Forensic Science International 208 (2011) 10–17

Contents lists available at ScienceDirect

Forensic Science International

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

Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison

Itiel E. Dror [a,b,*], Christophe Champod [c], Glenn Langenburg [c,d], David Charlton [e,f], Heloise Hunt [a], Robert Rosenthal [g]

[a] Institute of Cognitive Neuroscience, University College London, London, United Kingdom
[b] Cognitive Consultants International Ltd., United Kingdom
[c] Ecole des sciences criminelles, Institut de police scientifique, University of Lausanne, Lausanne, Switzerland
[d] Minnesota Bureau of Criminal Apprehension Forensic Science Services, St. Paul, MN, United States
[e] School of Applied Sciences, Bournemouth University, United Kingdom
[f] Fingerprint Bureau, Sussex Police, United Kingdom
[g] Department of Psychology, University of California Riverside, United States

ARTICLE INFO

Article history:
Received 24 June 2010
Received in revised form 4 October 2010
Accepted 9 October 2010
Available online 3 December 2010

Keywords:
Latent fingerprinting
Human cognition
Fingerprint analysis

ABSTRACT

Deciding whether two fingerprint marks originate from the same source requires examination and comparison of their features. Many cognitive factors play a major role in such information processing. In this paper we examined the consistency (both between- and within-experts) in the analysis of latent marks, and whether the presence of a 'target' comparison print affects this analysis. Our findings showed that the context of a comparison print affected analysis of the latent mark, possibly influencing allocation of attention, visual search, and threshold for determining a 'signal'. We also found that even without the context of the comparison print there was still a lack of consistency in analysing latent marks. Not only was this reflected by inconsistency between different experts, but the same experts at different times were inconsistent with their own analysis. However, the characterization of these inconsistencies depends on the standard and definition of what constitutes inconsistent. Furthermore, these effects were not uniform; the lack of consistency varied across fingerprints and experts. We propose solutions to mediate variability in the analysis of friction ridge skin.

Cognitive processes underpin much of the work carried out in many forensic disciplines which require examination of visual images. Fingerprints, bite and shoe marks, tire tracks, firearms, hair, handwriting and other forensic domains all hinge on comparative examination involving visual recognition. Although human experts are the 'instrument' in judging whether two patterns originate from the same source, understanding the factors that shape such judgements in forensic science has been relatively neglected. In the past it has been misconceived that "fingerprint identification is an exact science" ([1] p. 8); and this perception goes across all forensic disciplines [2]. The recent National Academy of Sciences report further highlights that "the findings of cognitive psychology... the extent to which practitioners in a particular forensic discipline rely on human interpretation... are significant" and that "...Unfortunately, at least to date, there is no good evidence to indicate that the forensic science community has made a sufficient effort to address the bias issue" ([3] p. 8–9).

The task demands imposed on the examiners require them to search through a rich stimulus, filter out noise, and determine characteristics and 'signals' for comparison (see [4,5] for discussion of signal detection theory (SDT) applied to fingerprint evidence). This initial analysis and determination of 'signals' can take place before the actual comparison between stimuli (e.g., the latent mark left at a crime scene and the comparison print of a known suspect). Scientists have long accepted that observations, including those in their own scientific research, encompass errors. A study examining 140,000 scientific observations reported in published research not only revealed that erroneous observations were made, but that they were systematically biased in favour of the hypothesis being researched [6]. Many different forms of contextual and cognitive influences affect our perception and bias it in a variety of ways [7]. Previous research on fingerprinting specifically examined potential cognitive contextual influences on comparing prints and decision making as to whether or not they originated from the same source [8–20].

* Corresponding author at: Institute of Cognitive Neuroscience, Department of Psychology, University College London, London, United Kingdom.
E-mail address: i.dror@ucl.ac.uk (I.E. Dror).

0379-0738/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2010.10.013

Bias in different aspects of forensic decision making has been examined in a number of studies (see review articles [21,22]). Specifically focusing on the initial analysis phase of the mark, before being actually compared to any prints, Langenburg [23] found that examiners generally reported more minutiae than novice controls. Furthermore, although the examiners varied in how many minutiae they observed in the initial analysis, they were more consistent than the novice control group (in 8 out of the 10 latent marks used in this study). These results were in agreement with those of Evett and Williams [1]. Following Langenburg study, Schiffer and Champod [24] found that training and experience increased the number of characteristics reported, and at the same time reduced the variability among observers. Schiffer and Champod also reported that the number of characteristics observed during the analysis phase was not affected by contextual information about the case or by the presence of a comparison print. Consequently, they concluded that the initial analysis stage (pre-comparison) is relatively robust and relatively free from the risk of contamination through contextualisation of the process.

Although Langenburg [23] and Schiffer and Champod [24] show that these inconsistencies decrease with training and experience,[1] they also make the point that "quite important variations do subsist between examiners" ([24] p. 119). All the studies consistently show that there is variability in the number of minutiae observed in the analysis stage, and that these inconsistencies are attenuated but not eliminated, during the initial training and experience in fingerprint examination. Furthermore, as reported by Schiffer and Champod [24], even in the relatively robust stage of analysis "a clear subjective element persists". A further study [25] suggests that the combined presence of contextual pressure and availability of the target comparison print influences the evaluation stage (following the analysis and comparison), but this effect varies among different marks.

