Fingerprint Comparison Expertise

Jacqueline R. Claydon

School of Psychology

University of Kent

Word Count 40696

A thesis submitted for the degree of Ph.D. in the Division of Human and Social Sciences

December 2022

Abstract

Forensic fingerprint examiners are awarded ‘expert’ status within courts of law and are presumed to have specialist knowledge and ability that would not be found within the general population, yet questions arise as to whether expertise in forensic fingerprint comparison is a scientifically valid and reliable process. To explore this assumption, a novel online fingerprint aptitude test was created (Chapter 2) to measure the abilities of untrained controls (‘Novices’), fingerprint examiners-in-training (‘Trainees’), and qualified fingerprint examiners (‘Experts’) in a variety of fingerprint tasks (Chapter 3). Analyses focussed on group and individual differences in performance. Accuracy in latent print comparison differentiated the abilities of these three groups, and Experts demonstrated a level of performance that scarcely any Novices were able to match. This thesis therefore proposes a cognitive theory of fingerprint comparison expertise that reflects superior performance by fingerprint examiners in the most challenging aspects of fingerprint comparison. In addition, this theory suggests that there should be a clear separation in the performance of fingerprint examiners and untrained observers, with any overlap in abilities not a marker of expertise. This theory of expertise was further expanded by the application of a battery of perceptual tasks, designed to reflect the varied cognitive components of fingerprint comparison (Chapter 4). This test battery data demonstrates that fingerprint expertise is underpinned by an ability in feature matching, and mental rotation with non-fingerprint stimuli, and a role for global processing during latent print comparison. Conversely, perceptual processes such as visual short-term memory and visual search did not demonstrate the same relationship with fingerprint identification. This thesis concludes with suggestions for future research and recommendations for incorporating these findings into police selection processes for fingerprint examiner recruitment.

Acknowledgements

I would firstly like to express my sincere gratitude to my supervisor, Markus Bindemann, for his support and guidance with this research, and in particular for his encouragement in keeping this project on track throughout the challenges presented by a global pandemic. My research was supported by a studentship awarded by the School of Psychology, for which I am also extremely grateful. I would like to express my thanks to the forensic fingerprint examiners and trainees from the 'large UK police service' who took part in this research, without whom this project would have been impossible. I am also indebted to the senior examiners and staff with whom I was able to collaborate in the creation of the fingerprint test that was such a vital element of this research.

Lastly, I would like to thank my family and friends for their support throughout the past seven years at the University of Kent. Special thanks must also go to my Mum for her encouragement, wisdom, and the endless cups of tea whilst I have been working on this thesis.

Declaration

I declare that this thesis is my own work carried out under the normal terms of supervision.

Jacqueline R. Claydon

TABLE OF CONTENTS

Chapter 1

Expertise in Fingerprint Comparison: A Review

1.1. Introduction

Forensic person identification by fingerprint comparison has a long history in the UK, dating back to the foundation of the first Fingerprint Bureau at Scotland Yard in 1901. Despite advances in knowledge accompanying the discovery of DNA, fingerprints are still used extensively within the UK criminal justice system with the latest data showing those of 8,012,521 individuals are held on IDENT 1, the UK fingerprint database (Home Office, 2019). Fingerprint comparison has been widely regarded as an infallible method of identification, due to the presumed uniqueness of individual fingerprints (CPS, 2019; Galton, 1892), although several high-profile miscarriages of justice involving erroneous identifications have challenged this perception. In the US, for example, the fingerprints of Brandon Mayfield were wrongly identified as being present on detonator caps used in the Madrid train bomb (OIG, 2006), and in the UK the fingerprints of Scottish detective Shirley McKie were found within a murder scene which she claimed to have never entered (Cole, 2008). Both Mayfield and McKie faced criminal proceedings that were only halted when the fingerprint identifications were challenged and found to be erroneous.

Whilst identification errors undoubtedly have severe implications for the wrongly accused who may be unable to restore their tainted reputations (Gould & Leo, 2010), flawed criminal investigations also harm the victim and their relatives, and undermine public confidence in the criminal justice system (Poyser & Milne, 2010). Not least, there is the possibility that the real perpetrator remains at large, and the crime is unpunished. The impact of flawed forensic testimony was considered in a review by Saks and Koehler (2005) and found to be a common factor in 63% of wrongful convictions, with false or misleading forensic evidence accounting for 27% of those cases.

There has been growing criticism of the reliability of forensic comparison evidence within the scientific and legal communities. A National Academy of Science (National Academies of Sciences [NAS], 2009) report into standards and practices within forensic science found considerable variation in the level of performance and accreditation between laboratories, and a widespread failure to consistently provide robust evidence to link a specific individual with evidence from a crime scene. Following from NAS (2009), the President's Council of Advisors on Science and Technology (PCAST, 2016) was commissioned to conduct a wide-ranging review into the validity of forensic comparison, focusing on methods relating to DNA, bitemarks, firearms, footwear, hair, and latent fingerprint analysis. The review recommended greater foundational validity in forensic comparison with processes that are repeatable, reproducible, accurate and reliable. In relation to fingerprint comparison, the PCAST report considered that "many examiners can, under *some* circumstances, produce correct answers at *some* level of accuracy" (p.95). However, the false positive rate in which fingerprints were incorrectly identified as being from the same source was higher than the report authors had expected, and the influence of contextual and confirmation bias in identification decisions was deemed problematic.

In conclusion, both NAS (2009) and PCAST (2016) considered that although recently published research had made positive inroads into establishing a scientific basis for fingerprint comparison, a cross-disciplinary response to future research between forensic practitioners and cognitive scientists was urgently required. This should focus on determining error rates and obtaining empirical evidence to account for examination procedures and their associated cognitive processes.

1.2. The Nature of Fingerprints

Fingerprints, also referred to as friction ridge skin patterns, begin to form in humans between the tenth and fourteenth weeks of gestation (Kücken & Champod, 2013), with their appearance influenced by factors such as foetal position and flow of amniotic fluid (Jain et al., 2002). This random nature of foetal fingerprint development means that even monozygotic twins have different friction ridge skin patterns despite being genetically identical, although a shared genetic basis means there will be some similarity between parents and children and between siblings (Jain et al., 2002). Because the pattern of ridges is formed in the dermal layer of the skin during their development, this means they are permanent throughout life unless altered by injury (Meuwly, 2014), and will be fully restored when only the surface layer of the skin is damaged (Champod et al., 2004).

It would be impossible to record and measure the fingerprints of the world population, therefore the uniqueness of fingerprints can only be predicted with theoretical modelling. This is based on the probability of inter- and intra- similarity in fingerprint features and posits that if a forensic expert matches twelve corresponding points on a pair of fingerprints the likelihood of an erroneous identification is so small to be ignored (Pankanti et al., 2002). However, the lack of empirical evidence to support fingerprint uniqueness means this can only be an assumption rather than an objective measure (Kadane, 2018; Saks et al, 2005).

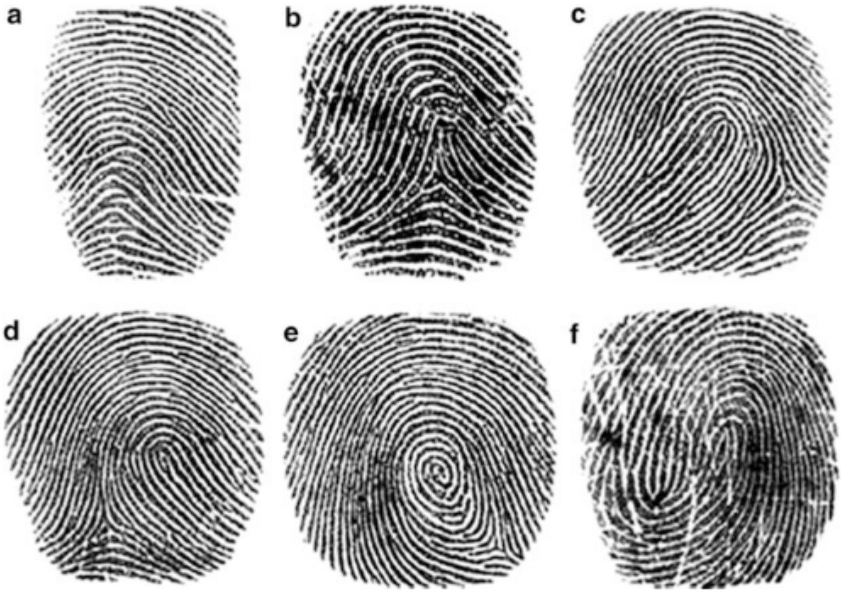
1.2.1. Fingerprint Features

To be an effective means of verifying the identity of an individual within the population, any physical marker need to be universally available, in a permanent form, readily collectible, and distinctive (Jain et al., 2002). The wide variation in permanent features found within fingerprints together with their largely random morphogenesis, an embryological process that causes organs and cells to develop their shape, means they are an ideal resource for forensic person identification. For fingerprint comparison, friction ridge skin is generally classified according to three levels that relate to the visibility of features.

Level 1 Features. Friction ridge skin comprises alternating valleys and ridges with a typical fingerprint having up to 150 ridges (Jain et al., 2006). When viewed during fingerprint examination, ridges typically appear as dark pixels and valleys as the lighter pixels, although this is dependent on image processing (Bigun, 2014). Level 1 features relate to the general pattern of the friction ridge skin and are classified according to whether they appear as arches (simple or tented), loops (left, right or twin) or whorls (Figure 1.1). It is estimated that only 1% of fingerprints would not demonstrate one of these identifiable patterns (Bigun, 2014).

Figure 1.1

Example of Level 1 Features

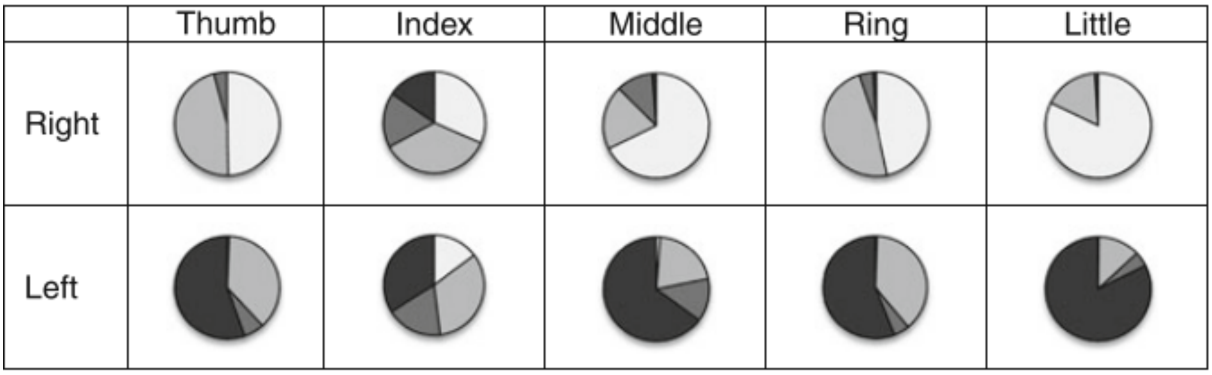


Note. Figure shows: a) Arch, b) Tented Arch, c) Left Loop, d) Right Loop, e) Whorl, f) Tented Loop (from Bigun, 2014).

Level 1 features are used for classifying fingerprints by pattern or ridge flow. They are visible without magnification (Champod et al., 2004), but are not unique and would therefore be examined in conjunction with other levels of features for identification purposes (Jain et al., 2006). The centre of the pattern is known as the core and the point at which three ridges meet in a triangular pattern is referred to as the delta (Meuwly, 2014). Most fingerprint patterns are loops (57%), followed by whorls (35%) and arches (7%), with prevalence of pattern type varying according to finger (Figure 1.2).

Figure 1.2

Prevalence of Fingerprint Patterns by Finger

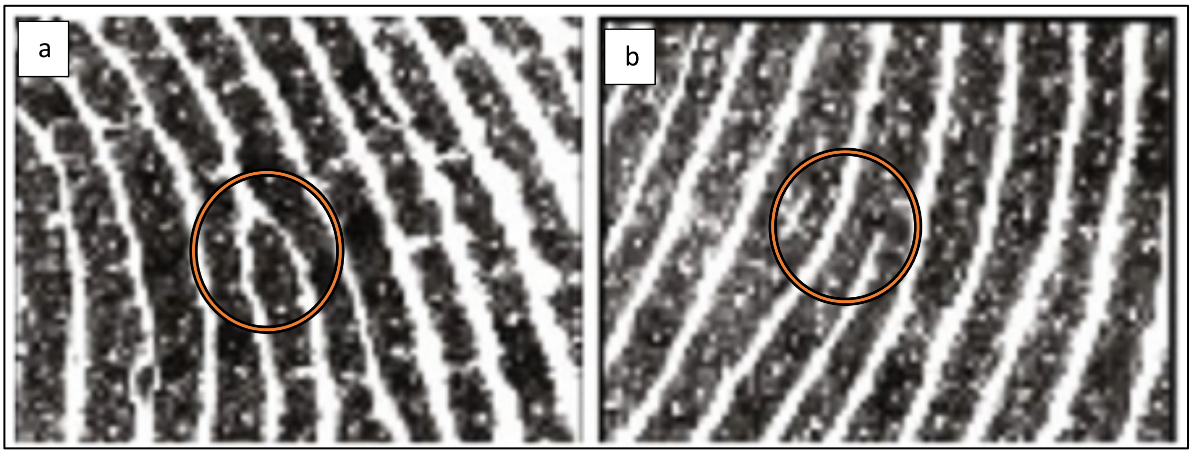


Note. Colours indicate left loop (black), right loop (white), arch (dark grey), whorl (light grey). From Meuwly (2014).

Level 2 Features. These are referred to as minutiae and are typically 0.1 to 0.5 mm in size (Bigun, 2014). In their basic form these are classified as ridge endings and bifurcations, that is, they are the point at which a ridge or valley ends, or a point at which the ridge divides into two parts (Figure 1.3). The identification of a fingerprint donor is possible from the examination of Level 2 features (Jain et al., 2006) and they require 5x to 10x magnification to be visible (Champod et al., 2004). Minutiae positions are useful in determining the spatial relations between features as they form reference points between ridges: counting ridges in a fixed line between minutiae is a means of identifying matching or non-matching features (Bigun, 2014). Most of the variability within fingerprints is accounted for by the infinite combinations of minutiae (Meuwly, 2014), and accordingly these are the features most widely used to compare fingerprints, both by human examiners and automatic systems (Bigun, 2014).

Figure 1.3

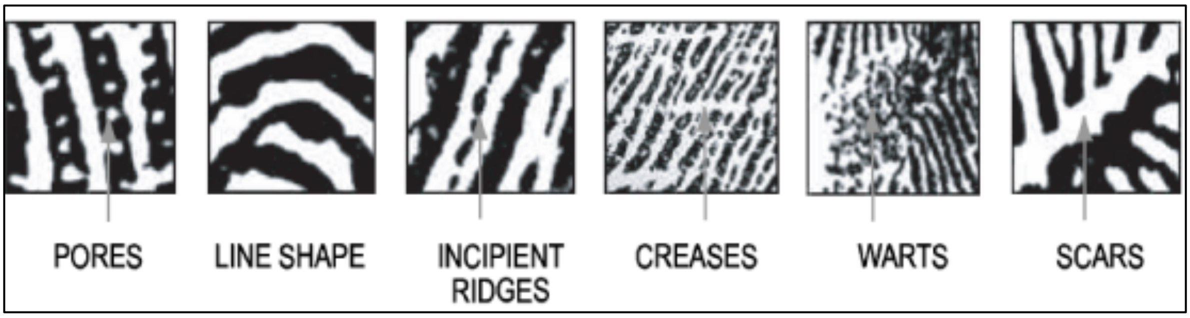
Fingerprint Image Depicting, a) Ridge Ending in a Valley Bifurcation, b) Valley Ending in a Ridge Bifurcation (Taken from Bigun, 2014).



Level 3 Features. These features are examined in conjunction with those classified under Levels 1 and 2 (Meuwly, 2014) and refer to qualitative rather than quantitative details within ridge formations (Jain et al., 2006), which are only visible under high magnification (Champod et al., 2004). Level 3 features typically include those shown in Figure 1.4 although they are not defined, and there is no consensus between examiners as to what constitutes a Level 3 feature (Anthonioz et al., 2008).

Figure 1.4

Examples of Typical Level 3 Fingerprint Features (Taken from Jain et al., 2006).



Overall, the infinite combinations of Level 1, 2 and 3 features that may be visible within each individual fingerprint, coupled with arbitrary influences on friction ridge skin development during gestation, tend to support arguments in favour of fingerprint distinctiveness. Within forensic settings, the persistence and durability of fingerprints (Champod et al., 2004), and the assumed uniqueness of their arrangement of features (CPS, 2019; Scientific Working Group on Friction Ridge Analysis, Study and Technology [SWGFAST], 2013), allow fingerprint examiners to provide evidence regarding the source of a fingerprint (Forensic Science Regulator [FSR], 2017).

1.3. Fingerprint Examination

Within the UK, accredited fingerprint examiners are regarded as expert witnesses whose knowledge and experience allows them to provide opinion testimony to a court regarding the source of a fingerprint found at a crime scene (CPS, 2019). The purpose of fingerprint examination is to determine whether two areas of friction ridge skin are from the same person or different people. When a person is arrested, sets of fingerprints (known as Ten Prints) are obtained using either a fingerprint scanner or by taking inked impressions. Ten Prints comprise rolled impressions of each finger, an impression of the four fingers and the thumb of each hand taken simultaneously, and palm prints (Home Office, 2019). Examiners compare Ten Prints, referred to as exemplars, with latent marks obtained from crime scene surfaces to identify the source of the marks (FSR, 2017).

