threatening their appearance at two discrete points in time.

> "In terms of scars, if it's a deep-seated scar... yeah, then that's something that would be very useful because you might see that there's a scar on your fingerprint form, but obviously if the fingerprints are taken ten years ago and the 'lift' is taken ... ten minutes ago they might have scarred their finger in the meanwhile so there might be a scar on your [latent] print and it'll be perfectly okay on the [fingerprint] form, so it depends if the permanent scar exists in both [finger]prints." (Participant 2; Male; 42 years old; 17 years of experience)

4. Discussion

The results of Experiment 1 have been instructive in describing, in qualitative terms, the process undertaken by a fingerprint expert when examining a pair of fingerprints, and the characteristics that they relied upon. Across the 12 experts interviewed, all described the ACE-V method without prompting, and all showed a high level of consistency in the description both of the stages of the method, and the of the fingerprint characteristics under examination. No idiosyncrasies were detected once variation in the manner of wording had been accounted for by the coding scheme. As such, these results provided two important benefits: They provided a level of transparency to the ACE-V method, and they provided a level of confidence that the method was consistently understood and described across the fingerprint analysts' community at least within this site. The current results also sit well alongside the research of Charlton et al. [37] whose qualitative approach focussed more on the emotional motivations and reactions of experts during their task. Together, these two qualitative studies provide a valuable insight into the inferential process, and the task-related pressures, faced by the experts.

Whilst greater transparency and consistency of usage of the ACE-V method will certainly help in cases involving an admissibility challenge in court, it is worth reflecting on the comments of Champod [36] who noted that the absence of transparency, and indeed, the absence of consistency of usage, need not indicate that the experts' conclusions are unreliable. Indeed, he is at pains to stress that the ACE-V method is a process of 'good practice' rather than a prescriptive regime. Viewed through this lens, the demonstration here of transparency and of consistency of the ACE-V method is valuable only in terms of a perception of the *process* but should not affect the evaluation of the *outcome*. This said, demonstration of consistency across experts in the description of their process is worthless if they do not actually apply that process towards a reliable outcome. The purpose of Experiment 2 was to see whether the ACE-V method was consistently applied, regardless of experience, to yield reliable decisions.

5. Experiment 2: Method

5.1. Design

A 2×3 within-participants design was used in which the effects of trial type ('same', 'different') and fingerprint pattern (whorl, radial loop, ulnar loop) were explored on a fingerprint matching task. Participant accuracy and speed of response were recorded and represented the dependant variables.

5.2. Participants

The expert participants were the same as those used in Experiment 1.

5.3. Materials

Fingerprint images were drawn from the Biosecure Database which consisted of 8 samples of a single fingerprint from 100 individuals. The 100 individuals were classified by the authors into fingerprint pattern types yielding three dominant groups corresponding to plain whorls ($n = 27$), radial loops ($n = 31$), and ulnar loops ($n = 35$). A total of 12 individuals were selected from each of the three pattern types (yielding 36 in total) and these represented the target fingerprints. Each target fingerprint was depicted by two images representing a good quality image (simulating the 'suspect fingerprint' obtained from the custody suite or the 10-card), and a relatively poor quality or partial print (simulating the 'latent print' obtained from the crime scene).

In addition, 6 individuals were selected from each of the 3 pattern types (yielding 18 in total), and these represented the foil fingerprints for use on 'different' trials.

Finally, 12 individuals were selected spanning the 3 pattern types and these represented the practice fingerprints for use during the orientation stage of the task. Four of these were used to construct 'same' trials, and thus two images were obtained as

above. The remaining 8 stimuli were paired to produce 4 'different' trials, with care taken to match the fingerprint pattern across the pairs.

The fingerprint matching task consisted of a practice phase of 8 trials (4 'same', 4 'different') and a main phase of 72 trials. In constructing the main phase of the task, each of the 36 target fingerprints was presented twice, once in a 'same' trial, and once in a 'different' trial. For 'same' trials, care was taken to ensure that identical images were never presented. Instead, the good quality target fingerprint was paired with the corresponding poor quality target fingerprint. This selection of stimuli ensured that (i) the task was not too trivial, (ii) the task more closely approximated that in the real world, and (iii) the use of simple image matching strategies on the part of the participant was minimised. For 'different' trials, care was taken to ensure that the foil fingerprint was always of the same pattern type (whorl, radial loop, ulnar loop) as the target fingerprint, again ensuring that the task was not too trivial. (See Fig. 1 for example stimuli for both 'same' trials and 'different' trials.)

Trials were presented, and data were recorded, via Superlab 4.5 running on a DELL laptop PC with a 17″ colour monitor and a screen resolution of 1024 × 768 pixels. Within this environment, each fingerprint measured approximately 8 cm high × 5.3 cm wide.

5.4. Procedure

Participants were briefed on the nature of the task, and provided informed consent prior to participation. Throughout the computer-based task, instructions were presented on-screen. These introduced participants to the task, and a set of 8 practice trials (half depicting 'same' trials) enabled them to locate the response buttons, and view the quality of the fingerprints they were to judge. In all trials, the two fingerprint images were displayed simultaneously and side by side on the screen, with the prompt question 'Same or Different?' below them. The images remained on screen until participant response, with participants pressing 's' if they considered the pair to come from the SAME individual, and 'd' if they considered the pair to come from DIFFERENT individuals. There was no option to provide an 'inconclusive' decision. Feedback was provided on these practice trials.

After an opportunity to clarify any queries, the main trials were presented. These consisted of 72 trials in total (half depicting 'same' trials). The format of each trial was identical to the practice phase except that feedback was not provided. Throughout the task, participants were asked to prioritise accuracy over speed and were reminded of the consequence of inaccurate decisions in the real world. The experimenter remained in the room throughout testing after which participants were fully debriefed and thanked for their time.