Dror et al. [11] suggested that as finger marks are more difficult (bottom-up), the more influence external factors (top-down) have on the observations. Bottom-up refers to the incoming data, where as top-down relies on pre-existing knowledge [26]. Top-down has many forms and manifestations, which include the context in which the data are presented, past experiences and knowledge, expectations, and so forth. Expertise is top-down, and as such experts rely more on top-down information. This allows efficient and effective processing of the bottom-up data, but also means it can distort and bias how the data are processed [27]. Variations in observation among different observers ("inter-observer" differences) and variations in observation for the same observer for the same task, taken at different times ("intra-observer" differences) are a well-known phenomenon in other fields involving expert decisions, such as radiologists or other medical technicians [28,29].

In the research reported here we examined three main issues:

1. The potential effect that a 'target' comparison fingerprint may have on the analysis of the latent mark.
2. The consistency in analysis among different examiners.
3. The consistency in analysis within the same examiner.

This paper further investigates and contributes to the studies on the analysis of fingerprints in the following ways:

1. Using actual latent fingerprint examiners, rather than forensic science or psychology students (such as in [11,25]).
2. Applying a within-subject (intra-observer) experimental design. This allows us to measure consistency in analysis, as we compare examiners to themselves. Such intra-observer measurements are extremely accurate and informative because they are not only statistically more powerful then inter-observer measures, but they allow us to confidently draw conclusions because the data cannot be attributed to individual differences, such as visual acuity, experience, strategy, cognitive style, and training.
3. Subjecting the experimental data to statistical procedures and standards (e.g., retest reliability) that quantify the consistency of latent fingerprint examiners in the analysis of latent marks.
4. Statistically differentiating between factors that contribute to inconsistencies in latent mark analysis; thus determining what portion of the variance is attributed to the examiners' performance and what portion is attributed to the latent marks themselves (using statistical effect sizes).
5. Suggesting a number of recommendations for dealing with issues surrounding latent mark analysis.

1. Effects of a 'target' comparison

The human cognitive system is limited in its capacity to process information. The information available far exceeds available brain power and cognitive resources, and therefore we can only process a fraction of the information presented to us. This mismatch between computational demands and available cognitive resources caused the development of cognitive mechanisms that underpin intelligence. For example, we prioritize what information to process according to our expectations (e.g., [30]). Expectations are derived from experience, motivation, context, and other top-down cognitive processes that guide visual search, allocation of attention, filtering of information, and what (and how) information is processed. These mechanisms are vital for cognitive processes to be successful. Expertise is characterised by further development and enhancement of such mechanisms [26,27,31,32].

Therefore, there are good scientific data showing that the presence of any contextual information may affect cognitive information processing. Various factors and specific parameters define the context, whom it may affect, how, and to what extent. Understanding these factors and parameters will help develop science-based training and best practices that will enhance objectivity in fingerprint analyses, as well as in other forensic comparative examinations involving visual recognition.

In the first experiment reported in this paper we used 20 experienced latent fingerprint examiners, to investigate whether the presence of a comparison 'target' print would affect the characteristics they observe in the latent mark. Each of the 20 experts received ten stimuli: five latent marks by themselves (solo condition) and five latent marks with the matching target print (pair condition). All the participants were instructed identically, requiring them to examine the latent marks and to count *all* the minutiae present in the image. The experimental conditions were counterbalanced across participants using a Latin Square design to minimize any affects due to the order of presenting the experimental trials [33].

We found that the presence of the accompanying comparison print affected how many minutiae were perceived by the expert latent print examiners. These differences were statistically significant ($t_{(9)} = 2.38$, $p = .021$; with an effect size, $r = .62$). Interestingly, as evident in Table 1, the presence of the accompanying matching comparison print mainly *reduced* the number of minutiae perceived. This is consistent with attention guided visual search, whereby our cognitive system operates within the contextual expectation. It is important to note that the reduced number of minutiae was perhaps due to the comparison print being from the same source (a match); if it had been a non-match,

[1] This is particularly noticeable at the earlier stage of professional development, when a trainee has some experience and training. It is by no means a continuous linear change; there is a strong initial effect, but then it levels off and may even decline.

Table 1
The mean number of minutiae observed when the latent mark was presented by itself ('solo'), within the context of a comparison print ('pair'), and the differences between these two conditions.

Latent mark	Solo	Pair	Difference
A	20.6	14.1	−6.5
B	13.4	9.9	−3.5
C	20.1	10.8	−9.3
D	9.8	9.7	−0.1
E	10.7	11.1	0.4
F	8.4	8.8	0.4
G	12.1	10.7	−1.4
H	15.6	10.5	−5.1
I	7.1	8.5	1.4
J	9.1	6.6	−2.5
MEAN	12.7	10.1	−2.6
SD	4.7	2.0	3.5

then it may have directed the perceptual cognitive system differently, possibly observing more minutiae. The importance of the finding is not whether the presence of the comparison print reduced or increased the number of minutiae perceived in the latent mark, but that the presence of a target comparison print had an effect on the perception and judgment of the latent mark.

This finding emphasises the importance of examining the latent mark in isolation, prior to being exposed to any potential comparison print. This is to maximize the 'clean' bottom-up and more objective analysis, driven by the actual latent mark, and to minimize external influences that may bias the process of analysing the latent mark itself. This is especially important when the latent mark is of low quality. Such recommendations are also appropriate for other forensic domains (e.g., DNA, see sequential unmasking [34]), as well as for scientific investigations in general: "Keep the processes of data collection and analysis as blind as possible for as long as possible" (Rosenthal [6] p. 1007).