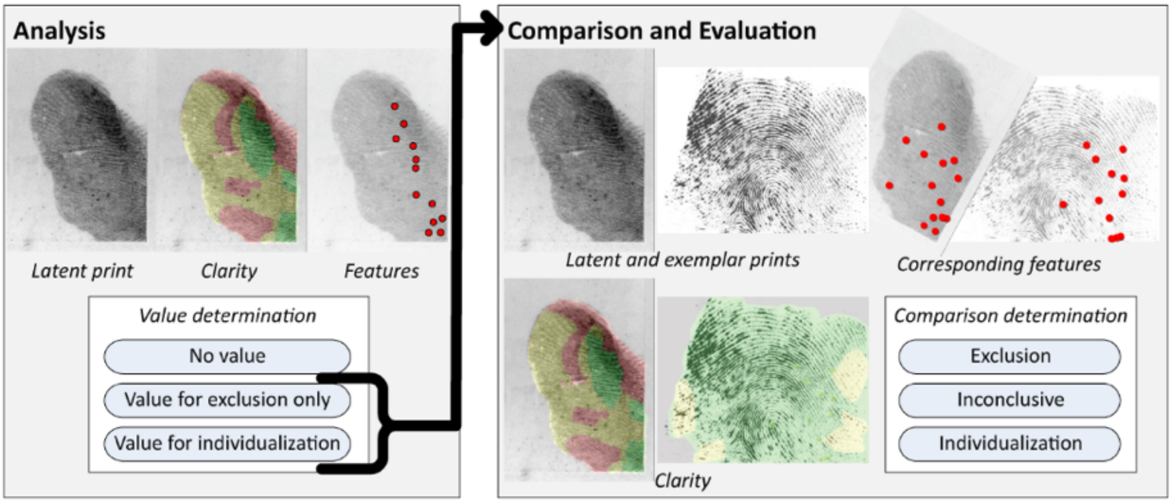
Within the UK and many other parts of the world the examination of friction ridge skin is conducted using the Analysis, Comparison, Evaluation and Verification (ACE-V) method. This was originally devised as an examination structure applicable to any forensic research (Huber, 1959), and has been expanded, notably by Ashbaugh (1999) and Champod et al. (2004), to reflect the fingerprint comparison process in more detail. In spite of its widespread adoption as a method of comparison, ACE-V has been criticised as an untestable scientific method which is lacking in formal structure (Mnookin, 2007), due in part to forensic examiners having insufficient experimental expertise to determine the scientific validity of the method (Haber & Haber, 2007). Nonetheless, ACE-V remains the most widely used method of latent print comparison (Ulery et al., 2011).

1.3.1. ACE-V Methodology

The ACE-V procedure is described in resources issued by the Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST, 2013) and the Forensic Science Regulator in the UK (FSR, 2017). When this process is carried out by an individual examiner it is referred to as ACE, and ACE-V refers to a subsequent verification procedure by an independent examiner. An example of ACE workflow is shown in Figure 1.5. Although the methodology for each phase of ACE is described sequentially, it is regarded as an iterative process with examiners returning to each element of ACE as required in order to reach a conclusion (Home Office, 2019; SWGFAST, 2013; Vanderkolk, 2011). Application of ACE is not subject to formal standards and is reliant on the skill and experience of the individual examiner (Ulery et al., 2014).

Figure 1.5

Example of ACE Workflow from Ulery et al. (2014).



Analysis. In this stage the latent mark is examined to determine whether it contains sufficient visible detail with which to proceed with a Comparison (Home Office, 2017). The examiner needs to consider the circumstances under which the mark was left, the effects of pressure and distortion on its appearance and the overall clarity of the image (SWGFAST, 2013; Vanderkolk, 2011). The examiner provides a determination of suitability which may be "value for identification" (VID), "value for comparison/exclusion" (VEO) or "no value" (NV) (SWGFAST, 2013).

Comparison. If the latent mark is of value, the examiner will compare it with a known source print (from Ten Prints) to determine areas of similarity or discrepancy in relation to features, sequences and spatial relations (SWGFAST, 2013). This is a process of mental comparison and assessment beginning with Level 1 features (SWGFAST, 2013), and may be accompanied by physical measurement of the images (Vanderkolk, 2011).

Evaluation. In this stage the examiner determines whether the latent mark is identified to an individual (individualization), excluded for an individual (exclusion), inconclusive, or contains insufficient detail with which to draw a conclusion (Home Office, 2017; SWGFAST, 2013). In reaching a conclusion, the examiner may apply the "One Discrepancy Rule": if one or more friction ridge detail appears on the exemplar but not on the latent mark, the comparison can be determined as an exclusion (SWGFAST, 2013). Conversely, if most features between the prints are in agreement, then a single *dissimilarity* will not necessarily result in an exclusion if it can be accounted for by distortion in the print or the presence of a scar (SWGFAST, 2013).

Verification. In the UK, the latent mark and fingerprints are subsequently compared by an independent examiner using ACE-V in order to verify or refute the conclusion of the original examiner (Home Office, 2019). This process can be blind, in which the verifier is unaware of the original outcome, or an open process in which the previous decision is known (FSR, 2017), with a recommendation that all individualizations should be subject to verification (SWGFAST, 2013). The PCAST (2016) report suggested the increased use of blind verification procedures would identify many of the errors committed during an examination.

Although procedures such as ACE and ACE-V are designed to maximise the likelihood of an error-free identification by fingerprint examiners (FSR, 2017; SWGFAST, 2013), the reliability of fingerprint examiners' identification decisions is difficult to determine, largely due to the ground truth as to the source of a crime scene mark being unknown within criminal investigations (Cole, 2008). The next section will focus on findings from empirical studies that have attempted to provide both quantitative and qualitative understanding of the accuracy of fingerprint comparison decisions by professional fingerprint examiners.

1.4. The Accuracy of Fingerprint Examiners

Although an examiner's identification decision may be supported by a confession from a perpetrator, or with additional forensic evidence such as DNA, it would be impossible to predict the accuracy of fingerprint examination on the basis of positive identifications, or individualizations, alone. A definition of accuracy needs to incorporate several components, namely whether expert performance is superior to novices, the reliability of comparison decisions in terms of likely error rates, and the probability of consistent and repeatable inter- and intra-examiner decision-making. These aspects of fingerprint examiner accuracy will be considered in the following section.

1.4.1. The Performance of Experts and Novices in Fingerprint Comparison

The abilities of fingerprint examiners and novice participants have been compared in several studies designed to gain an understanding of whether expertise could be explained by differences in task performance. They also serve as an indicator of the level of accuracy within expert and untrained populations.

In one study, expert and novice accuracy was compared in rating the similarity of pairs of simulated crime scene prints that were either matching, non-matching, or similar-non-matching (Tangen et al., 2011). This experiment was not designed to be analogous to case work, and the inclusion of similar-non-matching print pairs was a means of countering a weakness of typical proficiency tests in which all stimuli are either a matching pair or a pair of distractors, thus making them unnaturally easier to discriminate (Thompson et al., 2013). The inclusion of similar non-matching pairs also represents the most difficult comparison likely to be undertaken by examiners. With matching pairs, expert accuracy was 92.1% and novice accuracy was 74.6%, and with non-matching pairs, experts made no errors while novice accuracy was 77.0%. Even when comparing *similar*-non-matching pairs, expert accuracy was very high (99.3%), while novice accuracy was considerably lower at 44.8%, reflecting the misidentification of 55.2% of non-matching prints as being matching pairs. Therefore, the accuracy advantage for experts compared to novices could be accounted for by their ability to discriminate highly similar pairs of prints.

In a later study, genuine crime scene prints were used to compare the matching abilities of novices, new fingerprint trainees, examiners with intermediate experience and experts (Thompson et al., 2014). The print pairs were either matching, non-matching, or similar-non-matching. Across 45 trials, participants compared a crime scene latent with a rolled exemplar and provided either a match or no match evaluation. An interaction between expertise and trial type was revealed in match trials, with new trainees less accurate than the other groups. In non-match trials, both novices and new trainees were the least accurate groups. Comparing performance in similar-non-match trials showed experts and intermediate examiners were more accurate than both new trainees and novices, thereby supporting an earlier conclusion by Tangen et al. (2011) that expertise reflects an ability to identify very similar differences in fingerprints. There was little difference in the performance of experts with an average of 17.5 years of experience and intermediate examiners with an average of 3.5 years of experience.

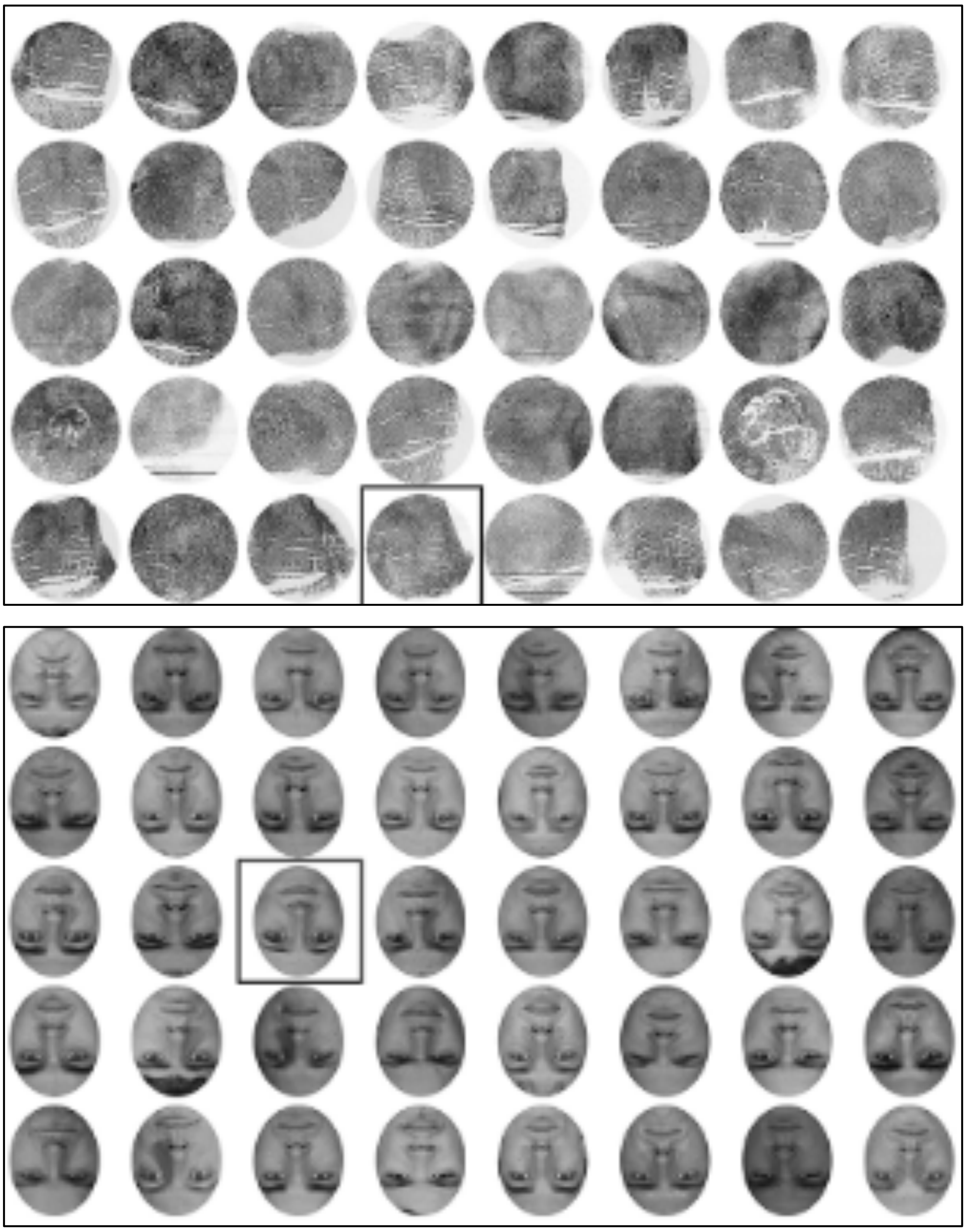
To examine whether an increased awareness of fingerprint comparison would lead to improved matching performance, an explanation of ACE methodology and descriptions of fingerprint features were given to a group of naïve participants, referred to as the ‘trained’ group (Stevenage & Pitfield, 2016). The performance of the trained group, a novice group, and a group of fingerprint examiners was compared in a series of tasks to identify fingerprints as the ‘same’ or ‘different’. Analyses revealed that experts were more accurate (99.5%) than the trained (86.4%) and novice participants (82.2%), with a significant difference between the trained and novice groups reflecting a small benefit of training. In addition, the response times of the trained group were longer than novices, but equal to the experts in ‘different’ trials, suggesting an increased awareness of task demands following training.

In a further comparison of expert and novice performance, observers judged whether a single fingerprint belonged to the same person who had provided the other four fingerprints in an array (Searston & Tangen, 2017a). The aim was to determine whether expertise in fingerprint comparison could extend to coarser levels of fingerprint categories, and whether familial resemblance within fingerprints could be discriminated. Experts were more accurate (75.5%) than novices (68.7%) in identifying whether or not the prints were from the same family but were less confident in their decisions, with neither group demonstrating a response bias. The authors posit this as evidence of flexibility in perceptual expertise, with experience in fingerprint comparison generalizing to fingerprints in general. Given that the novice group were undergraduate students with no experience with fingerprints, a performance advantage of only seven percent for the expert group supports the feasibility of comparing expert and novice accuracy in fingerprint comparison.

The performance of experts and novices was subsequently compared in a study to determine whether expertise with fingerprints generalized to expertise within an unfamiliar category (Searston & Tangen, 2017b). In Experiment 1, participants identified whether a fingerprint loop pattern was present in an array of whorl patterns (and vice versa), and whether an inverted female face was present in an array of inverted male faces (and vice versa). Examples of these arrays are shown in Figure 1.6. Inverted unfamiliar faces were chosen for the comparison category as they may be regarded as novel stimuli which is therefore perceived differently to familiar faces (Megreya & Burton, 2006).

Figure 1.6

An Example of a Fingerprint Array with a Loop Target Marked (Top), and an Inverted Face Array with a Female Target Marked (Bottom).



Experts were more accurate with fingerprints (92.0%) than novices (32.2%), with no difference in accuracy between experts and novices for faces, although experts were slower to identify faces than fingerprints. In Experiment 2, pairs of mated and non-mated fingerprints, and matching and non-matching inverted faces, were presented for 400ms prior to observers making a same or different finger or identity judgement. The results echoed those of Experiment 1, with experts more accurate within their own area of expertise, and with no difference in face-matching accuracy between the groups. These findings suggest that fingerprint expertise does not generalize to processing another class of stimuli and is therefore likely to be constrained within the domain of expertise.

Overall, these studies converge in showing that forensic fingerprint experts are more accurate than untrained observers in identifying matching and non-matching pairs of fingerprints. This accuracy advantage for experts increases when examining similar non-matching pairs of prints, thereby suggesting that identification accuracy may be underpinned by the ability to discriminate minute differences in pairs of fingerprints. Expert accuracy was higher than novices in unfamiliar fingerprint categorisation (familial similarity within prints), but this did not extend to greater accuracy with non-fingerprint stimuli (faces), thereby suggesting a domain-specific perceptual ability for fingerprint stimuli in examiners. Although novices were less accurate than fingerprint examiners, the results from these studies nonetheless demonstrate the feasibility of testing the performance of untrained observers in fingerprint comparison tasks, despite their inexperience with this class of stimuli.

1.4.2. Fingerprint Examiner Error Rates

Researchers have used black box studies, which measure procedural outcomes without considering *how* they have been reached, to determine the accuracy of forensic fingerprint examiners based on the number of correct and erroneous identifications. They have also explored inter- and intra-examiner consistency in comparison decisions and whether these reflect a reliable and repeatable procedure. These studies use known source fingerprint stimuli, which alleviates the ground truth problem, and focus on the decision-making process rather than any underlying cognitive processes. The following key studies have reported the accuracy and error rates of practicing fingerprint examiners in experimental conditions. The research methodologies and outcomes are described briefly below to allow comparison between studies.

In an early study, the performance of latent print examiners with less than one year of experience was compared against examiners ranging in experience from more than one year to thirty years (Wertheim et al., 2006). Observers were required to match ten latent prints to prints contained within eight full sets of fingerprints. All latent prints had sufficient information for individualization and were graded according to the perceived difficulty they would present to an examiner. Therefore, those with less experience examined easier prints. The less experienced examiners made 2% of identification errors, in which the wrong person was matched to the latent mark, and a further 1% of errors were deemed to be clerical, such as inputting the wrong response on the answer sheet. This compared with an error rate of 0.2% for the more experienced examiners, although they also made 1% of clerical errors. Overall, 48 of the 92 examiners made no errors, and 22 made at least one error.

The accuracy and reliability of comparison decisions was further considered in collaborative research with FBI scientists and featured 169 forensic latent fingerprint examiners (Ulery et al., 2011). Each examiner compared 100 print pairs using ACE. During Analysis, prints were rated as having no value (NV) for comparison, valid for identification (VID) or valid for exclusion (VEO). After comparison, examiners evaluated each print pair as either an individualization, an exclusion or inconclusive. Of the 4083 non-matching pairs that were VID, 6 pairs were incorrectly identified as matching pairs, thereby reflecting a false positive rate of 0.2%, or 1 error in 604 cases (PCAST, 2016). In relation to false negatives, 450 matching pairs from the 5969 deemed VID were incorrectly identified as non-matching, resulting in an error rate of 7.5%, with 85% of examiners making at least one such error (PCAST, 2016).