6. Results and discussion

Accuracy and speed of correct decisions for the expert analysts are shown in Table 1. From this it was clear than the experts performed exceptionally well, as would be hoped. Indeed, there were only 2 errors out of 864 decisions across the group, and these errors were false negatives in which a matching pair of fingerprints was classified as coming from different individuals. Of interest here was whether performance varied across the fingerprint pattern types, and whether performance was associated with the experience of the fingerprint expert. In order to examine these questions, accuracy of performance was combined across same and different trials to yield measures of sensitivity of discrimination (d') and response bias (C).

A one-way Analysis of Variance (ANOVA) was used to explore the first question, using sensitivity of discrimination (d') as the dependent variable, and pattern type (whorl, radial loop, ulnar loop) as the independent variable. This revealed no significant effect of pattern type ($F_{(2,22)} = 2.20$, $p > .05$) confirming that performance was equivalent regardless of fingerprint pattern. Similarly, a one-way ANOVA on the response bias score (C) confirmed no significant effect of pattern type on the level of bias in responding ($F_{(2,22)} = 2.20$, $p > .05$). Indeed a one-sample t-test (comparing the overall level of bias to zero) confirmed that there was neither a conservative nor a liberal bias in expert response ($t_{(11)} = 1.48$, $p > .05$). Instead, performance was accurate and high across the group.

In terms of speed of correct response, analyses were conducted on median response time (RT), which has the advantage of minimising the effect of skew often present in RT data. A two-way repeated-measures ANOVA was used to explore the effects of

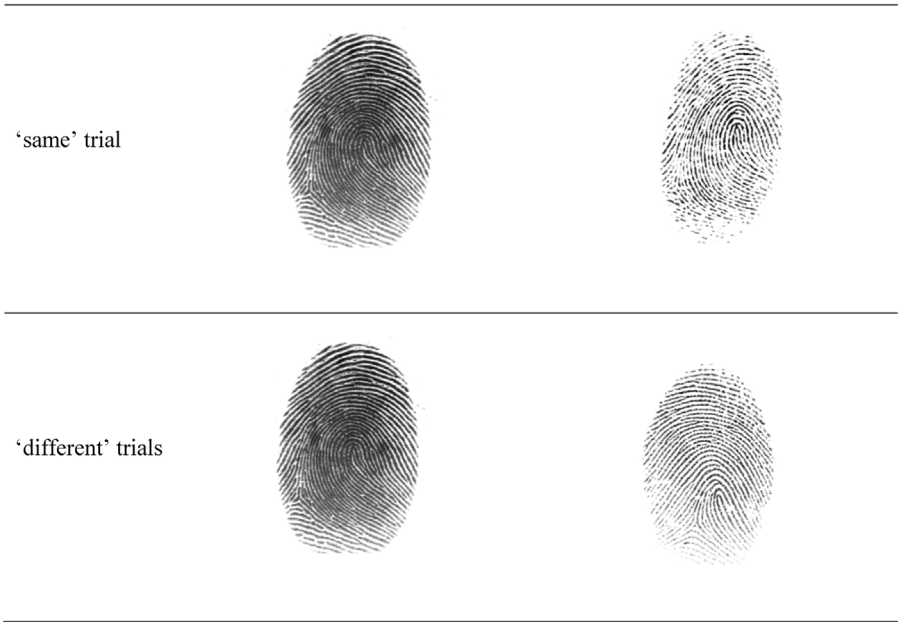


Fig. 1. Example stimuli used in a 'same' trial and a 'different' trial in Experiments 2 and 3.

Table 1
Accuracy and median speed of correct decisions (and standard deviation), together with measures of sensitivity of discrimination (d') and bias (C) for expert fingerprint analysts in Experiment 2.

	Whorls	Radial loops	Ulnar loops	Overall
Accuracy on 'matching' trials	1.00	.99	1.00	.995
	(.00)	(.03)	(.00)	(.01)
Accuracy on 'no match' trials	1.00	1.00	1.00	1.00
	(.00)	(.00)	(.00)	(.00)
Median RT on 'matching' trials (s)	14.85	15.56	17.15	15.85
	(8.61)	(10.34)	(10.46)	(9.42)
Median RT on 'no match' trials (s)	4.58	4.14	3.71	4.14
	(2.38)	(2.01)	(1.79)	(1.96)
Sensitivity of discrimination (d')	4.65	4.495	4.65	4.58
	(.00)	(.37)	(.00)	(.16)
Bias (C)	.00	.08	.00	.03
	(.00)	(.18)	(.00)	(.08)

fingerprint pattern (whorl, radial loop, ulnar loop) and trial type ('same', 'different') on speed of correct response. This revealed no main effect of fingerprint pattern ($F_{(2,22)} < 1$, *ns*), however, a main effect of trial type was apparent ($F_{(1,11)} = 23.88$ $p < .001$, partial $\eta^2 = .69$) with correct responses being significantly faster on 'different' trials than on 'same' trials. This was to be expected since the detection of a difference between the two fingerprints can terminate the trial faster than its absence. No significant interaction emerged to qualify this effect ($F_{(2,22)} = 2.42$, $p > .05$).

In order to address the second question, correlational analyses were conducted to determine whether performance was associated with experience as reflected by the number of years in the job. Pearson's correlations revealed no significant association between experience and d' ($r^2 = .54$, $p > .05$), between experience and response bias ($r^2 = .54$, $p > .05$), or between experience and speed of correct decisions on 'same' trials ($r^2 = .05$, $p > .05$) and 'different' trials ($r^2 = -.21$, $p > .05$). Taken together, these results revealed near-perfect performance with no bias. Experts appropriately took longer to report a 'match' than a 'no-match', but their degree of expertise as measured in years of service had no influence on any aspect of their performance meaning that those with 2 years of experience performed as well as those with 24 years of experience.