However, Dror points out that the comparison print can play an important role in helping examiners optimize their analysis by correctly guiding their cognitive resources and interpretation [35]. Therefore it seems reasonable to balance the vulnerabilities and cues presented by making the comparison print available to the examiner. A solution may be to first examine the latent mark in isolation, clearly documenting this more objective and uninfluenced analysis, but at the same time also allowing further analysis to be conducted later after exposure to the context of the target comparison print. Hence, the ACE approach needs to be initially applied linearly, making sure that the initial Analysis of the latent mark is done in isolation and documented, prior to moving to Comparison and Evaluation; yet still allowing flexibility, with transparency of when and why this took place, as well as procedures that control and limit the circumstances and extent for such retroactive changes so as to maximize performance but avoid (or at least minimize) circular reasoning and bias (for details, see [35]).

It is interesting and important to note that some latent marks were more susceptible to this effect than others. For example, Table 1 shows that latent mark D was basically unaffected by the presence of the comparison print, whereas latent mark B was quite dramatically affected (see Fig. 1, below, for the actual latent marks). It is clear from all the studies on latent mark analysis that findings are highly dependent on the specific fingerprints used. This suggests that we can (and probably should) tailor procedures and best practices to specific types of prints, rather than inflexibly applying identical procedures prescribed to all prints [35]. Such knowledge-based procedures will allow for higher quality work without requiring more resources, because it wisely and appropriately allocates resources to where they are needed.

The large variability in the effects of the presence of the comparison print on the latent mark analysis may explain why Schiffer and Champod [24] did not find such an effect: The latent marks they used may have been prints that are less (or not at all) affected by the presence of the comparison print, such as latent mark D in this study. An alternative (not mutually exclusive) explanation of why Schiffer and Champod did not find this effect is that they used students, and these effects may occur as examiners are more experienced and knowledgeable, and thus have expertise in how to utilise the information from the comparison print more effectively. The study reported here used experienced experts in latent print examination. It is also possible that experienced examiners tend to be more risk prone at calling minutiae, as opposed to students who will be more conservative.

It is also interesting to note that the largest differences were observed with the latent marks that had the highest number of

Fig. 1. Some latent marks were more affected by the presence of a target comparison print than other latent marks. For example, latent mark B (left panel) was more affected then latent mark D (right panel).

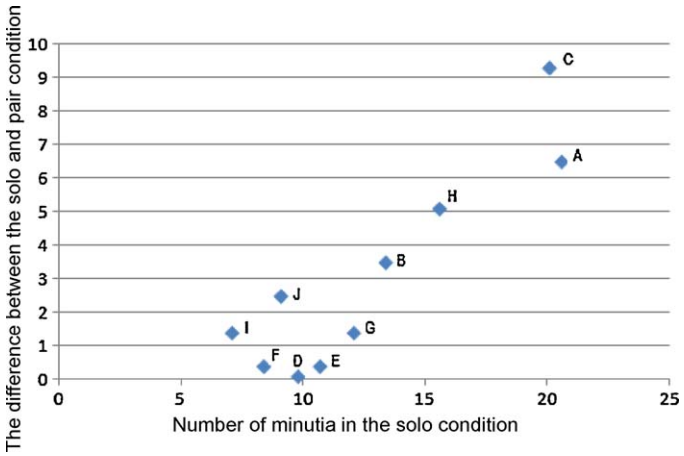

Fig. 2. The high correlation (0.9) between the mean number of minutiae observed when analysis was conducted when latent mark was presented by itself ('solo' condition) and the differences in analysis between the solo and pair conditions (absolute values, Table 1, right column).

minutiae observed in the solo condition (see, e.g., A, C, & H in Table 1). Overall, the correlation between the number of minutiae observed in the solo condition and the difference (absolute value) from those observed when shown in the pair condition was 0.9 (see Fig. 2). This may be due just to a ceiling effect, i.e., an artifact reflecting that as there are more minutiae marked in the solo condition, then there is more scope to reduce this number in the pair condition, and as the number of minutiae are lower in the solo condition, there is much less scope for a drop when they are presented in the pair condition.

An alternative, not mutually exclusive, explanation is the effect of motivational factors.[2] In the solo condition, examiners may be motivated to mark as many minutiae as they can, as they are not sure which ones may be useful and informative when they have a target print at the comparison stage. However, when the latent mark is analyzed while the target comparison print is available (as in the pair condition), examiners' motivation may drop when they get to a critical mass of minutiae they need for comparison purposes. Once they get to that threshold, they may be less likely to detect more minutiae. This effect further strengthens our suggestion that the initial analysis of latent marks should be done in isolation of a comparison exemplar print (especially when the latent mark is judged to be low quality, distorted, or has limited information available).

2. Inter-observer consistency

As we have shown, the presence of a 'target' comparison print can affect the perception and judgment of the latent mark in a number of ways. The next issue investigated was the consistency in the perception and judgment of minutiae in a latent mark across participants, even without the presence of a target comparison. The 'solo' condition data contain the answer to this question; it allows us to examine and compare the minutiae observed by different experts, and hence to report the variability in how latent print examiners may perceive and judge minutiae. Table 2 presents the relevant data, with the range of values for each latent mark (bottom row).

The apparent lack of consistency may reflect the absence of objective and quantifiable measures as to what constitutes a minutia, especially with latent marks that are of varying quality. However, these differences may also reflect individual differences between the examiners (arising from variations in eyesight, training, feature selection strategy, cognitive style, threshold criteria, etc.). It is important to understand the cognitive issues in latent mark analysis, and the variabilities in the analysis among and within examiners provides insights to the underling cognitive processing.