A further study to evaluate the accuracy of practicing fingerprint examiners (Pacheco et al., 2014) was commissioned by the National Institute of Justice in the US. This is cited by PCAST (2016) as evidence of performance but has not so far been published in a peer-reviewed scientific journal. For this experiment the researchers created their own fingerprint stimuli but did not incorporate similar-non-matches. Each examiner performed an ACE examination of 40 latent prints using 10 sets of known-source prints, with 70% of latent prints capable of being matched. A false positive rate of 3% was reported, equating to 1 error in 24 comparisons, with an error rate of 4.2% if inconclusive evaluations were included. The authors report 35 erroneous conclusions due to clerical error, and although they were unable to substantiate this claim, if these errors were excluded this would lower the false positive rate to 0.7%, or 1 error in 73 cases (PCAST, 2016).

Overall, these studies tend to show that, in laboratory settings, forensic fingerprint examiners tend to commit few errors, and are more likely to provide false negative than false positive identifications. However, the actual error rate fluctuates on a study-by-study basis, with some errors explained as clerical mistakes such as entering the wrong result onto an answer sheet (Pacheco et al., 2014), rather than cognitive failures. One important consideration is that although error rates are relatively low, these reflect erroneous identifications by several examiners rather than errors being committed by the same individuals. In forensic fingerprint comparison this *one* error can prove highly detrimental to the individual concerned, as well as undermining the credibility of forensic fingerprint examination decisions.

Whilst research undoubtedly shows that fingerprint examiners are able to perform at a high level of accuracy, accurate error rates are difficult to determine due to differences in experimental methodologies, type of stimuli presented, and the analyses used (PCAST, 2016). In addition, some studies allowed examiners to adopt elements of ACE during comparison (e.g., Pacheco et al., 2014 (as cited in PCAST, 2016); Ulery et al., 2011) but this was not incorporated into all experimental designs as these were not necessarily intended to be analogous to case work. PCAST (2016) recommends the error rates from Pacheco et al. (2014) and Ulery et al. (2011) are highlighted to jurors to allow them to determine the probative value of examiner testimony. Conversely, the authors of the studies do not suggest that erroneous identifications observed under experimental conditions would translate into reliable error rates for fingerprint comparison within forensic settings.

1.4.3. Consistency in Fingerprint Examiners' Decisions

Although empirical evidence suggests examiners demonstrate a high level of performance during fingerprint comparison, correct responses alone cannot be a true reflection of accuracy in the absence of data relating to the repeatability and reliability of decisions made by experts (PCAST, 2016). Such measures consider whether experts presented with the same fingerprint image are likely to reach the same identification decision, whether they use the same bases to reach the same conclusion, and if an individual examiner will make the same decision when examining the same image on separate occasions. To examine whether fingerprint comparison is a scientifically valid process, these key questions have been examined by researchers.

Inter-expert decision-making is reported to vary considerably in both the Analysis and Evaluation phases of fingerprint examination (Ulery et al., 2011). Consensus, in which 90% of a sample of 169 examiners provided the same determination, was reached for only 66% of the latent prints viewed during Analysis. During Evaluation, consensus was reached on only 73% of matching pairs and 56% of non-matching pairs of prints. With those fingerprints rated as valid for identification, the true positive rate for comparisons by individual examiners ranged between 29% to 94% and the true negative rate was between 5% and 100%.

To examine whether comparison errors were more or less likely to be repeated, seventy-two examiners from the original sample in Ulery et al. (2011) were re-tested seven months later using twenty-five previously examined fingerprint pairs (Ulery et al., 2012). Where an examiner had made a false negative or false positive error during initial testing, these were re-assigned, although observers were not aware that they would be examining previously encountered fingerprints. Ratings of comparison difficulty on a scale between "obvious" to "very difficult" were also provided. In trials with mated pairs, examiners repeated 89.1% of their individualization decisions and in trials with non-mated pairs, 90.6% of exclusion decisions were the same. None of the six false positive errors revealed during the initial test were repeated, and there were no new false positive errors. However, of the 226 false negative errors identified on the initial test, 30% were repeated subsequently. Much of the inter- and intra- examiner variability related to the comparison of pairs of prints that had low agreement between examiners, such as borderline or complex image pairs, rather than reflecting differences in the performance of examiners. In addition, the comparisons presented in this test were intentionally challenging, with higher difficulty ratings associated with lower repeatability and reproducibility.

In summary, these studies provide evidence of inconsistencies within fingerprint examination outcomes, with clear inter-examiner differences in the identification decisions reached. In addition, individual examiners demonstrated a range of true positive and true negative responses and did not necessarily replicate identification decisions when examining previously seen fingerprints. These studies show that, despite viewing the same stimuli, examiners' identifications are based on their self-determined threshold limit at which they are willing to commit to a decision. Converging evidence from research by Ulery et al. (2011)

therefore reinforces the importance of the Verification stage of ACE in identifying inconsistent decisions by fingerprint examiners.

1.4.4. Examiner Differences in Information Used for Comparison

Counting the number of minutiae present in a latent print has been identified as the key factor in value determinations by examiners during Analysis, and the means by which prints are classified as containing sufficient information to be valid for identification or exclusion, or to be of no value for comparison (Ulery et al., 2013, 2014). During the Evaluation phase, the relationship between the minutiae observed on the latent and those on the comparison print, and the location, number and type that are in agreement, is the key driver of an examiner's comparison decision (Neumann et al., 2013).

In research to examine the consistency of minutiae counting, examiners were asked to report all minutiae in a latent print presented alone ('solo condition') and to subsequently report the number of minutiae present in the same latent presented alongside a matching print ('pair condition') (Dror et al., 2011). Not only were different numbers of minutiae reported in the solo and pair conditions, which suggested a potentially biasing effect resulting from the target comparison, but considerable inconsistency between examiners was observed in the numbers of minutiae reported on each latent print. The largest difference related to latent mark A, in which the minutiae count ranged between 9 and 30, and the smallest difference in count was between 6 and 14 (mark D). This suggests either individual differences in the number of minutiae perceived or differences in the interpretation of the meaning of "minutiae".

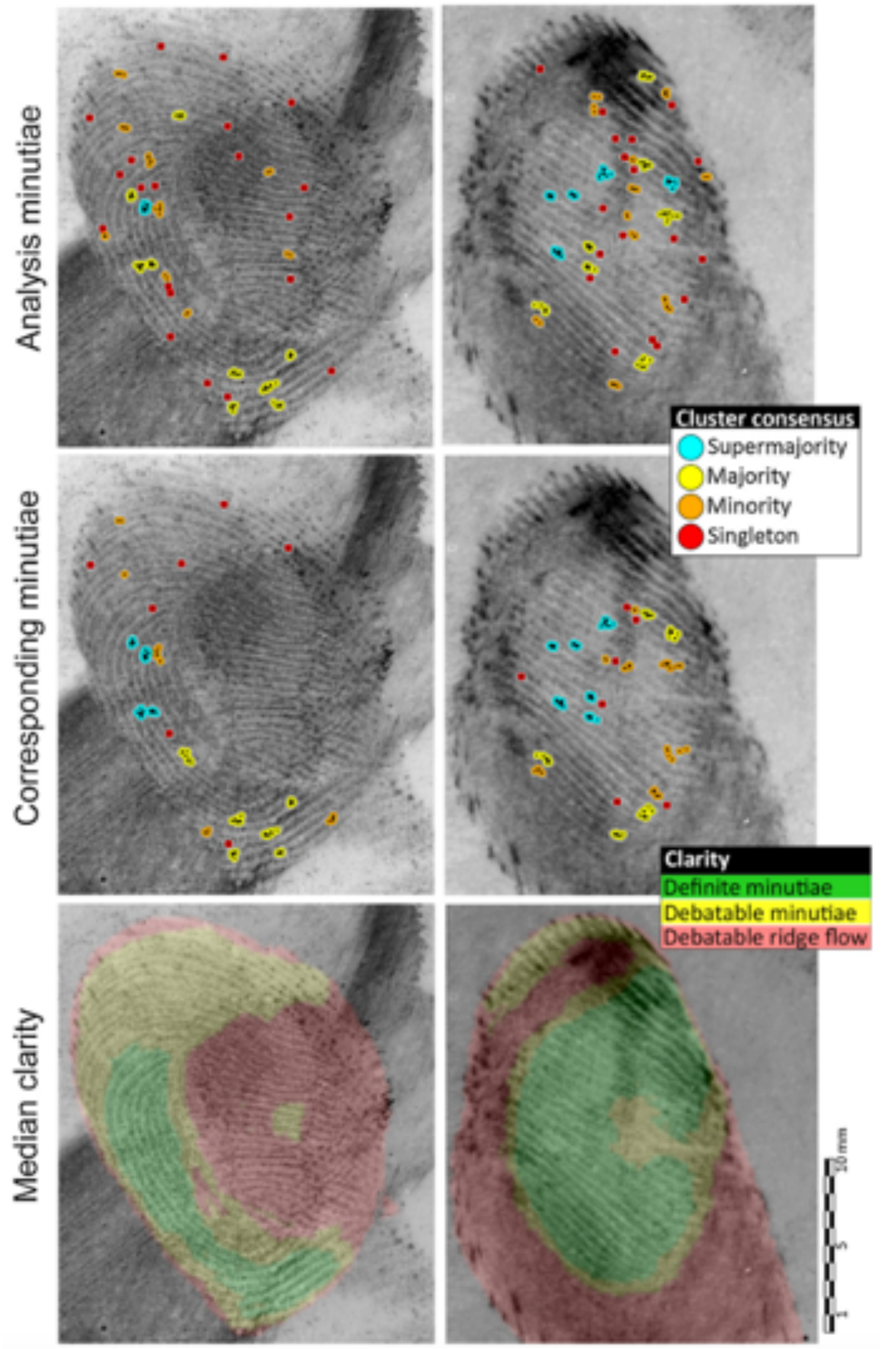
In a further measure of inter- and intra-examiner consistency, a new group of examiners was asked to report the number of minutiae on ten novel latent marks, repeating the task some months later (Dror et al., 2011). Retest-reliability within the same examiner was .86, and between examiners was .85, which appears to reflect a high level of consistency. However, only 16% of the experts reported exactly the same number of minutiae on identical fingerprint pairs on two separate occasions, and reliability in this instance only measured the number of minutiae counted but did not identify whether the same minutiae were counted on each occasion.

In a later study to assess the consistency of minutiae identification, fingerprint examiners were required to annotate fingerprint images according to specific mark-up criteria (Ulery et al., 2016) (see example in Figure 1.7). During Analysis, in which only the latent print was presented, examiners denoted the levels of clarity of different areas of the print, and the location and type of all minutiae. They also determined whether the image was valid for identification (VID), valid for exclusion (VEO), or of no value (NV). Images deemed to be VID or VEO were presented alongside an exemplar for Comparison, and this too was annotated according to clarity and corresponding minutiae. Examiners could revise their mark-up of the latent image and determination, and areas of agreement or disagreement between the latent and exemplar and clarity were marked. Finally, a comparison determination and confident rating was provided.

The overall mean probability of Analysis phase minutiae being reproduced by a second examiner was 63%, with lower mean reproducibility for unclear areas (47%) and higher reproducibility in clear areas (70%). During Comparison, if an examiner annotated a minutia on the latent print and the corresponding minutia on the exemplar, the probability of replication by a second examiner was 69% with clear minutiae and 47% with unclear minutiae. Although some variation between examiners was attributed to image quality, lack of consensus relating to the nature of minutiae and the application of different standards of mark-up were regarded as significant factors (Ulery et al., 2016).

Figure 1.7

Two Examples of Latent Mark-up from Ulery et al. (2016).



Note. Marked minutiae are shown as black dots, with clusters reflecting the number of examiners who marked that area of the print. Row 1 reflects Analysis; Row 2 reflects Comparison; Row 3 reflects clarity.

These studies again reflect clear inconsistencies in fingerprint comparisons conducted by experienced examiners, both at the level of the individual and between different examiners. Whether this is due to perceptual differences, in that each examiner perceives stimuli differently, or the semantic understanding of the term ‘minutiae’ differs, is difficult to determine, a fact which is acknowledged by the authors of these studies.

1.4.5. Differences in The Level of Sufficiency for Fingerprint Comparison

In the absence of a requirement to identify a minimum number of corresponding minutiae in a latent mark and an exemplar print, an individualization is reliant on the decision-making threshold of each examiner (SWGFAST, 2013). This is referred to as the ‘level of sufficiency’ and is the point at which an examiner identifies enough corresponding features to conclude that two fingerprints are from the same source.

In a study to compare inter-examiner levels of sufficiency, fingerprint examiners each examined twenty-two mated (matching) and non-mated (non-matching) latent and exemplar print pairs using ACE (Ulery et al., 2014). During Analysis, the latent image was annotated according to image clarity, location, and types of features present, along with a value determination of VID, VEO or NV provided. With latent marks rated as VID or VEO, these were presented alongside an exemplar for Comparison/Evaluation whereby both images were annotated according to clarity, features, and any correspondence or discrepancy between images. A comparison determination (individualization, exclusion or inconclusive) was provided along with a rating of comparison difficulty. Examiners were able to revise their Analysis determination at any stage.

During Analysis, seven minutiae was the approximate threshold for value determinations, with only 1% of latents rated as VID when fewer than seven were annotated. However, a minutia count of up to twenty-seven was observed for latents rated as VEO, noticeably when image quality was low. During Comparison/Evaluation, seven minutiae was also the threshold for determination: when the number of corresponding minutiae exceeded seven, most comparisons were rated as individualization and only 1% of individualizations were based on fewer. Some of the examiners adhered to a 12-point standard, in line with their usual working practices, which was reflected in a decision threshold of twelve corresponding minutiae. Overall, the difference in the number of individualizations between examiners using a 12-point standard (69%) and those with their own level of sufficiency (62%) was not significant.

For the majority of image pairs, there was considerable variation between examiners in relation to minutiae counts and determinations. Although the count of an individual examiner was predictive of their own determination, it was only a weak predictor of the determination of another examiner, with image quality and non-minutiae (Level 3 features) only explaining a small proportion of variance. The researchers were unable to conclude whether inter-examiner differences were dependent upon an examiner's interpretation of the available information with a fingerprint, or whether they reflected differences in the reporting of analyses, with a key factor being the absence of a common definition of minutiae. They suggest examiners may make quick and instinctive determinations during Comparison which influence the features they annotate. Conversely, having reached a determination, examiners may change their annotations in order to audit their decision. The authors do not suggest that such revisions are done to falsify decision- making but are an example of the difficulties encountered when trying to document the cognitive processes involved in fingerprint expertise.

1.5. The Nature of Expertise in Fingerprint Comparison

Expertise can be regarded as a level of performance attained by the most competent practitioners that is reflected in an ability to undertake a familiar task in an intuitive, automatic, and effortless manner (Kahnemann & Klein, 2009). The examination of fingerprints and the identification of relevant features is solely reliant on the expertise of the examiner rather than any quantitative measure of comparison (Hicklin et al., 2019).

Although many countries still require numerical points of similarity between questioned pairs of fingerprints, within the US and the UK a similarity judgement relies on the internal criterion of the individual examiner who needs to consider whether the two prints are “more similar than any other close nonmatch that they have encountered” (Busey & Parada, 2010, p. 156). This cognitive process suggests the development of expertise resulting from exposure to a wide range of different fingerprint exemplars, thus allowing categorization of novel prints based on previously encountered examples (Thompson et al., 2014). Fingerprint expertise is believed to be domain-specific and does not therefore translate to higher accuracy when processing unfamiliar categories such as inverted faces, suggesting it is constrained by prior experience within a particular class of stimuli (Searston & Tangen, 2017a).

An exemplar theory of categorization was tested in a series of experiments in which it was predicted that examiners would draw on their experience to discriminate between fingerprints, even when the information contained within them was restricted (Thompson et al., 2014). In a same-or-different identification task using upright and inverted fingerprints in artificial noise (Experiment 1), fingerprint examiners were more accurate than novices (*M* = 87.2% vs *M* = 71.9%) with matching, non-matching, and similar-non-matching pairs, regardless of orientation. Examiners were also more accurate than novices in identifying whether a target print, displayed for five seconds, was the same or different to a test print presented after a five second delay (Experiment 2), and when reporting a same or different identification for fingerprint pairs displayed for two seconds or sixty seconds (Experiment 4). The researchers suggest the high performance of experts with fingerprints containing limited information can be attributed to an element of non-analytical and instinctive processing arising from familiarity with the stimuli class.

Although fingerprint examiners report adherence to the ACE methodology during comparison (Stevenage & Pitfield, 2016), anecdotally they describe an intuitive comparison decision followed by retrospective justification, with ACE as a method of evidencing their evaluation (Hall, personal communication). Thus, if examiners can accurately report prints as an individualization or an exclusion, and a component of expertise is reliant upon non-analytical processing, it may be infeasible and difficult for an expert to explain how an identification decision has been reached (Thompson et al., 2014).

1.5.1. Visual Processing by Fingerprint Examiners

Studies have shown that the experts demonstrate a clear accuracy advantage over novices when examining the same sets of fingerprints, and researchers have considered whether their expertise can be accounted for by differences in their visual processing of fingerprint images.

Configural Processing. Within the face perception literature, configural processing refers to the observer’s visual processing of the basic arrangement of facial features. This is important for determining that a face is being perceived, with variations in spacing and the position of the facial features allowing the observer to discriminate between individual faces (for a review of terminology see Piepers & Robbins, 2012). During fingerprint examination, the arrangement of features, and their spatial relationship to other features, may also be relevant in determining whether or not these stimuli are from the same source. The use of relational or spatial information between fingerprint features may therefore reflect differences in configural processing, and this was explored in a comparison between experts and novices (Busey & Vanderkolk, 2005).