These results were very promising in terms of a validation of the ACE-V method. Not only did experts consistently describe its use (Experiment 1) but it appeared to be applied to provide highly reliable decisions regardless of the level of analyst experience (Experiment 2). That said, it is possible that the experts in Experiment 1 reported on aspects of their process that they thought they *ought* to report on, and either consciously or unconsciously withheld other details. Equally, it is possible that their high levels of performance in Experiment 2 could have rested on procedural elements that remain unspecified. Given this, a final test of the ACE-V method was to determine whether it was transparent enough to be conveyed to a set of naïve student participants and whether, in so doing, the performance of this trained group would approximate that of the expert analysts. This test of sufficiency was the purpose of Experiment 3.

7. Experiment 3: Method

7.1. Design

The fingerprint matching task used in Experiment 2 was repeated within Experiment 3 incorporating two additional groups of participants. As such, a $2 \times 3 \times 3$ mixed design was used in which the effects of trial type ('same', 'different') and fingerprint pattern (whorl, radial loop, ulnar loop) were varied within participants as before, and the effect of expertise (experts, trained, novices) was varied between participants. The expert data from Experiment 2 were incorporated into the present design but were analysed for a different purpose within Experiment 3. As before, participant accuracy and speed of response were recorded and represented the dependant variables.

7.2. Participants

The 12 experts used in Experiments 1 and 2 represented the expert group here. In addition, 54 student participants (30 females, 24 males) took part either on a volunteer basis or in return for course credit. Ages ranged from 18 to 30 years (mean = 21.85 years, SD = 2.8). Students were randomly allocated to either a 'trained' group ($n = 28$; 13 females, mean age = 21.92 years) or a 'novice' group ($n = 26$; 17 females, mean age = 21.76 years). None of the student participants had prior experience or expertise with fingerprint analysis, and all had normal, or corrected-to-normal, vision.

7.3. Materials

The materials for the fingerprint matching task were identical to those used in Experiment 2. However, a fingerprint training tool was developed for Experiment 3 as a way of instructing the 'trained' group. This took the form of a Powerpoint presentation and consisted of the experts' comments from Experiment 1 in which they described fingerprint characteristics, and their ACE-V method of analysis. In particular, the training tool provided an explicit focus on the stages of latent print analysis with clear description of the three levels of detail under consideration. In order to avoid potential misinterpretation, quotations from experts were used to provide all explanation throughout the Powerpoint presentation, and these were illustrated by the images that had been used in the experts' 'think aloud' task. The final fingerprint training tool was verified as an accurate and useful tool by a subset of the experts from Experiment 1.

7.4. Procedure

Prior to the fingerprint matching task, participants experienced training, or no training, as dictated by the participant group to which they had been assigned. Those in the training group were provided with the Fingerprint Training Tool in the form of the Powerpoint presentation, and they reviewed the presentation at their own pace. They were informed that the information would help them to complete the subsequent experimental task, and the experimenter remained in the room throughout this period in order to answer any questions and assist with understanding.

Following this, the 'trained' and 'novice' groups completed the fingerprint matching task in exactly the same way as the experts in Experiment 2. As before, participants were asked to prioritise accuracy over speed, and were reminded of the consequence of

inaccurate decisions in the real world. The experimental task took no more than 15 min after which participants were thanked and fully debriefed.

8. Results and discussion

Accuracy and speed of correct decisions are shown for all three participant groups in Table 2 and in Fig. 2. As was the case in Experiment 2, sensitivity of discrimination (d') and response bias (C) were calculated and were used in subsequent analyses. Of interest here was whether the provision of training would enable the trained group (i) to outperform their novice counterparts, and (ii) to perform on a comparable level relative to the experts.

Given the lack of any effect of fingerprint pattern in Experiment 2, and the lack of any theoretical predictions concerning fingerprint pattern, data were collapsed across fingerprint pattern here for simplicity of reporting.[2] Given the predominantly high performance of experts, degrees of freedom were adjusted where relevant when examining significance levels to take account of instances of unequal variances when comparing across groups.

A one-way ANOVA was used to explore the effect of participant group on sensitivity of discrimination (d'). This revealed a main effect of group as expected ($F_{(2,63)} = 92.45$, $p < .001$, partial $\eta^2 = .75$). Post-hoc comparisons confirmed the benefit of training in that the trained group outperformed the novice group ($t_{(52)} = 3.77$, $p < .001$). However, the trained group still performed significantly worse than the expert group ($t_{(33.53)} = 18.1$, $p < .001$) indicating a performance gap between the experts and those students trained in the ACE-V method.

A one-way ANOVA was again applied to explore the effect of participant group on response bias. This too revealed a main effect of group ($F_{(2,63)} = 3.71$, $p = .03$, partial $\eta^2 = .11$). In contrast to the previous analysis, post hoc comparisons revealed no difference in the level of bias between trained participants and novices ($t_{(52)} = 1.19$, $p > .025$) but also revealed no difference in the level of bias shown between trained and expert groups ($t_{(30.59)} = 2.32$, $p > .025$). This suggested that the trained group sat between the experts and novices without differing from either. A clearer picture was provided when the bias shown by each group was compared to zero via Bonferroni-corrected one-sample t-tests. These revealed no significant levels of bias amongst the experts ($t_{(11)} = 1.48$, $p > .016$), or amongst the trained group ($t_{(27)} = -2.01$, $p > .016$). However, a response bias was evident in the novice group ($t_{(25)} = -5.33$, $p < .001$) with a greater tendency for the novice group to say 'same' than 'different'. Overall, these results suggested that training was effective both in improving discrimination and in reducing bias amongst the trained group compared to the novice group. However, whilst the trained group approximated the experts in showing no response bias, their sensitivity of discrimination remained significantly below that of the experts indicating a performance gap between the two groups.