Evett and Williams [1], Langenburg [13,23], and Schiffer and Champod [24], all found inconsistencies among examiners regarding the number of minutiae observed. Evett and Williams suggest that this "confirms the subjective nature of points of comparison" (p. 7), and Langenburg [23] and Schiffer and Champod report that these variations are larger with novices.

As fingerprint examination advances, more objective measures and standards will ensure greater consistency among examiners. The potential influence introduced by a 'target' comparison print was addressed earlier. Another issue is the calibration of the threshold for determining whether a minutia is a 'signal.' Different examiners may be using different threshold criteria, and hence the large variance in how many minutiae different latent fingerprint examiners report on the same latent mark (similar problems occur in other forensic domains; see, for example, the lack of agreement on colour description used to determine the age of a bruise [35]).

A simple training tool could help deal with this problem.[3] A set of latent marks can be made available for examiners to analyse. After analysis, personal feedback will be provided to the examiner as to how consistent they are with other examiners. For example, it may state that 'your analysis resulted in similar minutiae as most examiners (and hence no need to calibrate thresholds), or it may state that 'your analysis resulted in a larger (or much larger, or smaller, as the case may be) number of minutiae relative to most examiners (and hence the examiners may consider changing their thresholds). The idea is that this would be a private measure, with results and feedback confidentially available only to the individual examiner. The full technical details of such a training calibration tool and its implementation are beyond the scope of this paper, but they are straightforward. Some more conceptual issues that need to be addressed are which latent marks should be used for this purpose, and how to make sure the feedback is taken on board and examiners do indeed calibrate their judgements. These must be scientifically based decisions. Furthermore, a fundamental issue that needs to be addressed is that the calibration is done to the 'correct' threshold, because ensuring different examiners use the same criterion, does not mean they are using the 'correct' one.

3. Intra-observer consistency

Judgment and subjectivity affect the number of minutiae characteristics reported, resulting in inconsistency among experts on how many minutiae are present within a specific latent mark. This study and the other studies [1,23,24] all consistently show that these variations are further dependent on the actual latent mark and exemplar prints in questions (i.e., some produce higher inconsistency than others). Furthermore, Dror and Charlton [8,9] report that some examiners are more affected by context than others. To ascertain the role of individual differences (such as experience, motivation, training, feature selection strategy, thresholds, cognitive style, personality) vs. the contribution of lack of objective quantifiable measures for determining characteristics in analysis of latent marks, we conducted an intra-observer (within-expert subject) experimental design.

[2] We mean a cognitive motivation, not an intentional motivation, or lack thereof, to conduct proper analysis. That is, how motivated and driven is the cognitive system to spend resources on processing and evaluating additional information. Generally, our cognitive system is efficient and economical in the sense that it does the minimal amount of processing needed to get the job done. This enables it to best utilize the brain's limited cognitive resources.

[3] This idea was first presented by Arie Zeelenberg, and referred to as Fingerprint Analyses Consistency Tester FACT (finder).

Table 2
The number of minutiae observed by each examiner for each latent mark (inter-observer). The minimum number per latent mark ('Min'), the maximum number per latent mark ('Max'), the standard deviation ('SD') and the range of minutiae observed for each latent mark (presented on the bottom row).

	Analysis of the latent marks									
	A	B	C	D	E	F	G	H	I	J
	22	9	15	8	9	3	8	11	7	10
	21	11	25	7	10	9	9	10	6	5
	19	9	18	10	7	9	15	19	6	6
	21	21	29	14	12	9	8	9	4	8
	17	16	15	11	16	9	7	12	5	5
	20	14	22	9	10	7	13	18	7	9
	22	17	15	10	10	8	11	24	8	11
	9	9	19	6	9	8	18	16	9	10
	30	15	25	10	12	12	19	22	12	17
	25	13	18	13	12	10	13	15	7	10
Min	9	9	15	6	7	3	7	9	4	5
Max	30	21	29	14	16	12	19	24	12	17
Mean	20.1	13.4	20.1	9.8	10.7	8.4	12.1	15.6	7.1	9.1
SD	5.49	4.01	4.93	2.49	2.45	2.32	4.25	5.15	2.23	3.54
Range	21	12	14	8	9	9	12	15	8	12

Within-expert experimental design examines intra-observer effects, comparing an examiner's responses at one time to their own responses at another time, thus controlling for individual differences (see Dror and Charlton [8,9]). The study reported here examined the consistency in analysis of latent marks within the same expert examiner. A new set of expert examiners was used. They were asked to report *all* the minutiae present on ten latent marks. A few months later, they were asked to do the same exercise, thus receiving the same identical instructions at time 1 and at time 2. The experts overall analyzed 200 latent marks, 100 latent marks twice. Table 3 presents the actual data: 10 latent print examiners, each making 20 analyses in total, analysing 10 latent marks (A–J), at $Time_1$ and at $Time_2$. In contrast to Table 2 where we examined the overall range and consistency obtained across examiners, here we focus on comparing the results of each examiner to his or herself, specifically looking at the degree to which the experts were consistent with themselves.

Analysis of variance (ANOVA) of the data from Table 3 showed that *examiners* differed significantly from each other in the number of minutiae reported: $F_{(9,81)} = 8.28$, $p = .<0.001$, effect size correlation *eta* = .69. This analysis also showed that the number of minutiae observed differed significantly from each other depending on the *latent mark*: $F_{(9,81)} = 57.30$, $p = <0.001$, effect size correlation *eta* = .93. Note the larger effect size for the contribution of the latent marks compared to the effect size for the contribution of the examiners. Most important is the Retest Reliability reported in Table 3 (right column) which is a statistical measure for quantifying consistency; see also the stem-and-leaf Plot and the Five Point Summaries of retest reliabilities in Fig. 3.