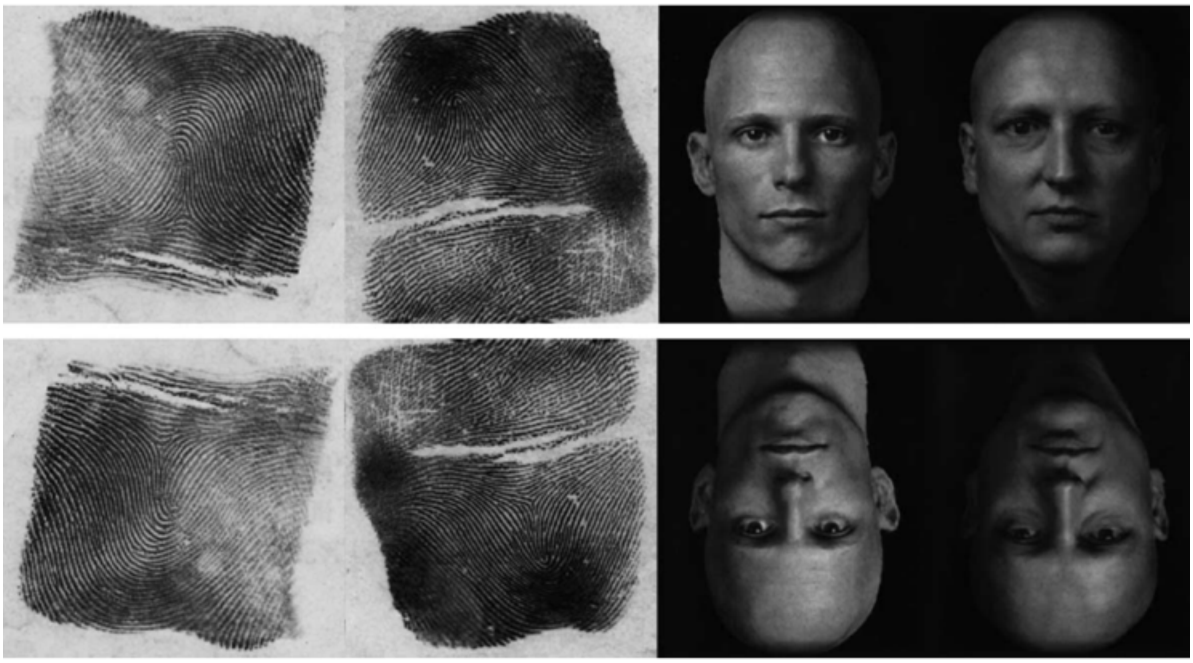
In Experiment 1, examiners and novices viewed a single target fingerprint fragment for one second. After a short (200ms) or long (5200ms) delay, two test fragments were presented and, if present, the target fragment needed to be identified. Fragments were either full or partially masked and were presented with and without artificial noise. Experts out-performed novices and were unaffected by delay, but their performance with partial images in noise was significantly lower than when viewing full images in noise. Busey et al. (2005) suggest this is evidence of configural processing, with partial prints in noise lacking sufficient contextual or relational information with which to make a reliable comparison.

In Experiment 2, EEG was used to record the brain activity of experts and novices while they identified upright and inverted fingerprints and faces as the same or different. Faces were included as they comprise a consistent arrangement of features (eyes above nose, nose below mouth) which is comparable to the arrangement of features within a fingerprint. Therefore, experts may exhibit similarities in their configural processing of both fingerprints and faces. Example stimuli are shown in Figure 1.8.

For both groups, accuracy was reduced when viewing inverted faces but not fingerprints, and behavioural data revealed no differences between experts and novices. EEG recording showed a delayed N170 response to inverted faces in both groups, which in previous research has been regarded as an indicator of disrupted configural processing of faces (e.g., Richler & Gauthier, 2014) and a marker of expertise (Gauthier & Tarr, 1997). With inverted fingerprints, the delayed N170 component was only apparent for examiners, thereby providing converging evidence of configural processing by experts and thus indicating a perceptual element to expertise.

Figure 1.8

Example Stimuli from Experiment 2 Showing Upright and Inverted Fingerprints and Faces

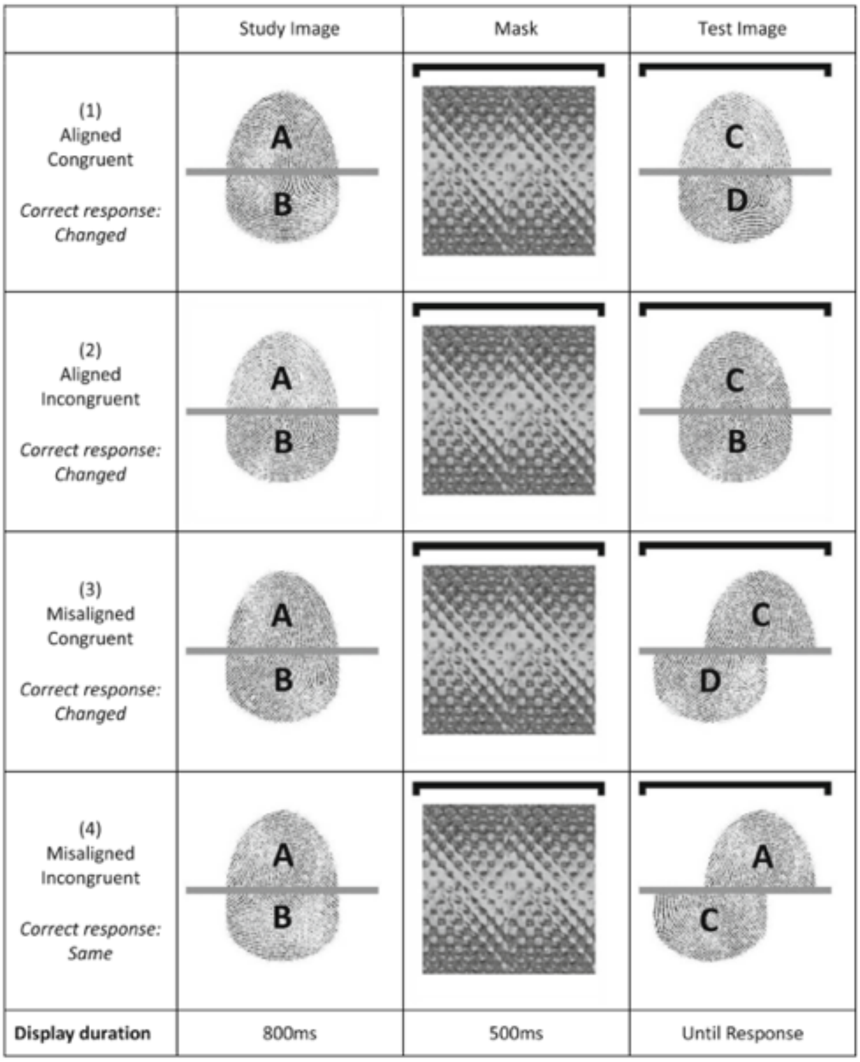


Holistic Processing. Within the face perception literature, the term ‘holistic’ typically refers to the visual processing of parts of a face as a whole rather than by its constituent parts (e.g., Farah et al., 1998; Maurer et al., 2002). This is regarded as an automatic process that develops through experience with the stimulus class (Richler & Gauthier, 2014; Wong & Gauthier, 2010). Examiner accuracy had previously been observed with full images rather than partial halves of the same image (Busey et al., 2005). This led researchers to consider whether expertise could be explained by holistic processes that integrate information from different regions of the prints (Vogelsang et al., 2017).

In a composite task (see example in Figure 1.9), examiners and novices were directed to match either the top or bottom half of a test fingerprint with a target print that had been presented 500ms previously. Test images were either aligned or misaligned (Experiment 1) or were inverted (Experiment 2). If observers were unable to ignore the part of the image that was not cued, it was predicted that this would interfere with, or facilitate, their processing of the test image. Misaligned and inverted images were used as they are believed to disrupt holistic processing.

Figure 1.9

Example Trial Sequence from Experiment 1 with Top Half of the Fingerprint Cued



	Study Image	Mask	Test Image
(1) Aligned Congruent *Correct response:* *Changed*	A B		C D
(2) Aligned Incongruent *Correct response:* *Changed*	A B		C B
(3) Misaligned Congruent *Correct response:* *Changed*	A B		C D
(4) Misaligned Incongruent *Correct response:* *Same*	A B		A C
Display duration	800ms	500ms	Until Response

Note. Capital letters on images are for explanation and were not on the trial images.

Overall, experts were more accurate than novices and there was no main effect of orientation. In composite and misaligned tasks, there were no differences in holistic processing between examiners and novices. With inverted images, expert accuracy declined with incongruent fingerprints but not when the images were misaligned, and between-group differences were only marginal. The authors concluded that this provided weak evidence in support of holistic processing by examiners who may use different perceptual processes to experts within other domains whose knowledge and experiences allows them to individuate or name objects (e.g., type of car, type of face, identity of person).

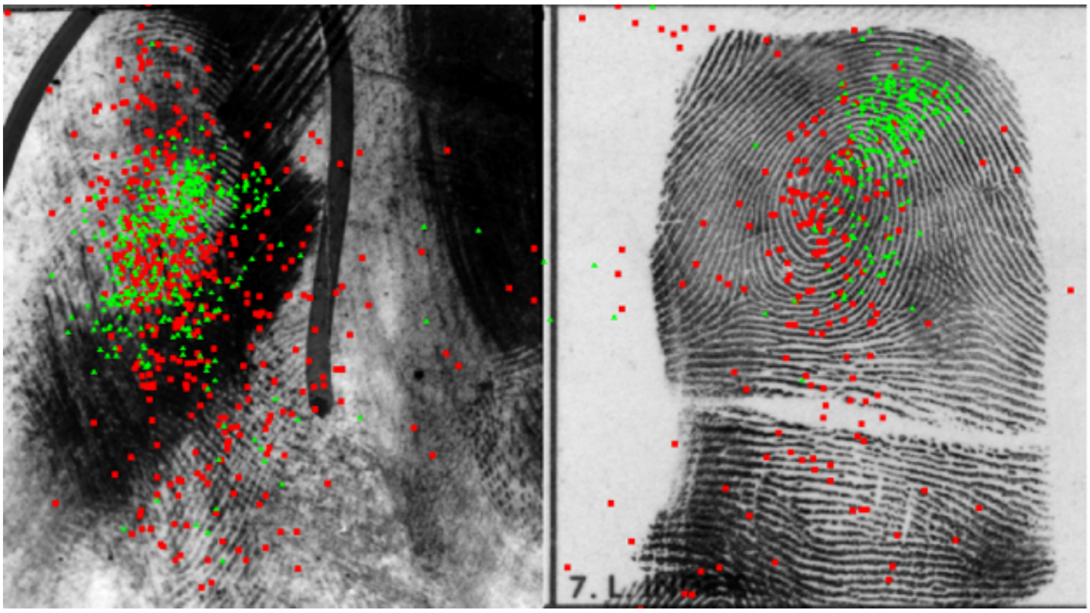
1.5.2. Eye Movements During Fingerprint Comparison

Fingerprint comparison is a visually demanding task, with examiners using eye movements to attend to those areas of an image containing relevant and highly diagnostic information. Eye movements (saccades) and fixations to visual stimuli are believed to be task-specific (Rayner, 1998) and purposeful (Henderson et al., 2005), and indicative of underlying cognitive processes (Henderson, 2003). As perceptual processes occur without conscious awareness and are difficult to verbalise (Yu et al., 2011), the key to understanding expertise may therefore lie in examining the gaze behaviour of experts and untrained observers during the comparison process.

To examine the viewing strategies of forensic fingerprint examiners, the eye movements of experts and novices were recorded whilst they compared latent and exemplar prints (see Figure 1.10) with viewing time restricted to twenty seconds (Experiment 2; Busey et al., 2011). Observers evaluated the prints as the same or different and were also able to provide a “too early to tell” rating. Although there were no between-group differences in the duration of fixations, examiners looked at the latent prints more than the exemplars, and their saccades were shorter when viewing both prints. Experts’ eye movements were also more consistent than novices and they appeared to fixate those areas that were likely to be informative whereas novices tended to focus on clearer, higher quality areas of the prints.

Figure 1.10

Examples of Expert and Novice Fixations When Viewing a Latent (left) and Exemplar Print (right)



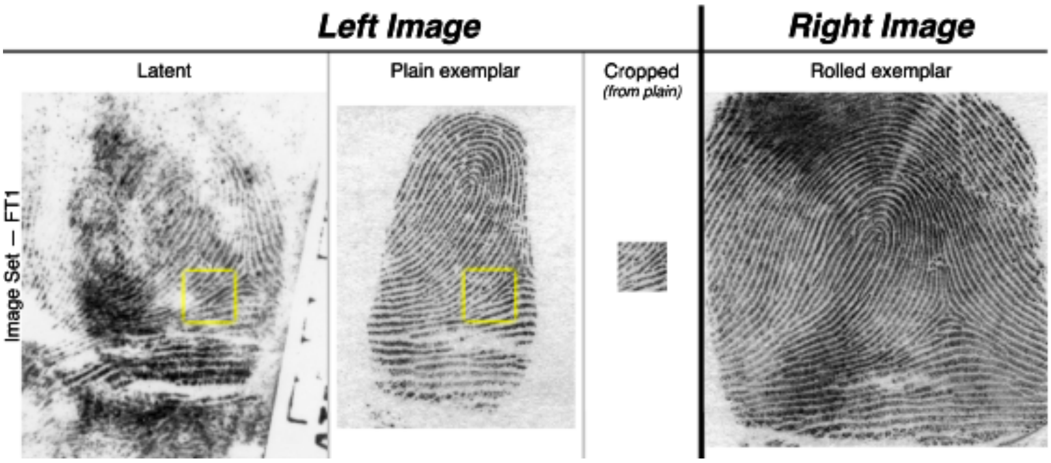
Note. Green areas are fixations by experts and red areas denote fixations by novices.

Overall, experts were more accurate than novices and made no false positive errors, although they were conservative in their decision-making with a high number of match trials rated as “too soon to tell”. The eye movements of examiners when viewing the latent print suggested an accumulation of information into working memory with which to make a comparison with the exemplar, with shorter saccades reflecting a purposeful evaluation of the images based on prior knowledge and experience (Busey et al., 2011) .

A further study considered the role of holistic and configural processing during target localization in fingerprint examination (Hicklin et al., 2019). This is the identification of relevant details within a latent print which are then located, or ‘localized’, on a corresponding area of an exemplar. Examiners viewed a target area outlined on a latent or plain print, or a cropped image containing only the target area, which they needed to locate on a corresponding exemplar. Examples of stimuli are shown in Figure 1.11.

Figure 1.11

Example of a Latent, Plain or Cropped Images (left images) Shown with a Mated Exemplar (right image).



Note. Yellow squares on the left image depict the target area that needs to be located on the right image.

When examiners viewed a *full* latent or *plain* exemplar, the target area on the matching rolled exemplar was accurately located within the first few fixations, with one to three seconds spent on each image before switching. However, when only a *cropped* area of the latent was available, examiners took longer to locate the target area in the exemplar, spent more time looking at each image before switching (two to six seconds) and failed to locate the target area in 6 out of 675 trials.

Across all trials, examiners spent longer looking at the exemplar, with repeated returns to the latent, perhaps reflecting short-term memory fade as the comparison progressed. The authors considered whether contextual information in the latent and plain images supported examiners' expectations about the location of the target area, or whether peripheral vision incorporated global information about ridge flow to guide the first fixation to the target print. They suggested comparison errors may arise from erroneous localizations, with the absence of localization when comparing non-matched pairs (exclusions) perhaps requiring a more holistic viewing strategy.

1.6. Fingerprint Examiner Recruitment

For some forensic comparison roles, self-selection based on perceived ability may be an important factor in the choice of career. For example, a person may be aware of a skill in recalling faces seen only briefly, perhaps many years previously, and this leads them to seek out suitable roles in forensic facial examination (Noyes et al., 2017). In relation to fingerprint examination, it seems highly unlikely that someone will have the opportunity to practice this skill in everyday life, and career choice may therefore be based on perceptions of job security or an interest in crime rather than ability. Technological advances mean job vacancies in forensic examination roles are now exposed to a much wider pool of potential applicants, with the accompanying risk that large numbers of applications may be from highly unsuitable candidates (NAS, 2017). Therefore, methods of identifying those candidates with the potential to be forensic examiners are likely to become increasingly important (Bécue et al., 2020).

There is currently no standardised selection tool to assist fingerprint laboratories in identifying applicants with the potential to be proficient examiners (NAS, 2017), and detailed guidance surrounding the type of aptitude tests that may benefit recruitment is lacking (e.g., SWGFAST, 2013). A recent workshop, attended by forensic examiners, researchers, psychologists, scientists, and members of the criminal justice community, considered the issues surrounding the selection of personnel within the pattern evidence domain of forensic science more broadly (NAS, 2017). Pattern recognition was identified as the key aptitude required in potential examiners, together with cognitive skills in learning, retaining, and recalling information, an ability to focus attention for long periods of time, sound decision-making, and good vision. The workshop concluded that the forensic community needed assistance in validating the relevance of visual acuity and cognitive ability in forensic examination, together with the development of suitable tools to aide selection. They suggested that evidence from tests for form-blindness, a condition whereby observers are unable to discriminate shapes, curves, and angles, as a method of predicting performance in fingerprint examiners was promising (e.g., Byrd & Bertram, 2003; Osborn, 1939), and indicated the feasibility of producing scientifically valid selection tools to aide recruitment.