Finally, a 2 × 3 mixed ANOVA was used to explore the effects of trial type ('same', 'different') and participant group (expert, trained, novice) on median RT for correct decisions. This revealed a significant effect of trial type ($F_{(1,63)} = 63.38$, $p < .001$, partial $\eta^2 = .50$) with speed of correct response being significantly faster for 'different' trials than for 'same' trials. There was also a significant main effect of participant group ($F_{(2,63)} = 23.47$, $p < .001$, partial $\eta^2 = .43$) with experts taking longer over their decisions than the trained group, who in turn took longer than the novice group. These effects were qualified by the presence of a significant interaction between trial type and participant group ($F_{(2,63)} = 48.71$, $p < .001$, partial $\eta^2 = .61$). Post hoc comparisons revealed a main effect of participant group for speed of response on both 'same' trials ($F_{(2,63)} = 34.02$, $p < .001$, partial $\eta^2 = .52$) and 'different' trials ($F_{(2,63)} = 12.23$, $p < .001$, partial $\eta^2 = .28$). However, the pairwise comparisons showed these to be characterised differently. Importantly, the trained group took longer to respond than the novices for both trial types ('same': $t_{(52)} = 5.04$, $p < .001$; 'different': $t_{(52)} = 4.81$, $p < .001$) suggesting more care over decisions following training. However, the trained group responded at an equivalent speed to the experts only on 'different' trials ($t_{(38)} = -1.68$, *ns*), but took significantly less time than the experts when making 'same' decisions ($t_{(12.6)} = 3.43$, $p = .005$) suggesting that they did not wait to amass as much supporting evidence for their 'same' decisions as did the experts.

Table 2
Accuracy and median speed of correct decisions (and standard deviations) together with measures of sensitivity of discrimination (d') and bias (C) for experts, trained students and novice students in Experiment 3.

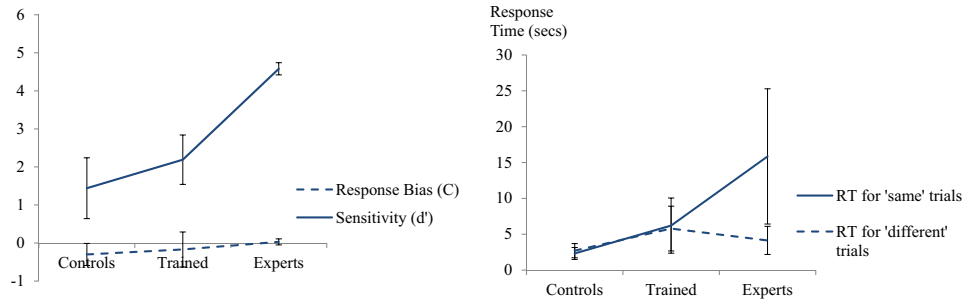
	Experts ($n = 12$)	Trained ($n = 28$)	Controls ($n = 26$)
Accuracy on 'matching' trials	.995 (.01)	.864 (.12)	.822 (.12)
Accuracy on 'no match' trials	1.00 (.00)	.795 (.15)	.644 (.16)
Median RT on 'matching' trials (s)	15.85 (9.43)	6.21 (3.85)	2.32 (.83)
Median RT on 'no match' trials (s)	4.14 (1.96)	5.79 (3.12)	2.71 (.99)
Sensitivity of discrimination (d')	4.58 (.16)	2.19 (.65)	1.44 (.80)
Bias (C)	.03 (.08)	−.17 (.46)	−.30 (.29)

Taken together, these results provided a mixed picture. On the one hand training was effective in supporting better performance in the trained group compared to the novice group. Sensitivity of discrimination was higher, bias in responding was removed, and trained participants appropriately took longer over their decisions. On the other hand, the performance of the trained group did not approximate that of the experts. In fact, there was still a significant performance gap in terms of sensitivity of discrimination, and in terms of taking sufficient time to reach a robust decision on 'matching' trials. These differences were perhaps to be expected given that expertise takes time to develop through practice, feedback, refinement, and repetition. Nevertheless, they suggest that mere knowledge of the ACE-V method is not sufficient in demonstrating expertise in a fingerprint matching task.

9. General discussion

The purpose of the current work was to provide some transparency to the process used by fingerprint experts, and to provide some reassurance as to the reliability of their decisions. In this regard, a mixed-methods approach has been valuable here, with Experiment 1 using a qualitative approach to better describe the method used, and Experiments 2 and 3 using a quantitative approach to test the reliability and the sufficiency of the approach in yielding accurate decisions. In this regard, the results of Experiment 1 provided reassurance in that the 12 experts spontaneously and confidently described a consistent method of analysis in line with the standards of good practice outlined in the ACE-V procedure. Moreover, they showed commonality in terms of

[2] When analyses were repeated with fingerprint pattern (whorl, radial loop, ulnar loop) as an additional factor, a main effect of fingerprint pattern emerged only when considering sensitivity of discrimination ($F_{(1.78,112.4)} = 11.52$, $p < .001$, partial $\eta^2 = .155$) with performance being better when judging whorls than ulnar loops ($t_{(65)} = 3.25$, $p = .002$), and when judging ulnar loops than radial loops ($t_{(65)} = 2.91$, $p = .005$) (see Fig. 1 in [50]). This effect of fingerprint pattern was unanticipated and is difficult to account for. However, the effect of fingerprint pattern did not qualify any other effects with any other dependent variables and thus is not reflected on any further.



A: Sensitivity and response bias across participant groups in Experiment 3

B: Response Time (in seconds) for correct decisions on 'same' and 'different' trials across participant groups in Experiment 3

Fig. 2. Graphical summary of the mean performance shown by control participants, trained participants and experts (together with standard deviation) in the fingerprint matching task in Experiment 3. Panel A summarises the sensitivity of discrimination and response bias measures. Panel B summarises the speed of correct responses.

the fingerprint characteristics that they relied upon, and they showed a common appreciation of the advantages and disadvantages associated with these characteristics. In terms of providing transparency over the method of fingerprint analysis, this study, combined with that of Charlton et al. [37], goes some way to providing a description of both process and pressure from the perspective of the analysts themselves. The concerns noted by Haber and Haber [20] over the lack of precision when describing the ACE-V are, however, upheld specifically in terms of a lack of specification over the degree of correspondence required when determining that two fingerprints are a 'match'. This is a point that will be returned to later.