It was interesting to see whether the inconsistencies occurred over the typical range of thresholds for potential decision (e.g., 8 vs. 17, see examiner 3, latent mark G), or in ranges that do not typically matter for identification (examiner 6, latent mark A, 25 vs. 34). Both cases have a difference of 9 minutiae, but the former variability is more likely to cross a decision threshold for identification, while the latter's range of values are more likely to all be above a decision threshold (of course, this cannot be determined with certainly from the data in the present study, as this analysis is on the latent mark alone, prior to comparison to a print). Do examiners even consider identification thresholds when conducting the initial analysis? Evett and Williams [1] reported that the number of minutiae participants observed was influenced by decision thresholds, e.g., "participants tended to avoid returning 15 points" (p. 7).[4] Categorical perception makes people perceive information according to psychological categories rather than by their actual physical appearance [36].

To further investigate and understand the inconsistency we calculated the absolute differences in the analysis between $time_1$ and $time_2$, for each examiner (1–10) for each latent mark (A–J), see Table 4 (see also the stem-and-leaf Plot and the Five Point Summaries of retest reliabilities in Fig. 4). A score of '0' reflects a potentially perfectly consistent analysis.[5] As evident in Table 4, there were only 16% such consistent analyses (this is a conservative value, best case scenario; the actual variability may be higher, see footnote 5). If we 'relax' our criteria for consistency, and characterize consistency as a difference of 0 or 1, then there are 40% consistent analyses; if we further relax our criteria for consistency to include differences of 0, 1, or 2, then there are 55% consistent analyses (or, stated differently, 45% of the analyses differed in at least more than two minutiae between the two analyses conducted by the same examiner – footnote 5 notwithstanding). These data raise questions about objective assessment even at the analysis stage (which seems to be more robust to influences and context than the other stages of fingerprint examination and decision making). The data reported here are conservative, as the variability may be much higher.

However, Analyses of variance (ANOVA) of the data of Table 3 showed that although examiners differed significantly from each other in their degree of consistency in judging fingerprints (*eta* = .44), and in the number of minutiae observed (*eta* = .69), they still showed a high degree of inter-observer reliability with each other ($r_{intraclass} = .85$) and with themselves (retest reliability $r = .86$). The examiners who showed the highest retest reliabilities also tended to show the smallest discrepancy between their two evaluations of the same fingerprints at $time_1$ and at $time_2$, $r = -.84$.[6]

The differences between $time_1$ and $time_2$ (see Table 4) show that some examiners are more consistent than others (see, e.g., examiner 10, who is relatively highly consistent vs. examiner 3). Indeed, analysis of variance (ANOVA) of the difference scores in

[4] The Evett and Williams study was conducted when a 16-point standard was in place in the UK.

[5] It is important to note that even if the examiner reported exactly the same number of minutiae, it does not necessarily reflect consistency because although they may have observed the same number of minutiae in $Time_1$ and in $Time_2$, these may have been a different set of minutia. Hence, reporting the number of minutiae (rather than their overlap) provides a best case scenario.

[6] When we computed the same statistics but on the relative difference (i.e., the difference as a function of the total number of minutiae, we obtained essentially the same statistical results).

Table 3
The number of minutiae observed by each examiner (1–10), for each latent mark (A–J), at time 1 and at time 2 (intra-observer). The last column shows the retest reliability statistic for each of the 10 examiners.

		Latent mark										Retest reliability (r_{12})
Examiner		A	B	C	D	E	F	G	H	I	J	
1	Time 1	27	15	17	9	9	7	16	13	7	13	.95
	Time 2	26	14	21	10	8	5	13	15	7	12	
2	Time 1	31	16	14	9	10	7	12	13	6	9	.85
	Time 2	23	13	19	10	9	9	10	8	8	11	
3	Time 1	19	11	13	5	9	5	8	12	6	10	.65
	Time 2	18	8	16	8	15	9	17	21	7	12	
4	Time 1	20	12	17	6	10	8	7	8	6	7	.92
	Time 2	22	9	19	11	10	9	8	8	6	8	
5	Time 1	19	11	19	6	10	13	9	14	8	12	.84
	Time 2	25	13	21	9	14	12	12	11	8	9	
6	Time 1	34	16	21	12	13	13	12	11	8	12	.80
	Time 2	25	12	23	11	17	7	12	16	9	13	
7	Time 1	21	9	19	9	12	9	10	18	6	10	.80
	Time 2	21	13	14	7	8	6	7	11	6	10	
8	Time 1	19	14	14	10	9	6	12	13	7	11	.87
	Time 2	22	13	18	10	15	8	13	17	5	11	
9	Time 1	19	11	11	7	9	4	8	15	5	2	.88
	Time 2	23	14	20	7	13	8	11	14	4	5	
10	Time 1	19	10	9	8	4	2	10	8	6	5	.91
	Time 2	20	10	9	7	8	3	6	7	6	5	

Table 4 showed that examiners differed significantly from each other in the consistency with which they judged the 10 latent marks: $F_{(9,81)} = 2.17$, $p = .032$, effect size correlation *eta* = .44. Are the more consistent examiners characterized by personality type and cognitive aptitudes? If so, we need to know how to select candidates with such cognitive profiles during recruitment. Or perhaps these examiners receive a certain type of training, or maybe they adopted more objective definitions? All these are important questions that may help pave the way to understanding how such variations can be minimized.