1.6.1. A Solution for Fingerprint Examiner Recruitment

Although novices can be taught how to compare fingerprints and show a benefit of training in terms of improved accuracy (e.g., Stevenage et al., 2016b), there is a paucity of research to examine whether individual differences in fingerprint examiners' ability supports the development of expertise. Individual differences are believed to have a genetic basis (e.g., Wilmer et al., 2010) and to be stable over time, reflecting a level of ability in a specific cognitive process that remains consistent despite variations or novelty of the task (Ackerman, 1987). For example, the identification of individual differences in the face processing domain has identified a group of observers known as “super-recognisers” (Russell et al., 2009). Their above-average performance in face perception is believed to reflect stable individual differences in ability (e.g., Bobak et al., 2016), and has led to their deployment by the Metropolitan police in identifying perpetrators from CCTV images and photographs (Davis et al., 2016).

One study to consider individual differences in fingerprint comparison used a longitudinal design, comparing data from fingerprint trainees in three monthly intervals as they progressed through their training (Searston & Tangen, 2017c). Participants undertook the same four tasks in each testing session. These comprised a visual search task using fingerprints and faces, and a fingerprint matching task, from Tangen et al. (2011), a speeded fingerprint matching task, and a person-to-fingerprint matching task taken from Searston and Tangen (2017a)

Searston et al. (2017c) found an improvement in performance across all tasks within a twelve-month period, with highest performance gains between the initial test and re-test at three months. In visual search and fingerprint matching tasks, high correlations between trainees’ performance in the initial and subsequent tests suggested individual differences in ability that were stable and predictive of later performance. In speeded matching tasks, initial performance was significantly related to ability at three months but not in subsequent months, and there was no relationship between initial and subsequent tests in person-to-fingerprint matching tasks. Overall, the findings suggest that some of the variation in perceptual ability may be due to individual differences, with the best performing trainees displaying consistently high performance throughout the duration of the study. Identifying those applicants with demonstrably superior visual skills is likely to be highly relevant when recruiting personnel to these roles.

1.7. Conclusion

This chapter has examined key studies in relation to the methodology and accuracy of fingerprint examination, and the nature of expertise in practicing examiners. The current literature shows that examiners are able to match fingerprints with a high degree of accuracy within experimental settings (e.g., Tangen et al., 2011; Thompson et al., 2014; Stevenage et al., 2016a), albeit with considerable inter- and intra-individual differences both in their performance (Ulery et al., 2011, 2012) and the information used to reach a comparison outcome (Dror et al., 2010; Ulery et al., 2014, 2016). Expertise may be partly accounted for by the ability of examiners to differentiate minute differences in pairs of fingerprints that are indistinguishable to novices (Tangen et al., 2011; Thompson et al., 2014). The ability to rapidly locate target areas in an exemplar (Hicklin et al., 2019; Yu et al., 2011), combined with high accuracy when matching low-quality prints presented only briefly, provides converging evidence that expertise may also rely on an element of non-analytical processing of fingerprints (Thompson et al., 2014). Research surrounding the cognitive processes underlying expertise is currently limited (PCAST, 2016), and the perceptual skills required for effective fingerprint comparison are alluded to rather than established (NAS, 2017).

One's own expertise in fingerprint comparison may be a difficult concept for a practitioner to explain (Thompson et al., 2014; Yu et al., 2011), and eye movement data has provided some evidence of different perceptual processes of examiners in the absence of a verbal account (Busey et al., 2005, 2011; Vogelsang, 2017). Further research to examine the cognitive processes of examiners may lend support to theories of non-analytical processing of fingerprints and this could explain some of the inconsistencies observed between examiners. In addition, the study of individual differences in performance is becoming increasingly important, not only in the recruitment of suitably skilled examiners (Searston et al., 2017b), but in providing insight into the cognitive factors that underpin fingerprint examination.

1.8. The Structure of this Thesis

The purpose of this thesis is to investigate the nature of expertise in forensic fingerprint examiners. Existing research has provided evidence of an accuracy advantage for professional examiners in fingerprint comparison, notably when compared with the performance of untrained observers (e.g., Stevenage et al., 2016a, 2016b; Tangen et al., 2011; Thompson et al., 2014). However, studies have also identified inconsistencies in the examination procedures used by examiners, with a wide range of individual differences in their identification decisions and approaches to fingerprint comparison (e.g., Dror et al., 2010; Ulery et al., 2011, 2012, 2014, 2016). A key argument developed in this thesis is that these individual differences are of paramount important for understanding fingerprint comparison expertise, for several reasons.

Firstly, the ‘expert’ status awarded to forensic fingerprint examiners in the criminal justice system should reflect an ability in their field of expertise which is not found in the population at large. However, research findings clearly show that fingerprint examiners are not a homogenous cohort of people, and their group advantage in fingerprint comparison may not necessarily be a true reflection of performance by the individuals comprising the group (see Ackerman, 1987; Davis et al., 2016; Russell et al., 2009, for examples of individual differences in performance in cognitive tasks). Accordingly, *any* court-practicing fingerprint examiner should not find that they are outperformed by a naïve observer in a fingerprint comparison task. Given the evidential weight afforded to an examiner’s subjective identification decision this is an important consideration, and one which this thesis will aim to explore with the creation and data from a novel fingerprint aptitude test.

A further aspect of expertise, and one which has not so far been studied to any degree, is the link between fingerprint comparison ability and cognitive skill. Fingerprints are complex stimuli and, given the few opportunities to examine fingerprints in everyday life, are unfamiliar to all but professional fingerprint examiners. It may therefore be expected that people who routinely undertake fingerprint comparison may have acquired advanced cognitive and perceptual skills in line with their accumulation of expert knowledge of fingerprints. Conversely, there may be observers within the general population who already possess these superior cognitive skills, and who may be naturally better suited to fingerprint comparison roles. To gain a greater understanding of the nature of expertise in fingerprint comparison, this thesis will explore which cognitive abilities underpin accuracy in fingerprint matching tasks.

The first experimental chapter describes the creation of an online fingerprint aptitude test, designed for participation by untrained observers and forensic fingerprint examiners to facilitate a direct comparison between these groups. The primary intention in creating a new test was to incorporate a wide range of fingerprint stimuli within one test. Previous studies have used fingerprints in a single format, either inked or rolled prints, simulated crime scene prints, partial prints, or latent prints (see, e.g., Searston & Tangen, 2017a, 2017b, Stevenage et al., 2017; Tangen et al, 2011; Thompson et al., 2014). However, to advance our understanding of the role of expertise in fingerprint comparison it is important to determine the level at which different groups of observers perform in fingerprint comparison tasks. Incorporating fingerprint stimuli that vary in format and complexity allows this comparison to be made. The varied stimuli for the fingerprint aptitude test were therefore comprised of inked and rolled impressions, partial and whole palm prints, and latent prints. These were provided by volunteer donors and were obtained and processed by an experienced fingerprint examiner to ensure that the quality and level of difficulty of the fingerprints was appropriately graded. Untrained observers (‘novices’) undertook the fingerprint aptitude test on two occasions, separated by several days. This chapter reports the data from these novices and demonstrates that the test was reliable across time. Using different stimuli in each of the four blocks of the test was reflected in differences in accuracy, thereby indicating that this test was suitable for participation by a wider pool of observers.

Chapter 3 explores the performance of experienced forensic fingerprint examiners (‘Experts’), trainee examiners (‘Trainees’), and novice observers (‘Novices’) when undertaking the fingerprint aptitude test. The aim was firstly to determine the level at which these three groups perform in a range of fingerprint comparison tasks by focusing on quantitative differences between these groups. Secondly, individual differences in fingerprint comparison accuracy were explored by comparing the performance of each Novice with that of the Expert and Trainee groups. The findings are considered in relation to the nature of expertise in forensic fingerprint examiners, and whether these professional observers possess fingerprint comparison abilities that are beyond the remit of the broader population.

Fingerprint comparison expertise is further explored in Chapter 4. The aim of this chapter was to understand the cognitive abilities of Experts, Trainees, and Novices, and to identify whether these were associated with experience in fingerprint comparison or were present in the wider population. This chapter uses data from the fingerprint aptitude test in Chapter 3 to consider the relationship between cognitive ability and fingerprint comparison accuracy. Observers first undertook a battery of perceptual tests designed to reflect the cognitive processes that are likely to be engaged when undertaking fingerprint comparison. Thus, these were intended to capture differences in visual short-term memory, visual search, mental rotation, feature comparison, and perceptual attention. As Trainees and Experts were participating in their professional capacity of Forensic Fingerprint Examiners, a measure of intrinsic motivation was also included to explore the relationship between motivation and test performance. Accuracy data from the fingerprint test and the test battery was then entered into a series of regression analyses to explore the relationship between cognitive ability in non-fingerprint tasks and performance in fingerprint comparison.

In the final chapter, the results from the three experimental chapters are considered in relation to the role of expertise in fingerprint comparison accuracy, and the level at which untrained observers perform in these complex visual tasks. The nature of the cognitive skills associated with fingerprint comparison ability is also discussed, along with emerging evidence that performance in these fingerprint tasks may be underpinned by a domain-general cognitive ability. Implications for the recruitment of suitable applicants to fingerprint comparison roles are considered in light of these findings.

Chapter 2

A Fingerprint Aptitude Test

2.1 Introduction

Forensic fingerprint comparison requires the observer to compare sets of fingerprints to identify whether they were produced by the same person or are from different people, often with the aim of identifying the source of a finger mark left at the scene of a crime (FSR, 2017). This task is performed routinely by fingerprint examiners whose experience and training allows them to provide expert testimony as to the source of fingerprints within courts of law. Fingerprints are visually complex stimuli comprising an identifiable pattern of lines and features, interwoven with minutiae such as ridge endings, bifurcations, creases, or pores (Jain et al., 2006). It is through close examination of fingerprint features, and the spatial relations between them, that qualified fingerprint examiners can provide expert testimony as to the source of marks found at crime scenes (CPS, 2019).

The identification of a person by their fingerprints is an established form of forensic comparison dating back to the beginning of the last century (Barnes et al., 2011). However, reviews by the National Academy of Sciences (NAS, 2009) and the President's Council of Advisors on Science and Technology (PCAST, 2016) have questioned whether fingerprint examiner testimony has scientific validity. PCAST reported that many forensic fingerprint examiners are highly accurate in experimental conditions (Langenburg et al., 2012; Pancheo et al., 2014; Tangen et al., 2011; Vokey et al., 2009), but it is difficult to determine whether performance in lab-based studies translates to a reliable error rate within live casework as the ground truth as to the source of a fingerprint found within a crime scene is seldom known (Cole, 2008). Fingerprint examiners also demonstrate high accuracy when compared to novice observers (Searston & Tangen, 2017a; 2017b; Stevenage & Pitfield, 2016b; Tangen et al., 2011; Thompson et al., 2014), although these studies also show that the conclusions of professional examiners are not error-free.

Experiments to compare the performance of professional fingerprint examiners and untrained observers have used a variety of fingerprint stimuli, such as simulated crime scene prints (Tangen et al., 2011), genuine crime scene prints (Thompson et al., 2014), high- and low-quality fingerprints from a database (Stevenage et al., 2017), and cropped rolled impressions (Searston & Tangen, 2017a, 2017b). However, as each study adopted a different experimental design, and tested different sample populations, this makes it difficult to draw comparisons between studies and to conclude *how* fingerprint examiners are better than novices in fingerprint tasks.

A further consideration is whether the fingerprint stimuli used in previous studies effectively captured differences in comparison performance between professional fingerprint examiners and untrained observers. It seems likely that the professional fingerprint examiners tested would have outperformed novice observers in fingerprint comparison tasks by virtue of their workplace experience and training. However, the use of single format stimuli does not indicate the level at which fingerprint examiners might excel in these tasks. Incorporating a range of fingerprint stimuli into a single test would identify the comparison tasks that differentiate the abilities of professional fingerprint examiners and untrained observers.

For these reasons, a new fingerprint aptitude test with varied stimuli needed to be created. As new sets of stimuli were required for this research, a fingerprint consultant employed by a large UK police service collaborated in the creation of this test. They were able to draw on their extensive experience to obtain sets of fingerprints and simulated crime scene latent marks from volunteers for incorporation into the test. Importantly, this ensured the ground truth of the source of any latent marks was known. This also allowed the quality of each latent fingerprint and finger mark to be professionally graded for clarity, complexity, and suitability for inclusion in the test. The resulting test comprised of four blocks of different stimuli, designed to capture fingerprint comparison performance in pattern matching, fingerprint image matching, palmprint matching, and latent print matching. The requirement to conduct remote testing during the COVID-19 pandemic also necessitated that this would need to be suitable for online participation.

A further important consideration was the suitability of this test to measure fingerprint comparison aptitude in untrained observers as there needed to be a realistic prospect that novice observers could provide the correct response. There is undoubtedly an awareness of fingerprints within the general population, although non-experts do not have a detailed knowledge of the constituent features of fingerprints, patterns, and minutiae that forensically trained observers hold. Nonetheless, previous research has shown the feasibility of measuring fingerprint comparison accuracy in observers with no prior training (Searston et al., 2017; Tangen et al., 2011; Thompson et al., 2014) or with only limited pre-test training (Stevenage et al., 2016). Therefore, to ensure that performance could be compared directly across groups varying in fingerprint expertise in subsequent studies, observers were required to match patterns and fingerprints in this test without any of the identification and magnification aids of the type deployed by experts.

This chapter describes the creation of this fingerprint aptitude test and its validation, which was assessed by testing untrained observers on two occasions. The analyses focus on identification accuracy in pattern, fingerprint image, palmprint, and latent print matching, the internal consistency of the test in terms of its ability to measure fingerprint comparison aptitude, and test-retest reliability.

2.2 Test Construction

This test was constructed in four blocks and created for online participation using Qualtrics software. The content of each block was designed to reflect a specific skill in fingerprint comparison, with varying degrees of difficulty. Each block of trials was preceded by task-specific instructions and an example trial with correct answer was always shown. Only one trial was presented on each page and was visible until a keyboard response was entered. Directly below each test array, participants were asked to rate their level of confidence in their response being correct from 1 (Not Confident) to 5 (Extremely Confident). All trials required the participant to match a target to a sample array or grid, or to identify that the target did not match any of the exemplars. The same order of stimulus presentation was maintained for participants in all trials, and the blocks were always completed in numerical order.

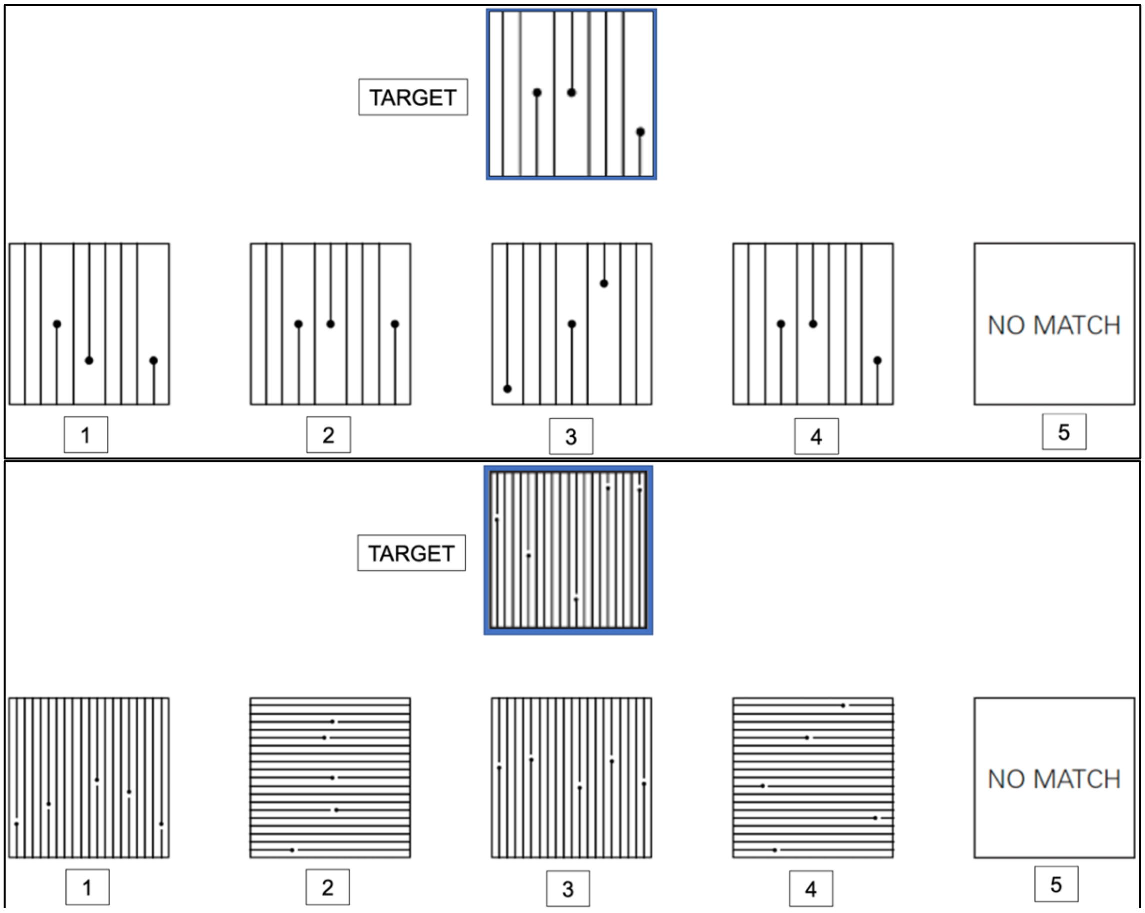
Block 1 – Visual Pattern Matching

This block comprised of fifteen trials in which participants were presented with a single target above an array of four exemplars and a box labelled ‘No Match’ (see Figure 2.1). Stimuli were manufactured black on white line drawings which were designed to measure simple pattern matching ability, analogous to observing, interpreting, and comparing characteristic features such as creases and minutiae during fingerprint comparison.