The work reported here importantly provides an empirical test of performance of the experts via a bespoke task conducted with the experts' awareness. The purpose was to examine the challenging premise that if the experts' process did not yield accurate decisions then it would be of limited value, and issues of transparency or commonality would become unimportant. In this respect, the results of Experiment 2 showed the performance of the experts to be near-perfect. Thus, taken together, the data suggested consistency of approach (Experiment 1) and reliability of outcome (Experiment 2). It is sensible, however, to express a degree of caution with this simple conclusion: the experts may of course have *reported* a consistent approach whilst actually using a variety of undisclosed methods. Moreover, their near-perfect performance in Experiment 2 may well have rested on the analytic approach as described, but it could also have rested on additional methods either with or without their awareness. These concerns are reflected to a degree in the results of Experiment 3, to which we now turn.

9.1. Addressing the performance gap between experts and trained students

The results of Experiment 3 are important in several regards. First, they reveal the effectiveness (or otherwise) when training a group of naïve students on the method of fingerprint analysis described by the experts. This provides an indication of the transparency of the process, and the ease with which it can be understood by lay people. Second, by comparing the performance of this trained group with both an untrained control group, and with the expert group, it is possible to determine the degree to which the described method is sufficient to produce reliable decisions. Put another way, it is possible to determine whether the level of performance shown by the experts may depend on skills or knowledge in addition to that reflected by their ACE-V description. In this light, Experiment 3 provided a very interesting picture. On the one hand, the trained group out-performed the untrained novices producing better discrimination, less bias, and taking appropriately longer before reaching their decisions. However, on the other hand, the trained group remained worse compared to the experts in terms of discrimination, and length of time taken to reach a conclusion on 'matching' trials. This performance gap is important, and four potential explanations are explored to account for it.

First, it is possible that despite their training, the trained group simply could not do the fingerprint matching task to the degree shown by the experts. As an explanation this perhaps is merely a restatement of the problem rather than an account of the cause. Nevertheless, it reminds us that the experts may simply have a greater facility when matching fingerprints through practice, familiarity, and considerably more experience with the task. In this regard, the absence of any effect of years of experience amongst the expert group suggests that there may be a threshold level of practice required to demonstrate competency which the trained students had not achieved, but which all the experts exceeded. This said, it would be far more interesting to reflect on what this expertise may enable rather than merely to offer a threshold explanation to account for the current data.

Second, whilst all participants were reminded of the consequences of poor performance in the real world, it is possible that the trained group performed poorly relative to the experts simply because they were not invested in the outcomes. In contrast, the experts may have been highly motivated to perform well both through their appreciation of real world consequences, and through a potential desire to showcase their expertise. This issue of motivation is one that is very difficult for any laboratory-based study to address, and indeed it may be applied to any overt testing of experts, and to any studies involving student participants. In the current study, it may well account for the apparent performance gap between experts and trained students, although the fact that the trained group outperformed the novice student group goes some way to weakening its adequacy as an explanation for the current pattern of data.

The third possibility is that the trained group were not just unable to do the task, but they were unaware of their inability. This is a metacognitive explanation rather than a cognitive one, and rests on the insight that people hold about their performance. In this regard, it is possible that the trained group did not realise that their performance was sub-optimal. Consequently, they terminated the (matching) trials earlier than the experts, having amassed less evidence on which to base their decisions. In a very real sense, this metacognitive awareness of how well one performs may be affected by having (or not having) a standard by which to judge the degree of correspondence required when determining two prints as a match. Experts will have drawn on their experience to provide this implicit standard whereas the trained group received no advice on this point. This is an intriguing explanation and points to

the importance of feedback when evaluating one's level of performance. The trained group will have lacked any significant level of feedback (other than that received through the practice trials), whilst the experts will have gained feedback through the peer verification of their daily task. This feedback will have enabled the experts to check and fine-tune their performance, giving them a much greater metacognitive awareness of their abilities.

The fourth explanation for the performance gap between trained students and experts suggests that the training provided here may have been sub-optimal in ways that go beyond mere advice and feedback regarding degree of correspondence. This possibility recognises the fact that the experts may not have described their process in its entirety, either because of a professional pressure to report on the standard process, or because there are elements of the process which may be hard to articulate or may be unconsciously driven. In either situation, and despite the training tool having been verified for its completeness and accuracy by a subset of the experts, the training may have been insufficient to enable expert level performance. This account of the performance gap is very interesting indeed as it suggests that there may be more to the experts' task than they have been able to describe. In this respect, innovative work conducted by Busey and colleagues [16,39,40] has revealed aspects of the experts' process that may be difficult to verbalise. Specifically, Busey and Vanderkolk [39] have used EEG recordings to demonstrate a delayed N170 component when fingerprint images are inverted compared to when upright. This delayed component mirrors that when viewing upright and inverted faces and has been taken as an indicator of configural rather than featural processing [41]. Importantly, this delayed N170 pattern was only shown by fingerprint experts and was absent in fingerprint novices. Similarly, Busey et al. [40] have used eye-tracking to determine where the expert looked during the fingerprint matching task. When exposure time was limited, the results suggested that experts showed more accuracy on the matching task overall compared to novices. However, they also spent a greater proportion of their time, compared to the novices, looking at the latent print over the comparison fingerprint, and showed much more consistency in the regions of interest that they focussed on. Of course, merely knowing what someone is looking at cannot reveal what they are doing, and Busey et al. [40] are keen to point out that the expert may gain their advantage through interpretation rather than mere acquisition of information. However, in documenting these markers of expertise, Busey and Vanderkolk [39] emphasised the years of training, and immersion in fingerprint analysis, that are required in order to show such qualitative differences in performance. Consequently, the performance gap between the experts and trained participants noted here may plausibly reflect the gap between knowledge, versus expert refinement of that knowledge.