However, the inconsistencies did not only vary between examiners, they were also dependent on the latent mark itself. The analysis of the variance also showed that latent marks differed significantly from each other in the consistency with which they were judged: $F_{(9,81)} = 2.82$, $p = .006$, effect size correlation *eta* = .49. This means that some latent marks are just more susceptible to issues of consistency than others. However, understanding and characterizing what constitutes such latent marks is not a simple matter, and we must be careful and not be hasty in determining how to *a priori* know which prints are susceptible to inconsistent analysis. With careful further research and converging studies, we should be able to learn and predict which latent marks are likely to be problematic.

Examiners		Latent Marks
1, 2, 5	.9	
0, 0, 4, 5, 7, 8	.8	7
	.7	
5	.6	8
	.5	4, 9
	.4	9
	.3	0, 5, 6
	.2	4
	.1	6
Mean	0.85	0.46
SD	0.085	0.22
Min	0.65	0.16
Median	0.86	0.43
Max	0.95	0.87

	Examiners	Latent Marks
Maximum	.95	.87
75th percentile	.912	.612
Median	.86	.425
25th percentile	.80	.285
Minimum	.65	.16
Mean	.85	.46
SD	.085	.218
S^2	.007	.048

Fig. 3. Stem-and-leaf plot of retest reliabilities of 10 fingerprint experts and 10 latent marks (top panel) and summaries statistics of retest reliabilities of 10 fingerprint experts and 10 latent marks.

Experts		Latent Marks
1	4.	
1, 2, 3	3.	4, 5, 6, 7
3, 7, 8	2.	4, 6, 9
2, 5, 6	1.	3, 7
	0.	7

	Experts	Latent Marks
Maximum	4.1	3.7
75th percentile	3.22	3.52
Median	2.75	2.75
25th percentile	1.58	1.6
Minimum	1.2	0.7
Mean	2.58	2.58
S	.922	1.049
S^2	.851	1.100

Fig. 4. Stem-and-leaf plot of absolute difference scores of 10 fingerprint experts and 10 latent marks (top panel) and summary statistics of absolute difference scores of 10 fingerprint experts and 10 latent marks.

Table 4
The differences in number of minutiae observed by the same examiner at different times. The bottom row is the mean difference per latent mark (A–J), and the right most column is the mean difference per examiner (1–10).

	Latent mark										Mean
Examiner	A	B	C	D	E	F	G	H	I	J	
1	1	1	4	1	1	2	3	2	0	1	1.6
2	8	3	5	1	1	2	2	5	2	2	3.1
3	1	3	3	3	6	4	9	9	1	2	4.1
4	2	3	2	5	0	1	1	0	0	1	1.5
5	6	2	2	3	4	1	3	3	0	3	2.7
6	9	4	2	1	4	6	0	5	1	1	3.3
7	0	4	5	2	4	3	3	7	0	0	2.8
8	3	1	4	0	6	2	1	4	2	0	2.3
9	4	3	9	0	4	4	3	1	1	3	3.2
10	1	0	0	1	4	1	4	1	0	0	1.2
Mean	3.5	2.4	3.6	1.7	3.4	2.6	2.9	3.7	0.7	1.3	2.58

This is an important step to remedy the problem. Once we know which latent marks are likely to cause consistency issues, we can recommend appropriate scientifically based procedures that attenuate the problem. For example, in latent marks of low quality, instructing a number of examiners to only mark minutiae that they have high confidence in. And then allow only use of those minutiae that have been marked across different examiners, thus using consensus in high confidence to determine the reliable features to use in such marks. Another approach is for mapping quality and clarity across a latent mark, so as to map high, medium, and low quality regions. Variability of feature selection may be lower if examiners are required to select only from the higher quality regions, but that may entail losing out on information. In this study we have identified a common phenomenon found in many expert domains, invite debate on the topic and its significance, and have suggested recommendations to deal with it.

Given that this present study has identified significant inter- and intra-observer variations during minutiae selection, it is relevant to ask: What impact can this have on the overall comparison decision making outcome? Is the lack of consistency a practical concern or an academic issue? The answer to these questions appears to be complex and depends on a number of factors. For example, in Evett and Williams [1] the variations in reported minutia did not totally predict the variations in overall decision outcome. In their study, Trials B, E, and F (which varied a lot in minutiae reported by some examiners), had 99%, 92%, and 100% consensus ($N = 130$) that the latent mark and the print originated from the same source. In other words, the variations (e.g., Trial F varied up to 42 minutiae), did not necessarily prevent experts reaching the same final conclusion. In contrast, other trials (such as Trial H) which had smaller variations, had less consensus on the final overall decision (in Trial H, e.g., 54% concluded they are likely from the same source, 38% reported insufficient detail to make a decision, and 8% reported they are not from the same source). Here the variation in feature selection appeared to be critical.

In Langenburg et al. [14] a similar trend was observed. In their Fig. 12, participants reported ranges (maximum differences) of 21, 17, and 12 minutiae respectively for Q1, Q4, and Q5 trials. However, trials Q1 and Q5 resulted in 100% consensus ($N = 43$) for the reported decision. Q4 on the other hand resulted in three errors, and the remaining participants nearly split on reporting "identification" or "inconclusive". Those that reported "identification" had a statistically significant higher likelihood of also reporting more minutiae. In this trial, it appeared that the number of minutiae observed directly correlated to the decision reported and was a critical part of the decision making process. Therefore, it is clearly a critical issue and variation needs to be researched and understood better.