Some of the matching images were rotated versions of the target thereby reflecting the need to mentally rotate images during fingerprint comparisons. The target was absent in two of the fifteen trials, and in the remaining trials the location of the target in the array was presented in a pre-set order to ensure it appeared equally in each position throughout the block. Each target and array measured 1000(w) x 397(h) pixels in total at a screen resolution of 127 ppi and was presented in the centre of the screen.

Figure 2.1

Example of Target-Present (top) and Target-Absent Trials (bottom) From Visual Pattern Matching (Block 1).

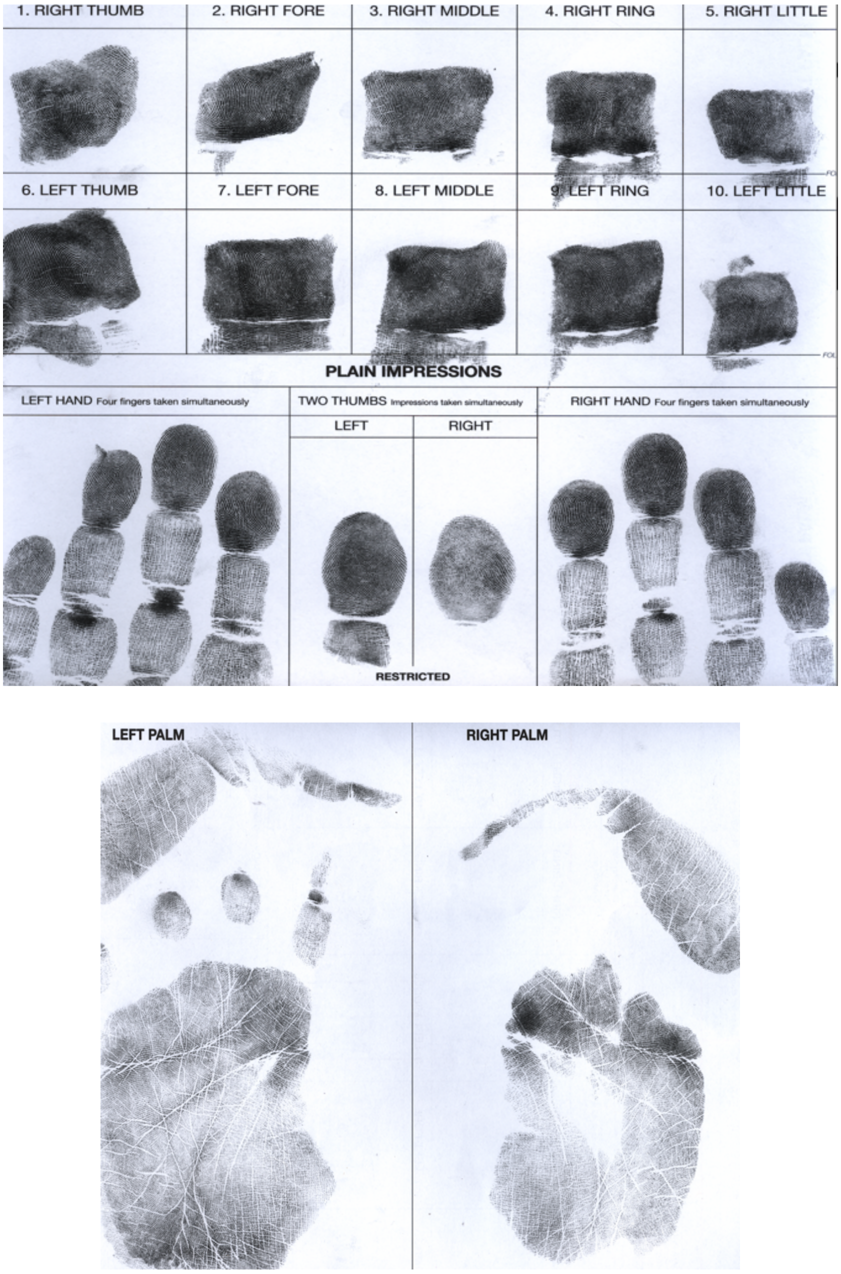


Block 2 – Fingerprint Image Matching

Fingerprint and palm images for use in Blocks 2 and 3 were created specifically for this research. They were taken from sets of rolled and plain impressions on standard police Tenprint forms supplied by known volunteer donors (see Figure 2.2 for an example). Clear images were selected to ensure they contained sufficient information with which to make a comparison, and the quality of all images was verified by a Fingerprint Consultant employed by a large UK police service. This block contained twenty trials and followed the same format as Block 1, with a single target image presented above an array of four exemplars and a box labelled ‘No Match’. The target and array measured 1000(w) x 562(h) pixels at a screen resolution of 127 ppi.

Figure 2.2

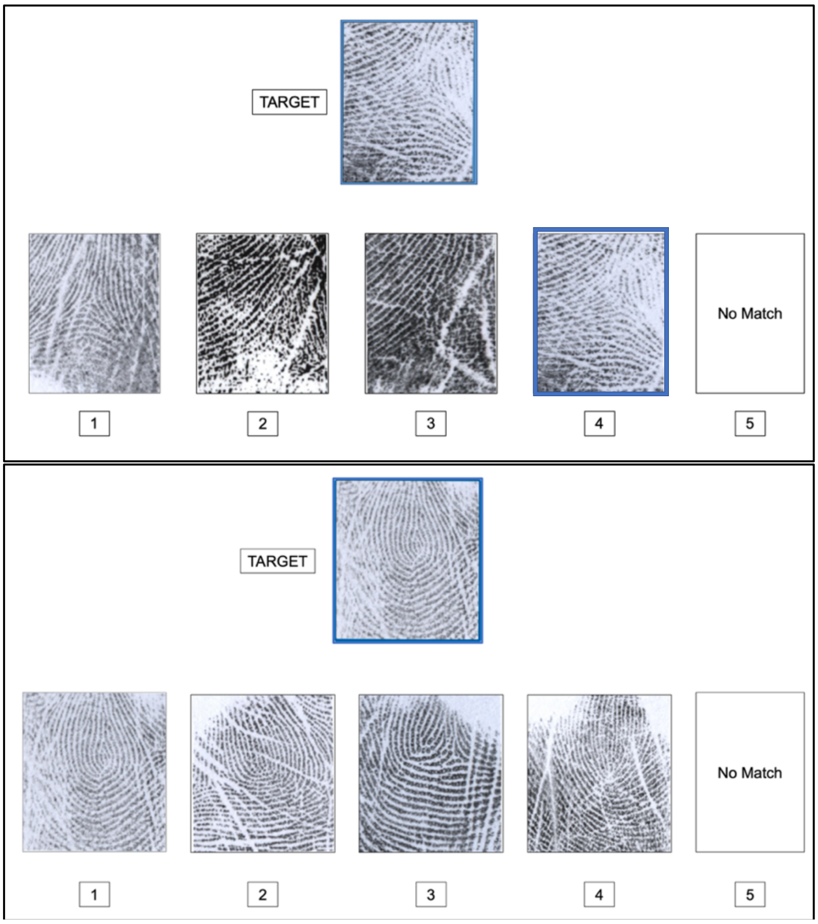
Example of Tenprints:Rolled and Plain Impressions (top) and Palmprints (bottom).



The target was either a rolled fingerprint or a cropped section from a palmprint. In target present trials, the matching counterpart was presented along similar exemplars. The target was absent in six of the twenty trials and the exemplars bore a close similarity to the target. Similarity was established by the police fingerprint consultant who created the stimuli for the test and was based on their extensive experience in fingerprint examination. The location of the target was pre-set across target present trials, to ensure that it appeared equally in each array position. The first fifteen trials measured basic image recognition, and in target present trials the target and matching counterpart were *identical* prints (Figure 2.3).

Figure 2.3

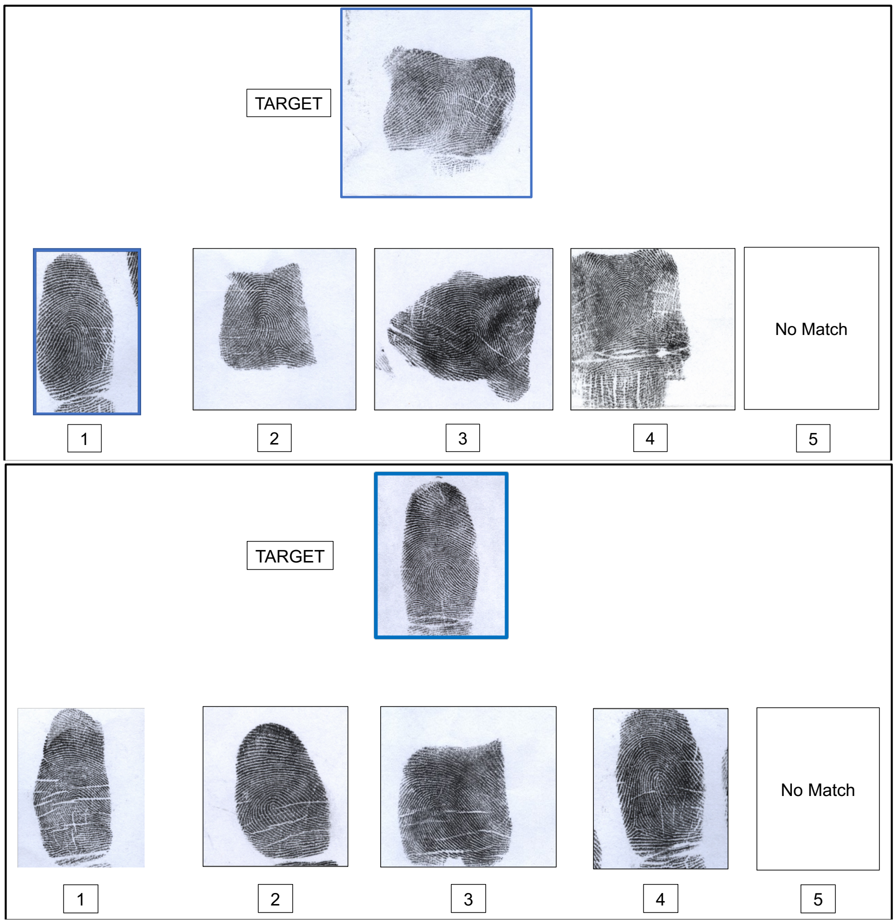
Example of Target-Present (top) and Target-Absent (bottom) Stimuli from Trials 1 to 15.



Although this task may appear relatively simple to complete, experienced fingerprint examiners advise that this can identify observers with low visual acuity, or a tendency towards inattention to task instructions or stimuli. For the final five trials the target was either a plain impression and its matching counterpart a rolled impression, or vice versa (Figure 2.4). In contrast to the first fifteen trials, the matching fingerprints in target present trials were therefore *non identical* images. Observers were instructed to pay close attention to the features of the target print in these latter trials rather than trying to locate an identical image.

Figure 2.4

Examples of Target-Present (top) and Target-Absent (bottom) Stimuli from Trials 16 to 20.

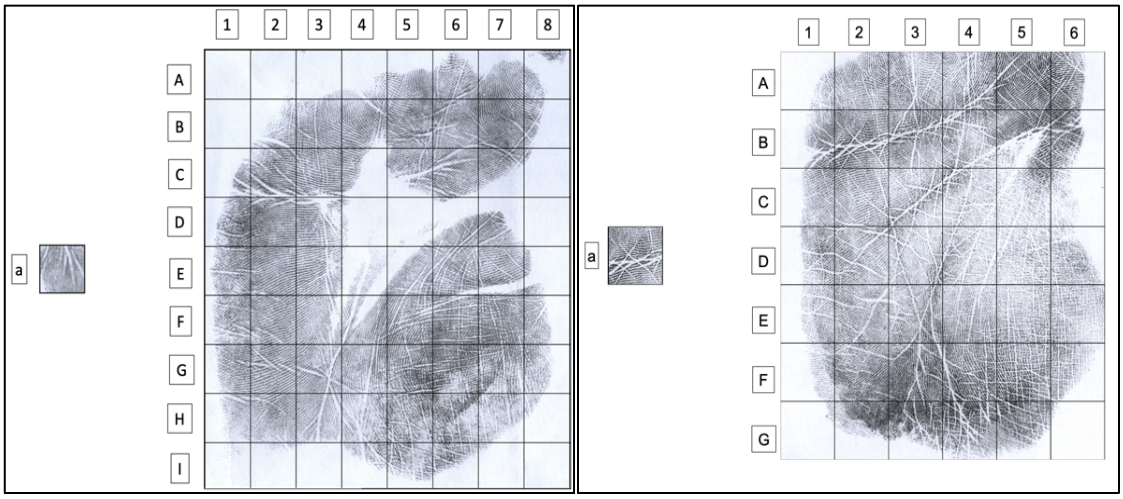


Block 3 – Palmprint Identification

In this block, the target was a cropped section of palmprint that needed to be either located on, or excluded from, a complete palmprint. In the first ten trials, an 8 x 9 grid was overlaid onto a palmprint, annotated with numbers and letters to allow identification of the target area, with the same image used in each trial. A different palmprint was used in the final ten trials, overlaid with a 6 x 7 grid annotated with letters and numbers. An example of the stimuli is shown in Figure 2.5.

Figure 2.5

Example of a Target with Grid Overlay on a Complete Palm, with Trials 1-10 (Left) and Trials 11-20 (Right).



Therefore, as the size of the palmprint was kept constant across trials, grids in the first ten trials were smaller than those in the final ten trials, with the different sizes likely to reflect differences in task difficulty. Perhaps surprisingly, experienced examiners suggested it would be easier to locate the target when smaller grids were applied. For this reason, these were presented in the first half of this block.

The target was presented to the left of the complete palm (annotated with a small letter for trial identification), scaled to the exact size of the overlaid grid, with one target per page. There were twenty trials in total. For each of the two palmprints, four of the targets were in their correct orientation, two were oriented to 180 degrees, one was oriented 90 degrees to the right, and one was oriented 90 degrees to the left. In two trials, the target was not present on the complete palmprint. The total size of the target and comparison palmprint as displayed on screen was 700(w) x 555(h) pixels at a screen resolution of 127 ppi.

Block 4 – Latent Fingerprint Matching

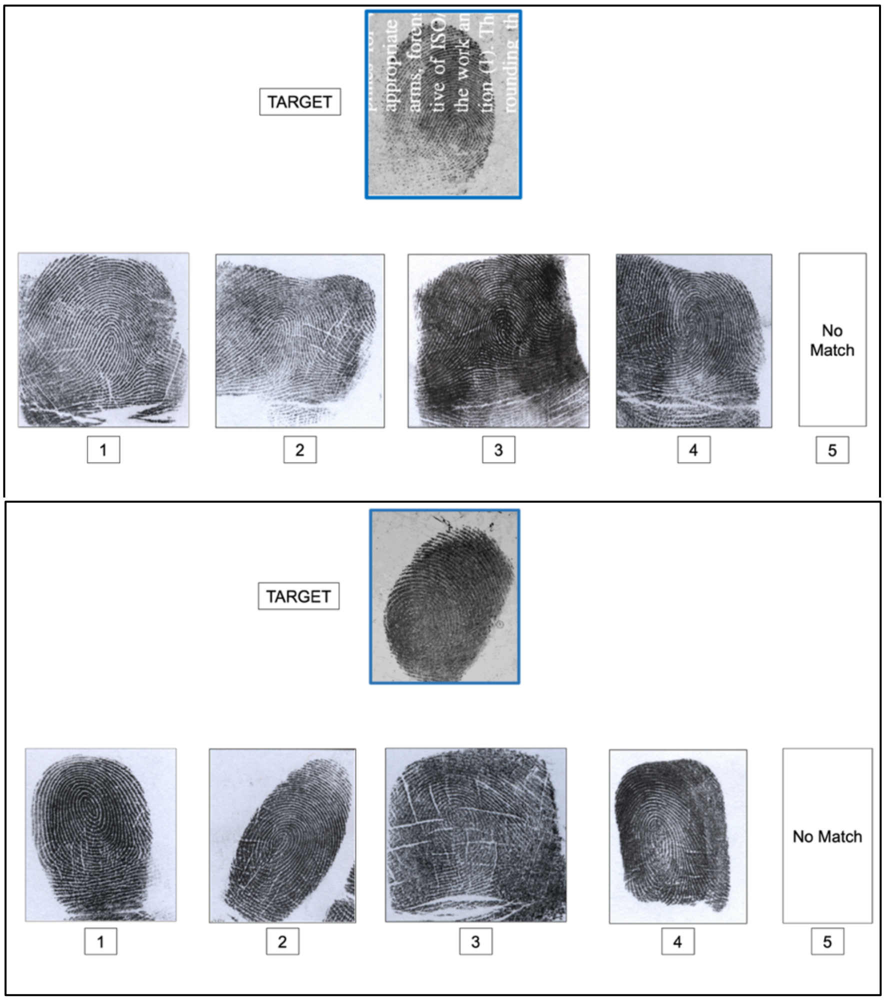
This section was designed to reflect the more challenging fingerprint comparisons undertaken by examiners. To facilitate this, latent print stimuli were created for incorporation into the trials, of the type left when an individual touches with a surface. The same volunteer donors who provided the Tenprint sets of impressions for Blocks 2 and 3 also handled glass, paper, and plastic surfaces to provide latent prints. These were developed by an experienced fingerprint examiner using standardised techniques.