The current data also suggested another difference between the performance of experts, trained participants and novices here, which centres on the performance in 'different' trials. Indeed, examination of the data reveals that the trained participants (and novices) were fairly good in correctly saying 'same' on 'same' trials, but were substantially worse than the experts at correctly saying 'different' on 'different' trials. This pattern of performance echoes that seen by experts and novices in an allied forensic science discipline – handwriting analysis. Indeed, the work of Kam et al. [42] revealed better performance by 7 FBI document examiners than by 10 control participants. However, their later work using a substantial population of 100 experts and 41 control participants revealed that the control participants were let down not through their ability to say 'same' to matching samples but through their inability to say 'different' to mismatching samples [43]. In fact, despite financial incentive to perform well, and financial penalties for each error made, the control participants made incorrect 'match' decisions on 38.3% of occasions compared to an error rate of 6.5% as shown by the experts. One might conclude that the novices did not know what differences to look for, and thus all samples looked highly similar causing good performance when the samples were indeed from the same person, but poor performance when the samples were from different people. Additionally, the novices may have overlooked perceived differences as unimportant when the samples were from different people. This observation reinforces Busey et al. [40] point regarding a difference in interpretation, as well as a difference in perception, when accounting for the performance gap between experts and novices

Taken together, these results suggest that the expert may demonstrate their expertise in ways that are not captured by the ACE-V method alone. Specifically, the experts may adopt a more superior configural processing approach, and may adopt a longer and more focussed looking pattern compared to novices meaning that they are able to acquire and make better use of the information that they are provided with. In this regard, the results offer reassurance in suggesting that fingerprint experts do possess skills beyond that of the novice. However, they also suggest that there would be value in even greater transparency and understanding regarding those additional aspects of the experts' process.

9.2. Future work

In terms of future work to define the expert's process more fully, one approach is to borrow from the methodological playbook used by face researchers, where a relatively mature literature exists in exploring expertise effects. Of particular interest may be the examination of categorical perception effects amongst experts. Categorical perception describes the capacity to differentiate highly similar stimuli such that an abrupt perceptual shift is reported between instances of one stimulus and instances of another [44]. The classic example is the perception of light in which we see discrete colour bands of red, orange, yellow etc., despite the fact that colour varies continuously along the dimension of wavelength. Categorical perception has also been demonstrated when distinguishing between the faces of identical twins [45] such that that once each twin has been learned, pictures of the twins can readily be distinguished into discrete categories. Categorical perception is an interesting phenomenon because it is underpinned by two effects – compression and separation. In the context of the twin study, compression related to a shift in perception such that images of the same twin came to be seen as more similar after learning than before. In contrast, separation related to a shift in perception such that images of the two different twins came to be seen as more different after learning than before. Put another way, compression relates to the greater perception of 'sameness' within a category whilst separation relates to the greater perception of 'difference' between categories.

Against this background, it may be expected that expert fingerprint analysts may show a classic categorical perception effect reflecting an abrupt perceptual shift when discriminating between instances of highly similar or confusable fingerprints. In contrast, novices (and to a lesser extent, trained participants) may show far less ability to differentiate the fingerprints. Of more interest, though, is whether the expertise gap is driven by a difference in compression, separation, or both. The current data, combined with the findings from handwriting analysis reviewed above, may predict that novices and trained participants may adequately perceive similarities within a category (similarities between matching fingerprints). However, they may err in terms of their ability to see differences between categories (differences between non-matching fingerprints). Thus, compression but not

separation may be expected. Such a prediction is consistent with the observation here of a greater number of false positive errors than false negative errors amongst the novice and trained groups compared to the experts.

Examination of categorical perception effects provides one tool, which may be of value in examining expertise effects in fingerprint analysis. However, if categorical perception is revealed, it would then be valuable to determine the basis of such effects. Two explanations have been discussed within the categorical perception literature, and consideration of both may be valuable. The first explanation relates to the benefit that comes from having the language to differentiate between stimuli or between elements of the stimuli [46]. Put more clearly, experts may perform better in their task because their expertise affords them the language to describe, remember, and discriminate between examples. [40] noted the benefit that language can provide in that complex visual patterns can be described, differentiated, ascribed psychological meaning, and communicated to others, through the capacity to label them. Novices, on the other hand, simply lack this language. The second explanation relates to the benefit that may come through improved (expert) perception of the stimuli themselves perhaps by noticing smaller changes along previously used perceptual dimensions (increased sensitivity of perception) or perhaps by noticing differences along more dimensions (increased richness of perception) [47,48].

The application of multidimensional scaling to similarity judgements provides one way to explore this issue. Multidimensional scaling can enable the extraction of a similarity space within which individual fingerprints are located according to their physical properties along the dimensions that describe the space. Similar fingerprints will thus lie close together whilst different fingerprints will lie far apart in the space. Importantly, the analysis will also allow a determination of the perceptual dimensions that an individual may use when viewing complex stimuli such as fingerprints. The increased sensitivity explanation would suppose that experts use the same number of dimensions as novices but use them better, whilst the increased richness explanation would suppose that the experts use more dimensions when analysing fingerprints compared to novices. Consequently, an analysis of the dimensions used by novices, trained participants and experts may represent a valuable step in future work and may provide greater transparency of the expert process, with a clear benefit for training tools.

Finally, whilst the suggestions thus far provide empirical ways to explore the expert *perception* of fingerprints, consideration must also be given to the area of expert decision making, particularly in terms of the standard of sufficiency when determining a match. Currently, this standard is left up to the expertise of the analyst, and this standard may be refined over time with feedback from peers. The trained participants in the current study received no advice, and no feedback, in this regard. Of course, a move towards a quantitative expression of similarity between fingerprints (perhaps through a likelihood ratio or an expression of match probability) would remove the need for an internal standard to support a thresholded decision. However, until such a time, greater transparency is required regarding these internal standards, affected as they are by both the unequal consequences of false positives over false negatives, and the desire to avoid mistakes at all, as discussed by Charlton et al. [37].