It appears as a general trend that the reduction of available minutiae in a finger mark, especially to a point where the amounts may hover around categorical decision thresholds (i.e., "identification" vs. "inconclusive"), can lead to different decisions. Therefore, a possible best practice would be to identify *a priori* which marks are likely to produce such decision variations and apply special procedure, such as previously discussed (use of consensus high confidence minutiae, quality mapping, conservative selection procedures, etc.). Further research is recommended here, particularly to determine which suggested variation reduction technique is appropriate and effective.

4. Summary and conclusions

Feature selection during the analysis stage of a latent mark is important because it sets the stage and the parameters for comparisons and decision making. Although this stage is relatively robust, it is still susceptible to observer effects. In this study we found that the presence of a comparison 'target' print may affect the analysis stage. Furthermore, there is lack of consistency in the analysis not only among different examiners (e.g., reliability among examiners $r = .85$), but also within the same examiners analysing identical latent marks at different times (retest reliability $r = .86$).

The characterization of experts' consistency depends on the standard applied. If we examine the purest test of consistency, i.e., how consistent examiners are with themselves, then the retest reliability of $r = .86$, though far from perfect is respectably high; but using another standard, we find that at best (see footnote 5) only 16% of experts observed the exact same number of minutiae when analysing the same latent mark (40% of the experts were within one minutia difference, and 55% were within a difference of two minutiae).

Our study goes beyond establishing that analysis of latent marks by experienced latent print examiners is inconsistent. First, it demonstrates that the presence of a comparison print can affect the analysis of the latent mark. Second, it shows that examiners are inconsistent among themselves; i.e., different examiners vary in their analysis. Third, it reveals that the consistency of examiners with themselves varies; some examiners are relatively consistent with themselves and others are not. Fourth, we found that the lack of consistency does not only depend on the examiner in question, but it also highly depends on the nature of the latent mark itself.

For each of these findings we suggest potential recommendations to mitigate the problems. First, given the effects of the comparison print, we suggest that initially the analysis of a latent mark should be done in isolation from the comparison print. Furthermore, we do not rule out reconsideration of the analysis after exposure to the comparison print, but stipulate that this process, should it occur, must be clearly and transparently

documented, and justified. Further research needs to consider other ways to deal with variation in the analysis stage. One suggestion may be, for example, that examiners should mark confidence levels in minutia detection; thereafter they can only reconsider low confidence judgements but cannot change those that were analyzed initially with high confidence (see Dror [35] for details).

Second, given that examiners vary among themselves in their analysis, we support the development of a simple calibration tool that enables examiners to adjust their threshold so as to meet the standards in the field.

Third, given that some examiners are more consistent with themselves than others, we are confident that with proper selection of examiners with the right cognitive profiles specifying the exact skills needed for latent fingerprint examination and with proper training, can reduce the examiners' contribution to inconsistencies in finger mark analysis.

Fourth, given that the latent marks themselves play a major contributing role to the inconsistencies, and that these contributions vary with different marks, we suggest that such marks be subject to a different analysis procedure. Namely this would require using only higher confidence consensus minutiae that a number of independent examiners agree on.

Determining characteristics in finger mark analysis is critical and measures must be taken to minimize inconsistency and increase objectivity. These issues are not limited to fingerprint examination, there are similar issues across the forensic disciplines, including DNA. We do note that the potential problems with inconsistent analysis may be acute only when the comparison and latent mark are near the threshold for identification (and thus one analysis may result in identification whereas another analysis does not; problems may also arise around judgments of 'inconclusive' when another analysis may be sufficient for identification). When the decision is considerably beyond the threshold of determination, then these issues may not have important practical implication (as both analyses, although inconsistent, still will result in the same overall decision).

Understanding the cognitive issues involved in pattern matching and decision making, and researching them within the realm of fingerprinting is a promising way to decrease expert variation, improve the reliability of fingerprinting, and to gain insights into the human mind and cognitive processes.

Acknowledgments

We would like first to thank all the latent print examiners who took part in our studies. Without such cooperation this research would not have been possible. We also want to thank Joseph Almog, Camille Bourque, Rebbeca Bucht, Thomas Busey, Gerald Clough, Ralph and Lyn Haber, Anthony Laird, Danielle Mannion, Wayne Plumtree, Norah Rudin, and Arie Zeelenberg for valuable comments on an earlier version of this paper. However, any opinions, findings, and conclusions or recommendations expressed in this paper are the sole responsibility of the authors.

Correspondence concerning this article should be addressed to Itiel Dror, Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, England (e-mail: i.dror@ucl.ac.uk); further information is available at www.cci-hq.com.