To be more representative of casework, some of the fingerprint images were distorted by movement of the finger during the making of the mark or had substrate or background interference. Others were rotated to be incongruent with the target, and one was shown in the reverse direction. The exemplars and matching impressions used to compile the arrays were from the Tenprint sets previously referred to. Some additional fillers depicting arch impressions were obtained from the NIST Special Database 302 (Fiumara et al., 2007) as this particular fingerprint pattern was not well represented in the Tenprints obtained from the donors. Example stimuli are shown in Figure 2.6.

In this final block the test format matched that of Blocks 1 and 2, with the target presented above an array of four exemplars together with a box labelled ‘No Match’. There were twenty trials, with the target displayed in a pre-set order to ensure it appeared equally in each position. In four trials the target was absent from the array. Each target and array measured 1000(w) x 563(h) pixels at a screen resolution of 127 ppi. As entire impressions needed to be retained for the purpose of testing fingerprint comparison ability, the sizes of the images in the arrays were irregular, however, the visible detail in each impression was comparable across images.

Figure 2.6

Example of Target-Present (top) and Target-Absent (bottom) Trials.



2.3 Method

Participants

Thirty students (2 male) from the University of Kent, with a mean age of 19.6 years (*SD* = 3.9, range = 18 to 40 years), participated in this study in return for course credits. The only requirement for participation was normal or corrected-to-normal eyesight. The research was conducted in line with ethical guidance issued by the British Psychological Society and was approved by the School of Psychology Ethics Committee at the University of Kent.

Procedure

This research was conducted online using Qualtrics software. Participants took the same Fingerprint Aptitude Test on two occasions, separated by a minimum gap of at least seven days (mean Time 1 and Time 2 interval = 11.2 days, *SD* = 3.1). The instructions and procedure were consistent on both occasions. Prior to commencing each test, participants were provided with instructions to calibrate their computer screen. This was done by placing a credit card against an onscreen template of a standard sized credit card (85.6 mm x 54.0 mm = 323.5 x 204.0 pixels) and adjusting the browser magnification until the card and template matched.

All responses were entered using the computer keyboard and participants needed to attempt all trials. Responses to any trials currently in view could be changed prior to entering the response, but previous answers could not be reviewed or amended. Task instructions emphasised accuracy over speed and participants were aware that each test may take up to two hours to complete. Onscreen instructions displayed between each block recommended to observers to take a short screen break prior to moving on to the next section.

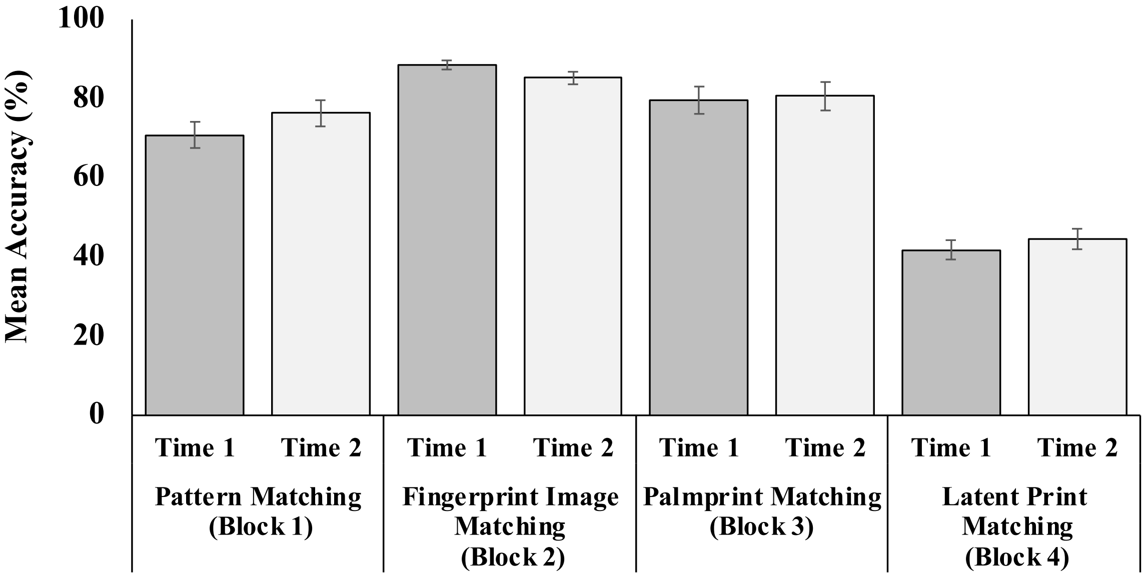
2.4 Results

Analysis of Accuracy Data

Participant Accuracy: Overall mean accuracy in both test sessions was similar for Time 1 (*M* = 70.00, *SE* = 2.02) and Time 2 (*M* = 71.60, *SE* = 2.02). To further compare performance across the two test sessions, mean percentage accuracy was calculated separately for each block (see Figure 2.7). This figure shows differences in accuracy between each block, with a noticeable decline in accuracy during latent fingerprint matching in the final block. Some smaller differences in accuracy were also apparent within each block at Time 1 and Time 2.

Figure 2.7

Mean Percentage Accuracy Scores in Each Block for the Test at Time 1 and Time 2.



Note. Error bars denote the standard error of the means.

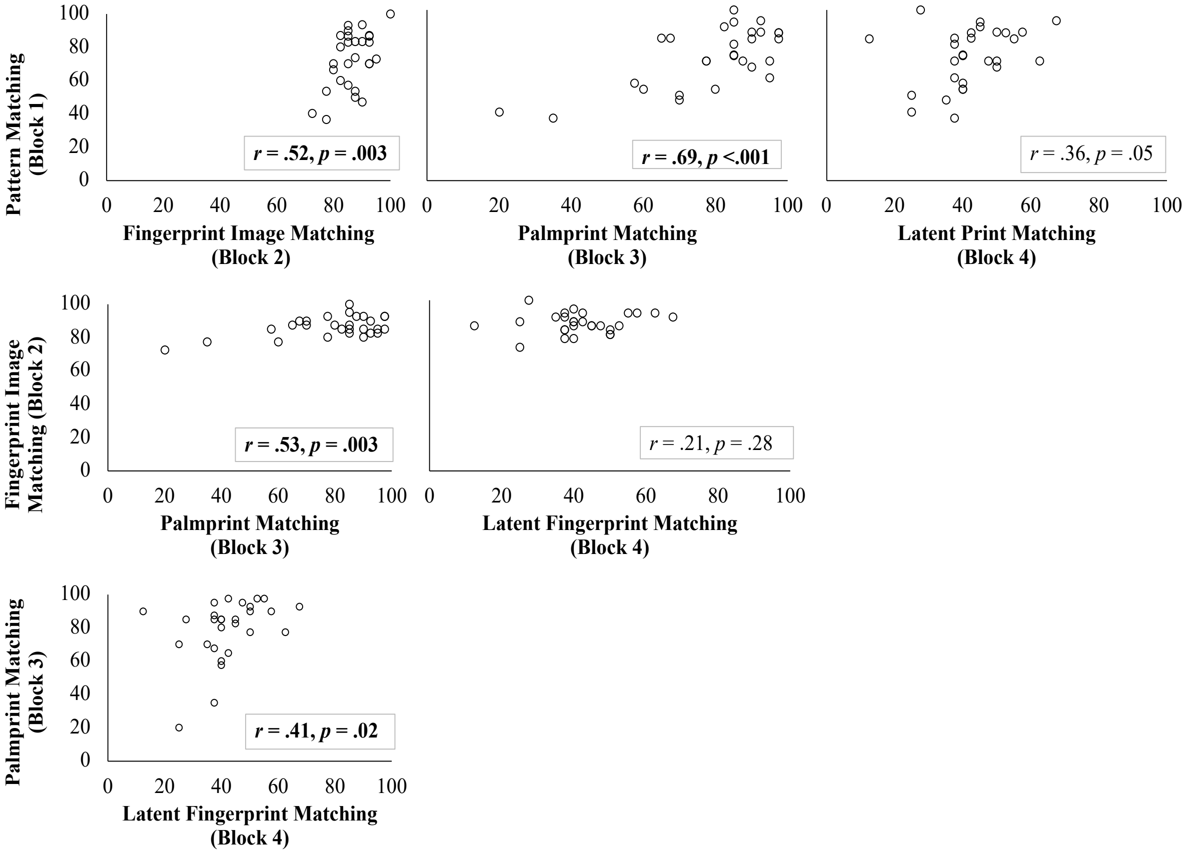
These differences in accuracy were explored with a 2 (Session: Time 1 or Time 2) x 4 (Block: 1, 2, 3, or 4) within-subjects ANOVA of the data. This revealed a main effect of Block, *F*(3, 174) = 141.81, *p* < .001, partial η^2 = 0.71. A series of paired *t* tests with Bonferroni correction showed that accuracy in Block 1 (*M* = 73.40, *SE* = 2.40) was lower than Block 2 (*M* = 86.80, *SE* = 0.94), *t*(58) = 6.06, *p* < .001, and Block 3 (*M* = 80.00, *SE* = 2.50), *t*(58) = 2.97, *p* = .03, and was higher than accuracy in Block 4 (*M* = 43.10, *SE* = 1.79), *t*(58) = 11.68, *p* < .001. Accuracy in Block 2 was higher than in Block 3, *t*(58) = 2.94, *p* = .003, and Block 4, *t* (58) = 22.05, *p* < .001. In Block 4 accuracy was lower than in Block 3, *t*(58) = 15.34, *p* < .001. There was no main effect of Session, *F*(1, 58) = 0.30, *p* = .60, partial η^2 = 0.01, and an interaction between Session and Block was not found, *F*(3, 174) = 1.28, *p* =.30, partial η^2 = 0.02.

In summary, there was no difference in overall accuracy, or accuracy within each block, between each test session. Analyses revealed differences in accuracy between all blocks of the test, thus reflecting the varied task demands of each block. Across the test, latent fingerprint comparison (Block 4) proved to be the most difficult task, with fingerprint image matching (Block 2) the easiest of the four blocks.

<u>Accuracy Correlations Across Blocks:</u> As analyses had revealed differences in accuracy between the four blocks of the test, the relationship between performance across blocks was then explored. There was no difference in accuracy within each block across Time 1 and Time 2, and mean percentage accuracy in each block was therefore calculated by combining data from both test sessions. Pearson correlation coefficients were computed using this data to assess the accuracy relationship between blocks of the test. Block-by-block correlations and *p* values are shown in Figure 2.8.

Figure 2.8

Correlations of Mean Percentage Accuracy Data Between Different Blocks of the Test.



Note. Significant Pearson correlation coefficients are shown in bold font.

In summary, these correlations show that accuracy in pattern matching (Block 1) was moderately associated with fingerprint image matching (Block 2), and palmprint matching (Block 3), with a weak correlation with latent print matching (Block 4). Accuracy in fingerprint image matching was moderately associated with similar accuracy in the more difficult task of palmprint matching but not latent print matching, and accuracy in palmprint matching was moderately related to performance in the more challenging latent print comparisons. Therefore, although accuracy was similar between some blocks of the test, the absence of correlations between fingerprint image matching and latent print matching, and the weak relationship between pattern matching and latent print matching, may reflect the dissociable aspects of these different fingerprint comparison tasks.

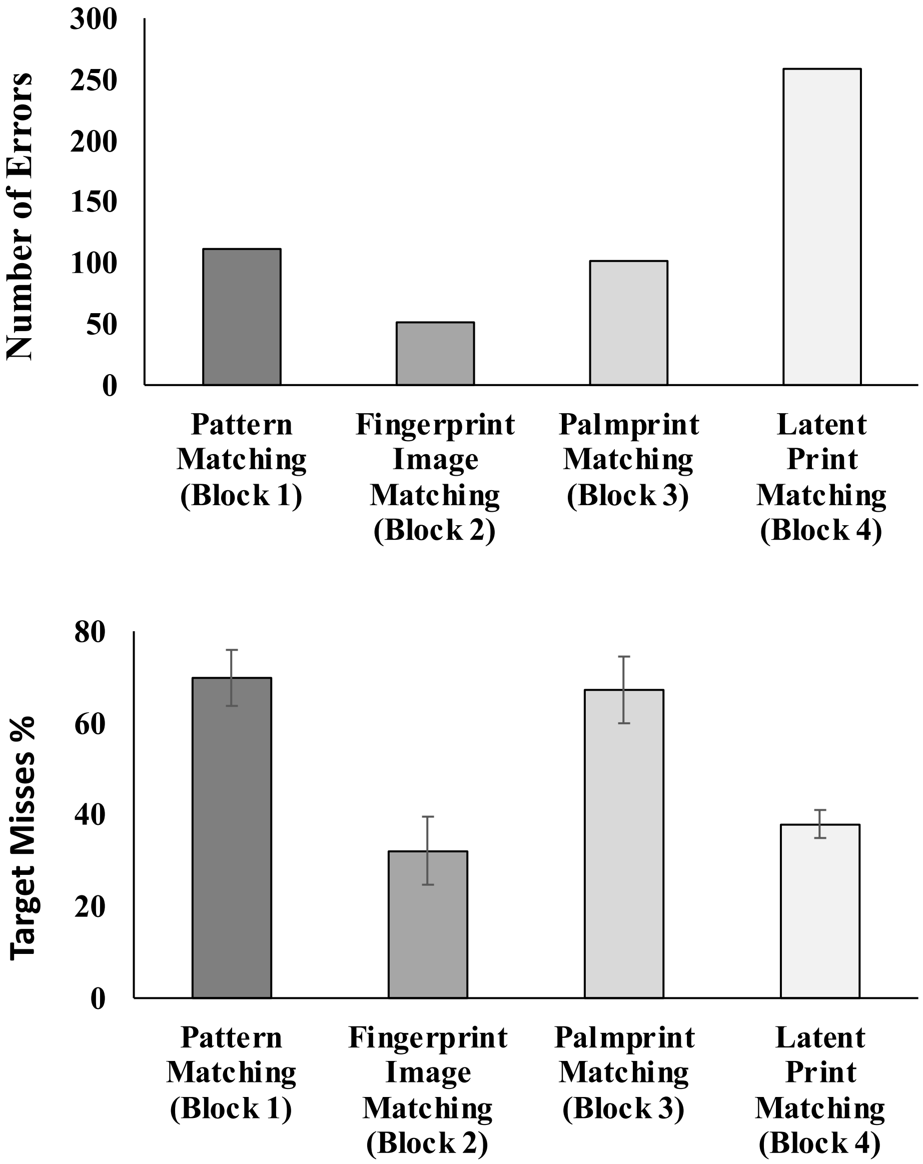
Errors in Target Present Trials: Observers entered a 'no match' decision if they could not locate the target in the array. In target-present trials these responses would therefore be regarded as *target misses*. To examine whether the target was missed in preference to erroneously identifying another exemplar, the number of errors made by observers in target-present trials in each block was first counted (see Figure 2.9). These data naturally reflect the accuracy data (Figure 2.7), with most errors committed during latent print matching in Block 4 and fewest errors during fingerprint image matching in Block 2.

The percentage of these trials in which the target was missed was then calculated for each observer and used to compute the mean percentage of targets missed within each block (see Figure 2.9). A one-way ANOVA of this data was then used to examine differences in the targets missed between each of the blocks of the test, $F(3, 116) = 10.20$, $p > .001$. This revealed that targets were missed more in Block 1 ($M = 70.00$, $SE = 6.17$) than Block 2 ($M = 31.10$, $SE = 7.40$), $t(116) = 4.41$, $p < .001$, or Block 4 ($M = 38.00$, $SE = 16.50$), $t(116) = 3.63$, $p = .002$. Targets were missed less in Block 2 than in Block 3 ($M = 67.40$, $SE = 7.29$), $t(116) = 4.13$, $p <.001$, and more in Block 3 than Block 4, $t(116) = 3.34$, $p =.006$. No other comparisons were significant, $p > .05$.

Therefore, in the first two blocks, the percentage of target misses was commensurate with the total number of errors: higher errors in Block 1 equated to higher target misses, and lower errors in Block 2 equated to lower target misses. Conversely, there were fewer target misses associated with higher errors in Block 4, and more target misses with fewer errors in Block 3.

Figure 2.9

Errors in Target Present Trials in Each Block (top) and Percentage of Target Misses in Each Block (bottom).