9.3. A word of caution

Before over-interpreting the current results as providing any indication of error rates or proficiency in the real world, it is valuable to note three caveats with the current work. First, expert performance was assessed with the experts' awareness, and this may have encouraged a different style of responding to that when unaware of testing or when conducting routine casework. Second, performance was assessed using stimuli which may be of higher quality compared to the experts' normal caseload, creating an overly-simplistic task. This methodological factor was necessary in order to avoid floor effects amongst student participants, but may well have inflated the measures of expert competency above those expected in the real world. Finally, performance was assessed using a test in which there were an equal number of 'same' and 'different' trials. Whilst this experimental approach brings robustness through having a good number of trials of each type, and through the avoidance of prevalence effects, an equivalent number of 'same' and 'different' trials may be quite unrealistic (see [28]). For all these reasons, the absolute levels of performance shown by the experts should not be taken as indicative of their levels of performance in their day to day task. Instead, the focus here is usefully placed on the relative performance of experts, trained students, and the untrained control group on a standard test.

9.4. Summary and conclusions

The current work has presented a qualitative exploration of the ACE-V method as described from the perspective of active trained fingerprint analysts. It has also provided a quantitative test of reliability of fingerprint decisions based on this method, and the sufficiency of the method as a training tool for novices. The results offer reassurance in suggesting a common appreciation and adherence to the ACE-V method across the analysts involved in this study, and in suggesting reliable decisions as an outcome. Nevertheless, the performance gap between experts and a trained group of students suggests that the experts possess skills over and above those captured by the ACE-V method, notably in terms of configural processing, and an appreciation of the degree of correspondence required before a 'match' should be concluded.

These findings offer specific guidance for fingerprint training packages, and suggest that even greater transparency of the expert process is required, including some quantification of the degree of correspondence when determining a fingerprint match. However, such a recommendation sits alongside the current debate over how decisions should be articulated. If the issue of degree of correspondence is to be quantified (rather than remaining as a categorical match/no-match judgement), it would demand some baseline statistics on the number of individuals whose fingerprints would be expected to share any particular number of features. This demands that we have proper statistics relating to the frequency of fingerprint characteristics in the population, or the frequency of matches by chance, both of which do not currently exist. These metrics are desirable if we are to move to a more quantified standard to determine a match, and if we are to move towards a more probabilistic way of reporting decisions (see [1,3,36]). However, they require that we address a research challenge that is quite considerable. Given that the current lack of transparency regarding standards of correspondence represented one reason for a performance gap between experts and trained participants here, the current data add weight to the call for this challenge to be addressed. The promise, well-articulated by Koehler [49], is that the testing of the reliability of fingerprint decisions will not damage its reputation in court. Instead it will provide the necessary evidence to support its robustness.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant (EP/J004995/1 SID: An Exploration of SuperIdentity) awarded to the first author.

Colleagues on this grant are thanked for helpful contributions to the current work. In addition, the authors are grateful to the two anonymous referees for their advice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.forsciint.2016.08.026.