References

[1] I.W. Evett, R.L. Williams, A review of the sixteen points fingerprint standard in England and Wales, The Print 12 (1) (1996) 1–13 (Also published in Fingerprint Whorld 21(82) (1995) 125–43; and in J. Forensic Ident 46(1), 49–73).
[2] I.W. Evett, Expert evidence and forensic misconceptions of the nature of exact science, Sci. Justice 36 (1996) 118–122.
[3] NAS, Strengthening Forensic Science in the United States: A Path Forward. National Academy of Sciences, Washington, DC, 2009.
[4] V.L. Phillips, M.J. Saks, J.L. Peterson, The application of signal detection theory to decision-making in forensic science, J. Forensic Sci. 46 (2) (2001) 294–308.
[5] J. Vokey, J. Tangen, S. Cole, On the preliminary psychophysics of fingerprint identification, Quart. J. Exp. Psychol. 62 (5) (2009) 1023–1040.
[6] R. Rosenthal, How often are our numbers wrong? Am. Psychol. 33 (11) (1978) 1005–1008.
[7] R.S. Nickerson, Confirmation bias: a ubiquitous phenomenon in many guises, Rev. Gen. Psychol. 2 (2) (1998) 175–220.
[8] I.E. Dror, D. Charlton, Why experts make errors, J. Forensic Ident. 56 (4) (2006) 600–616.
[9] I.E. Dror, D. Charlton, A. Péron, Contextual information renders experts vulnerable to make erroneous identifications, Forensic Sci. Intern. 156 (1) (2006) 74–78.
[10] I.E. Dror, J.L. Mnookin, The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensics, Law Prob. Risk 9 (1) (2010) 47–67.
[11] I.E. Dror, A. Péron, S. Hind, D. Charlton, When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints, Appl. Cogn. Psychol. 19 (6) (2005) 799–809.
[12] I.E. Dror, B. Rosenthal, Meta-analytically quantifying the reliability and biasability of fingerprint experts' decision making, J. Forensic Sci. 53 (4) (2008) 900–903.
[13] G. Langenburg, A method performance pilot study: testing the accuracy, precision, repeatability, reproducibility, and biasability of the ACE-V process, J. Forensic Ident. 59 (2) (2009) 219–257.
[14] 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 (3) (2009) 571–582.
[15] D.M. Risinger, M.J. Saks, W.C. Thompson, R. Rosenthal, The Daubert/Kumho implications of observer effects in forensic science: hidden problems of expectation and suggestion, Calif. Law Rev. 90 (1) (2002) 1–56.
[16] L.J. Hall, E. Player, Will the instruction of an emotional context affect fingerprint analysis and decision making? Forensic Sci. Intern. 181 (2008) 36–39.
[17] M. Saks, L.J. Concerning, E. Hall, Player 'Will the introduction of an emotional context affect fingerprint analysis and decision-making?', Forensic Sci. Intern. 191 (2009) e19.
[18] I.E. Dror, On proper research and understanding of the interplay between bias and decision outcomes, Forensic Sci. Intern. 191 (2009) e17–e18.
[19] R.B. Stacey, Report on the erroneous fingerprint identification bombing case, J. Forensic Ident. 54 (6) (2004) 706–718.
[20] K. Wertheim, G. Langenburg, A. Moenssens, A report of latent print examiner accuracy during comparison training exercises, J. Forensic Ident. 56 (1) (2006) 55–93.
[21] I.E. Dror, S. Cole, The vision in 'blind' justice: expert perception, judgment and visual cognition in forensic pattern recognition, Psychol. Bull. Rev. 17 (2) (2010) 161–167.
[22] W.C. Thompson, Interpretation: observer effects, in: A. Moenssens, A. Jamieson (Eds.), Encyclopaedia of Forensic Sciences, John Wiley & Sons, London, 2009, pp. 1575–1579.
[23] G. Langenburg, Pilot study: a statistical analysis of the ACE-V methodology – analysis stage, J. Forensic Ident. 54 (1) (2004) 64–79.
[24] B. Schiffer, C. Champod, The potential (negative) influence of observational biases at the analysis stage of finger mark individualization, Forensic Sci. Intern. 167 (2007) 116–120.
[25] B. Schiffer, The relationship between forensic science and judicial error: a study covering error sources bias and remedies, Ph.D. Thesis, Université de Lausanne, Lausanne, 2009.
[26] T. Busey, I.E. Dror, Special Abilities and Vulnerabilities in Forensic Expertise in Fingerprint Sourcebook, NIJ Press, Washington, DC, USA, 2010 (Chapter 15).
[27] I.E. Dror, The paradox of human expertise: why experts can get it, in: N. Kapur (Ed.), The Paradoxical Brain, Cambridge University Press, Cambridge, UK, in press, (Chapter 9).
[28] E.J. Potchen, T.G. Cooper, A.E. Sierra, G.R. Aben, M.J. Potchen, M.G. Potter, J.E. Siebert, Measuring performance in chest radiography, Radiology 217 (2000) 456–459.
[29] S. Bektas, B. Bahadir, N.O. Kandemir, F. Barut, A.E. Gul, S.O. Ozdamar, Intraobserver and interobserver variability of fuhrman and modified fuhrman grading systems for conventional renal cell carcinoma, Kaohsiung J. Med. Sci. 25 (2009) 596–600.
[30] C. Summerfield, T. Egner, Expectation (and attention) in visual cognition, Trends Cogn. Sci. 13 (9) (2009) 403–409.
[31] K.A. Ericsson, N. Charness, P.J. Feltovich, in: R.R. Hoffman (Ed.), The Cambridge Handbook of Expertise and Expert Performance, Cambridge University Press, New York, 2006.
[32] K.A. Ericsson (Ed.), Development of Professional Expertise, Cambridge University Press, New York, 2009.
[33] R. Rosenthal, R.L. Rosnow, Essentials of Behavioral Research: Methods and Data Analysis, third ed., McGraw-Hill Press, 2007.
[34] D.E. Krane, et al., Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation, J. Forensic Sci. 56 (2008) 1006.
[35] I.E. Dror, How can Francis Bacon help forensic science? The four idols of human biases, Jurimetrics: J. Law Sci. Techn. 50 (2009) 93–110.
[36] S. Harnad (Ed.), Categorical Perception: The Groundwork of Cognition, Cambridge University Press, New York, 1987.