References

[1] C. Neumann, J. Champkin, Fingerprints at the crime-scene: statistically certain or probably? in: Significance: The Royal Statistical Society, February 2012, 21–25.
[2] S. Pankanti, S. Prabhakar, A.K. Jain, On the individuality of fingerprints, IEEE Trans. Pattern Anal. Mach. Intell 24 (8) (2002) 1010–1025.
[3] J.L. Mnookin, The validity of latent fingerprint identification: confessions of a fingerprinting moderate, Law Probab. Risk 7 (2008) 127–141.
[4] Federal Bureau of Investigation, The Science of Fingerprints: Classification and Uses, US Government Printing Office, Washington, DC, 1985.
[5] D.R. Ashbaugh, The premise of friction ridge identification, clarity, and the identification process, J. Forensic Identif. 44 (1994) 499–516.
[6] S.A. Cole, More than zero: accounting for error in latent fingerprint identification, J. Crim. Law Criminol. 95 (2005) 985–1078.
[7] W.F. Leo, Fingerprint identification: objective science or subjective opinion? The Print 17 (2001) 1–3.
[8] I.E. Dror, S.A. Cole, The vision in "blind" justice: expert perception, judgment and visual cognition in forensic pattern recognition, Psychon. Bull. Rev. 17 (2010) 161–167.
[9] S.M. Kassin, I.E. Dror, J. Kukucka, The forensic confirmation bias: problems, perspectives and proposed solutions, J. Appl. Res. Mem. Cogn. 2 (2013) 42–52.
[10] OIG, A Review of the FBI's Handling of the Brandon Mayfield Case, Office of the Inspector General, Oversight and Review Division, US Department of Justice, 2006, pp. 1–330.
[11] A. Campbell, The Fingerprint Inquiry Report, 2011 Available at http://www.thefingerprintinquiryscotland.org.uk/inquiry/3127-2.html.
[12] National Academy of Sciences, Strengthening Forensic Science in the United States: A Path Forward, National Academies Press, Washington, DC, 2009.
[13] NIST, Expert Working Group on human factors in latent print analysis. Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach, US Department of Commerce, National Institute of Standards and Technology, 2012.
[14] Forensic Science Regulator, Cognitive bias effects relevant to forensic science examinations, FSR-G-217, Issue 1, 2015, 1–96.
[15] S. Stammers, S. Bunn, Unintentional bias in forensic investigation. POSTbrief, Number 15, Houses of Parliament, Parliamentary Office of Science and Technology, 2015.
[16] T.A. Busey, G.R. Loftus, Cognitive science and the law, Trends Cogn. Sci. 11 (3) (2007) 111–117.
[17] R.A. Huber, Expert witness: in defence of expert witnesses in general and of document examiners in particular, Crim. Law Quart. 2 (1959) 276–296.
[18] D. Ashbaugh, Quantitative–Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology, CRC Press, New York, 1999.
[19] C. Champod, C. Lennard, P. Margot, M. Stoilovic, Fingerprints and Other Ridge Impressions, CRC Press, Boca Raton, 2004.
[20] R.N. Haber, L. Haber, Scientific validation of fingerprint evidence under Daubert, Law Probab. Risk 7 (2008) 87–109.
[21] C. Champod, Presentation to the Fingerprint Enquiry Scotland, 25th November 2009, 2009 Retrieved from http://www.webarchive.org.uk/wayback/archive/20150428163143/http://www.thefingerprintinquiryscotland.org.uk/inquiry/files/ED_0005.pdf.
[22] G. Langenburg, A performance study of the ACE-V process: a pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ACE-V process, J. Forensic Identif. 59 (2) (2009) 219–257.
[23] I. Evett, R. William, A review of the sixteen points fingerprint standard in England and Wales, in: J. Almog, E.E. Springer (Eds.), Proceedings of the International Symposium on Fingerprint Detection and Identification, Ne'urim, Israel, (1995), pp. 287–304.
[24] K. Wertheim, G. Langenburg, A. Moenssens, A report of latent print examiner accuracy during comparison training exercises, J. Forensic Identif. 56 (1) (2006) 55–93.
[25] S. Gutowski, Error rates in fingerprint examination: the view in 2006, in: The Forensic Bulletin, Autumn 2006, 2006, 18–19.
[26] J.M. Tangen, M.B. Thompson, D.J. McCarthy, Identifying fingerprint expertise, Psychol. Sci. 22 (8) (2011) 995–997.
[27] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. U.S.A. 108 (19) (2011) 7733–7738.
[28] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS ONE 7 (3) (2012) e32800.
[29] G. Langenburg, C. Champod, P. Wertheim, Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons, J. Forensic Sci. 54 (2009) 571–582.
[30] M.J. Saks, Concerning L.J. Hall E Player. Will the introduction of an emotional context affect fingerprint analysis and decision-making? Forensic Sci. Int. 191 (2009) e19.
[31] I.E. Dror, D. Charlton, A. Péron, Contextual information renders experts vulnerable to making erroneous identifications, Forensic Sci. Int. 156 (2006) 174–178.
[32] I.E. Dror, D. Charlton, Why experts make errors, J. Forensic Identif. 56 (2006) 600–616.
[33] I.E. Dror, K. Wertheim, P. Fraser-Mackenzie, J. Walajtys, The impact of human-technology cooperation and distributed cognition in forensic science: biasing effects of AFIS contextual information on human experts, J. Forensic Sci. 57 (2) (2012) 343–352.
[34] I.E. Dror, R. Rosenthal, Meta-analytically quantifying the reliability and bias ability of forensic experts, J. Forensic Sci. 53 (4) (2008) 900–903.
[35] I.E. Dror, S.M. Kassin, J. Kukucka, New application of psychology to law: improving forensic evidence and expert witness contributions, J. Appl. Res. Mem. Cogn. 2 (2013) 78–81.
[36] C. Champod, Fingerprint examination: towards more transparency, Law Probab. Risk 7 (2011) 165–189.
[37] D. Charlton, P.A.F. Fraser-Mackenzie, I.E. Dror, Emotional experiences and motivating factors associated with fingerprint analysis, J. Forensic Sci. (2010).
[38] V. Braun, V. Clarke, Using thematic analysis in psychology, Qual. Res. Psychol. 3 (2006) 77–101.
[39] T.A. Busey, J.R. Vanderkolk, Behavioral and electrophysiological evidence for configural processing in fingerprint experts, Vision Res. 45 (2005) 431–448.
[40] T.A. Busey, C. Yu, D. Wyatte, J. Vanderkolk, F. Parada, R. Akavipat, Consistency and variability among latent print examiners as revealed by eye tracking methodologies, J. Forensic Identif. 61 (1) (2011) 60–92.
[41] B. Rossion, I. Gauthier, How does the brain process upright and inverted faces? Behav. Cogn. Neurosci. Rev. 1 (2002) 62–74.
[42] M. Kam, J. Wetstein, R. Conn, Proficiency of professional document examiners in writer identification, J. Forensic Sci. 39 (1) (1994) 5–14.
[43] M. Kam, G. Fielding, R. Conn, Writer identification by professional document examiners, J. Forensic Sci. 42 (5) (1997) 778–786, http://dx.doi.org/10.1520/JFS14207, J.ISSN 0022-1198.
[44] S. Harnad, Psychophysical and cognitive aspects of categorical perception: a critical overview, in: S. Harnad (Ed.), Categorical Perception: The Groundwork of Cognition, Cambridge University Press, New York, 1987.
[45] S.V. Stevenage, Which twin are you? A demonstration of induced categorical perception of identical twin faces, Br. J. Psychol. 89 (1998) 39–57.
[46] G. Lupyan, D.H. Rakison, J.L. McClelland, Language is not just for talking—redundant labels facilitate learning of novel categories, Psychol. Sci. 18 (12) (2007) 1077–1083.
[47] R.L. Goldstone, Influences of categorization of perceptual discrimination, J. Exp. Psychol. Gen. 123 (1994) 178–200.
[48] R.L. Goldstone, Perceptual learning, Annu. Rev. Psychol. 49 (1998) 585–612.
[49] J.J. Koehler, Fingerprint error rates and proficiency tests: what they are and why they matter, Hast. Law J. 59 (2008) 1077–1110.
[50] Data in Brief, Data from a fingerprint matching task with experts, trained students and untrained novices, Forensic Science International: Data in Brief (2016) (submitted).