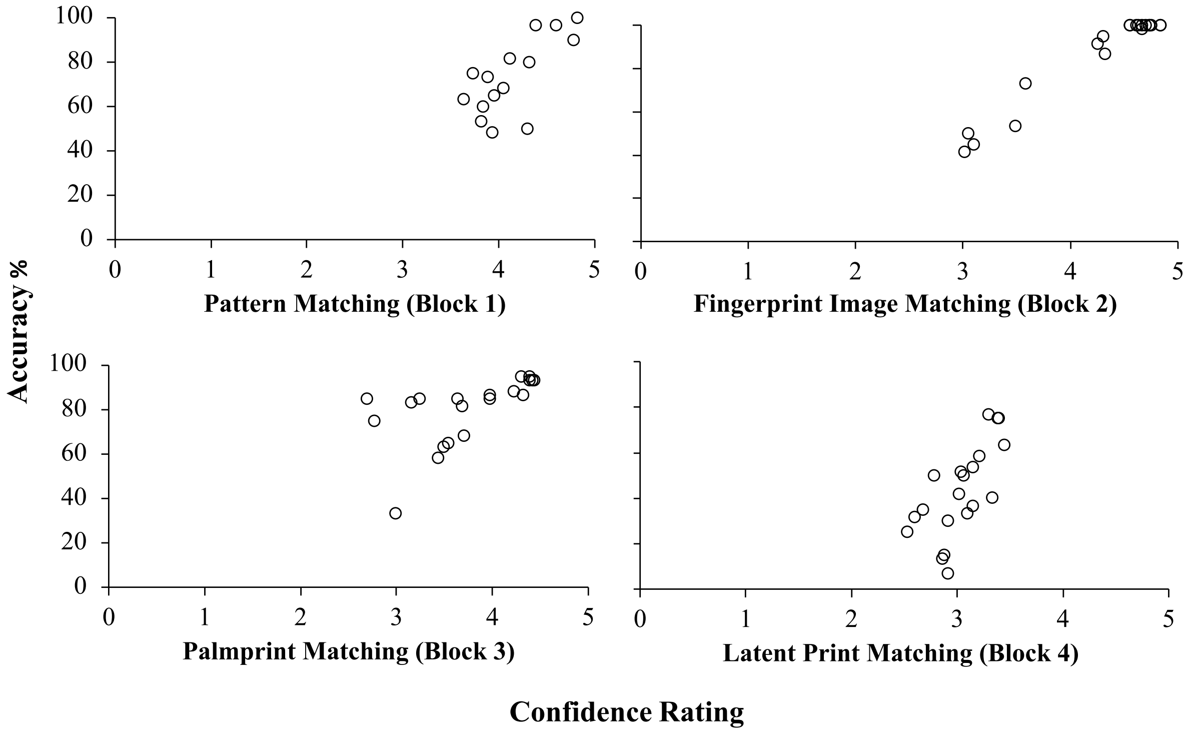
Note. In relation to the number of errors, in Block 1 there were 450 total trials, and in Blocks 2 to 4 there were 600 trials in each block. Error bars denote the standard error of the means.

Confidence Ratings

For each test item, participants were required to provide a rating of their confidence in their response being correct from 1 (Not Confident) to 5 (Extremely Confident). To explore the association between accuracy and confidence, the mean accuracy score and confidence rating was calculated for each item (see Figure 2.10), with data collapsed across both test sessions. This data was then used to compute Pearson correlation coefficients to examine the relationship between confidence and accuracy within each block. This revealed positive correlations in Block 1, $r = .710, p = .003$, Block 2, $r = .978, p < .001$, Block 3, $r = .593, p = .006$, and Block 4, $r = .694, p < .001$. Therefore, within all blocks of the test, higher accuracy was associated with higher observer confidence that the correct response had been provided.

Figure 2.10

Correlation of Mean Item Accuracy (%) and Confidence Rating (1 – 5) Within Each Block.



In summary, analyses of the data confirmed there were no differences in accuracy across both test sessions, and the observed differences in accuracy between the blocks reflected the varied demands of the test. Although performance in pattern matching (Block 1) and fingerprint image matching (Block 2) was not related to accuracy in latent print matching (Block 4), performance in palmprint matching (Block 3) was moderately associated with this task. There was a moderate to strong relationship between accuracy and confidence, with higher confidence associated with higher accuracy. In relation to errors in target-present trials, observers missed more targets in both the easiest (fingerprint image matching) and most difficult (latent print matching) blocks of the test.

Analysis of Response Time Data

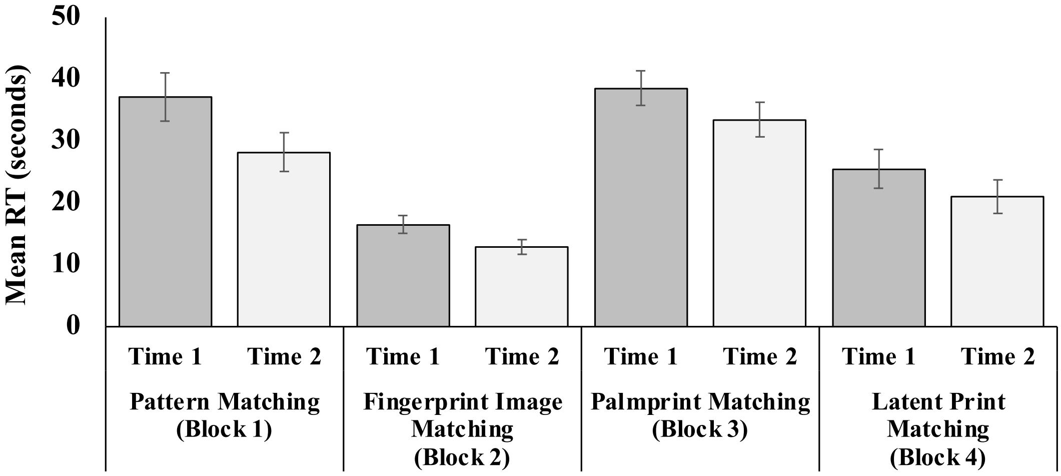
Response Times: For the fingerprint aptitude test, response time was recorded at the point at which the final response to each item was entered. This therefore reflected the time spent viewing each test item and entering the response and corresponding confidence rating. To explore whether differences in the overall time taken to complete each fingerprint test reflected different response times in each session, the mean response times were calculated in seconds for each block and are shown in Figure 2.11.

This data was analysed with a 2 (Session: Time 1 or Time 2) x 4 (Block: 1, 2, 3 or 4) mixed-model ANOVA. This revealed a main effect of Block, $F(3, 174) = 49.77$, $p < .001$, partial $\eta^2 = 0.46$. A series of paired Bonferroni *t* tests showed response times in Block 1 ($M =$ 32.60s, $SE = 2.50$) were slower than those in Block 2 ($M =$ 14.70s, $SE = 0.94$), $t(58) = 8.19$, $p < .001$, and Block 4 ($M =$ 23.30s, $SE = 2.05$), $t(58) = 4.08$, $p < .001$, but were not different to those in Block 3 ($M =$ 35.90s, $SE = 1.97$), $t(58) = 1.62$, $p = .67$. Responses in Block 2 were also faster than those in Block 3, $t(58) = 13.65$, $p < .001$, and Block 4, $t(58) = 5.69$, $p <$ .001,and faster in Block 4 than Block 3, $t(58) = 6.92$, $p < .001$. There was no main effect of

Session, $F(1, 58) = 3.17$, $p = .08$, partial $\eta^2 = 0.05$, and no interaction between Session and Block, $F(1,58) = 0.74$, $p = 0.53$, partial $\eta^2 = 0.01$.

Figure 2.11

Mean Response Times in Seconds for Each Block of the Tests at Time 1 and Time 2.



*Note***.** Error bars denote the standard error of the means.

In summary, there was no difference in the response times of the tests taken at Time 1 and Time 2. There were differences in response times between the blocks, with fingerprint image matching (Block 2) completed in the fastest time, whereas observers took less time to match latent fingerprints (Block 4) than they did to match patterns (Block 1) or to match palmprints (Block 3).

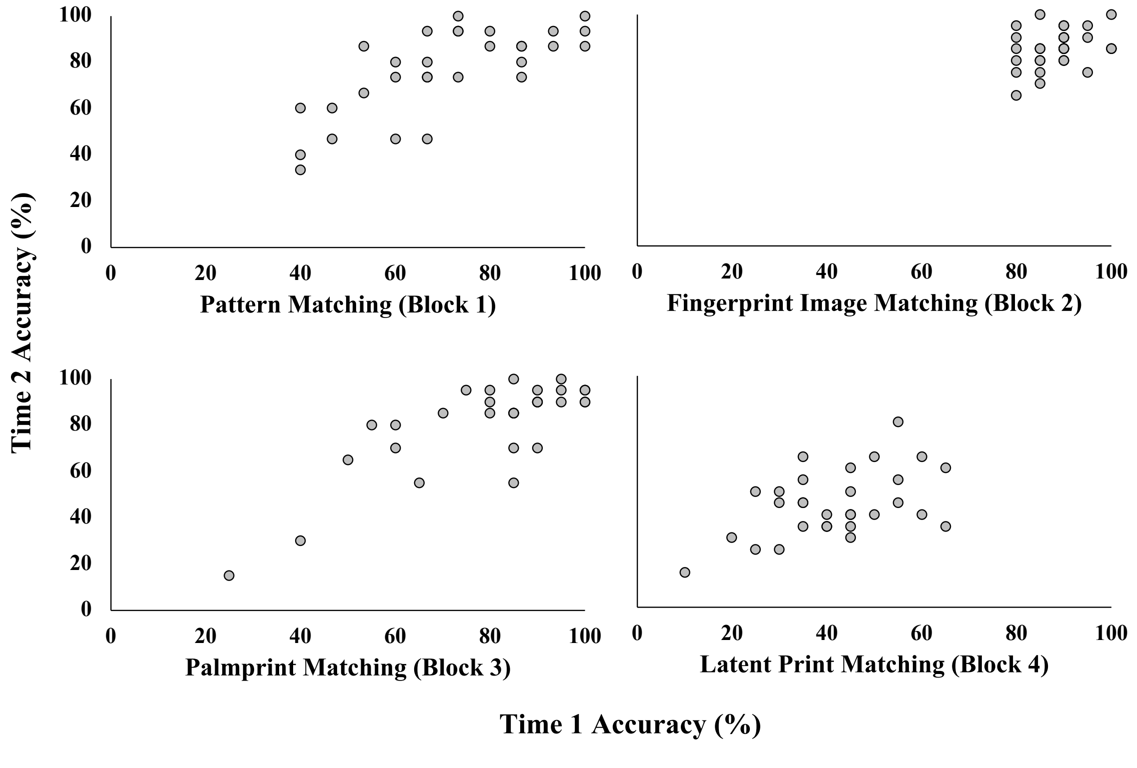
Test-Retest Reliability

Accuracy Correlations Between Test Sessions: To compare performance across both test sessions, the test-retest reliability of the fingerprint test was analysed using the mean accuracy score for each participant at Time 1 and Time 2 in each block of the test (Figure

2.12). Pearson correlation coefficients of this data revealed positive associations between the test scores at Time 1 and Time 2 in Block 1 ($r = .731, p < .001$), Block 3 ($r = .786, p < .001$), and Block 4 ($r = .484, p = .007$), but not in Block 2 ($r = .329, p = .08$). These correlations show that the fingerprint test has good re-test reliability in relation to Blocks 1, 3, and 4. The scatterplot for Block 2 shows very similar scores at Time 1 and Time 2, and the non-significant correlation is likely to reflect lack of variance in scores rather than differences across test sessions.

Figure 2.12

Correlations of Mean Subject Accuracy (%) Within Each Block at Time 1 and Time 2.

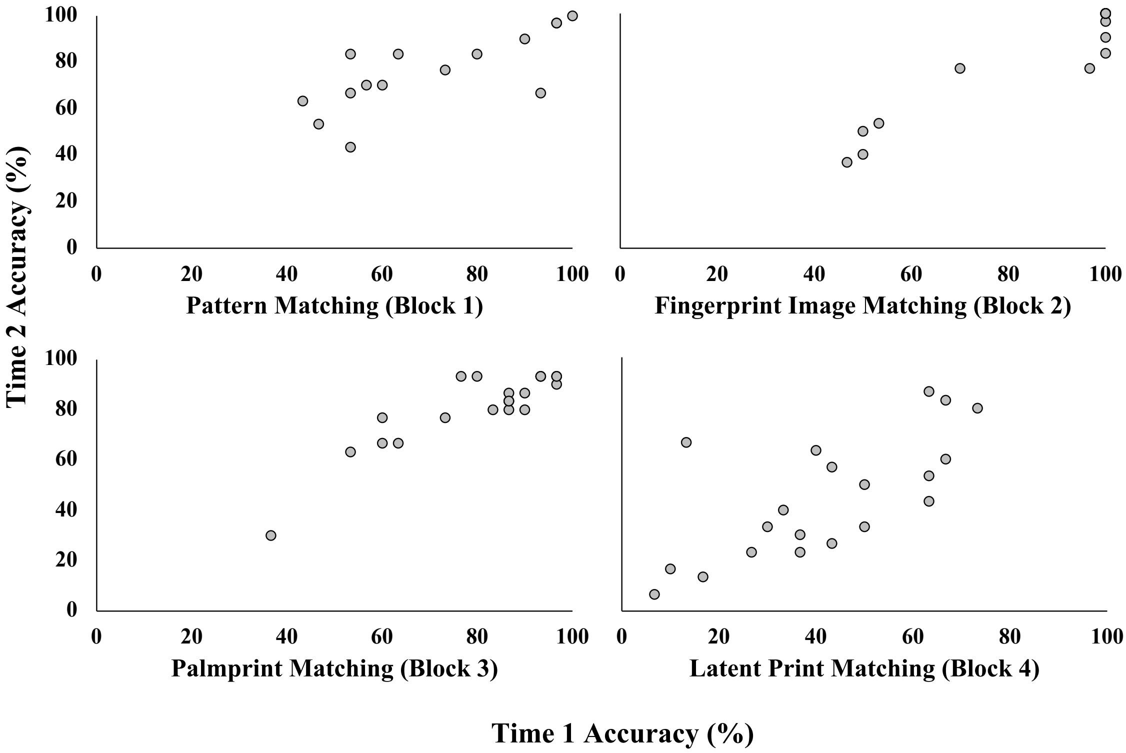


<u>Item-Level Reliability:</u> To explore whether the positive correlations for accuracy in both sessions also reflected similar responses to each test item, differences between accuracy scores were also examined at the item level for Time 1 and Time 2. For this purpose, mean

percentage accuracy scores were calculated for each item on a block-by-block basis for each test session and are shown in Figure 2.13

Figure 2.13

Correlation of Mean Item Accuracy (%) Within Each Block at Time 1 and Time 2.



Note. In Block 2, five trials scored 100% at Time 1 and Time 2.

This data was then used to compute Pearson correlation coefficients to assess the relationship between the mean accuracy scores within each block at the two timepoints. This revealed strong positive correlations between Time 1 and Time 2 within Block 1, $r = .750, p < .001$, Block 2, $r = .905, p < .001$, Block 3, $r = .880, p < .001$, and Block 4, $r = .700, p < .001$. These analyses suggest that participants responded similarly to the test items in each block at Time 1 and Time 2.

Cronbach’s Alpha: Finally, Cronbach’s Alpha was computed using raw item accuracy data from Session 1 for each block of the test. Block 1 contained 15 items (α = 0.72), Block 3 contained 20 items (α = 0.84), and Block 4 contained 20 items (α = 0.70). Cronbach’s Alpha could not be computed for Block 2 items due to the lack of variance in scores. Overall, these analyses show the fingerprint aptitude test had high re-test reliability, and good internal consistency in relation to test items in Blocks 1, 3, and 4.

In summary, the analyses show that test-retest performance was consistent, both across individuals and items, and the fingerprint test also demonstrated good internal consistency in relation to the majority of test items.

General Discussion

This chapter described the creation of a fingerprint aptitude test in which stimuli comprised line drawings, analogous to viewing patterns within fingerprints, and fingerprint images of varying degrees of complexity. Data from novice observers who participated in the test on two occasions, separated by an interval of around eleven days, was analysed to determine the validity and reliability of the fingerprint test. As this was a newly created test and different fingerprint stimuli were incorporated into each of the four blocks of the test, analyses measured whether these variations in stimuli were manifested by differences in accuracy and response times between each block, its internal reliability in so far as it measured fingerprint accuracy, and whether data from two testing sessions would confirm its repeatability.

In relation to repeatability, the analysis showed no difference in observers’ performance within each block of the test across both sessions. This was further supported by analysis of the data at both the subject and the item level which showed that accuracy correlated positively within each block across both test sessions, thereby reflecting a similar pattern of responses on both occasions. Given the interval of several days between tests, and the number of test items, it seems unlikely that this is accounted for by observers recalling their previous responses and repeating them during the second test. Analysis of the response time data also found no difference within each block across both sessions. In relation to the time taken to complete the trials in each block, analyses revealed consistent levels of performance across both sessions. Examination of the accuracy and response time data therefore shows that the test reliably measures aptitude for fingerprint comparison over time.

A further key element in the design of the fingerprint aptitude test was its ability to measure performance in a range of tasks relevant to forensic fingerprint comparison. As such, the content of each test block was designed to reflect these varied processes. Block 1 stimuli comprised of manufactured line drawings containing features analogous to fingerprints, with lines to represent creases and ridges and additional characteristics such as dots to represent minutiae. Although identical image matching is typically accomplished with relative ease (e.g., Jenkins et al., 2011), accuracy was lower than during fingerprint image matching (Block 2) and palmprint matching (Block 3), but higher than in latent print matching trials (Block 4).

In contrast to Block 1, comparisons in fingerprint image matching in Block 2 predominantly focussed on locating an exemplar that was identical to the target. At-ceiling accuracy for several test items shows this task could be accomplished with relative ease by observers, and accuracy was higher than in Block 1. This may reduce the effectiveness of this block in discriminating between test-takers. However, fingerprint experts advise that these fingerprint image matching tasks may identify those observers with low visual acuity, or a tendency towards inattention to task instructions or stimuli. Poor performance in these tasks may also indicate vision issues relevant to latent print comparison such as form blindness, an impaired ability to distinguish minute differences in stimuli relating to shape or size (Byrd & Bertram, 2003).

In the palmprint matching trials in Block 3, observers needed to locate a cropped section of palm on a complete palmprint. In some trials, this could be achieved by locating the identical image, in a similar method to that deployed in Block 1. In other trials, observers needed to mentally rotate an image to locate its counterpart. This task is routinely undertaken by forensic experts who place a print in its upright orientation to facilitate comparison with an exemplar. Accuracy here was higher than that observed in fingerprint image matching (Block 2) and latent print matching (Block 4), and lower than accuracy in pattern matching (Block 1). Despite the increasing complexity of these palmprint matching trials, the results show that observers were able to complete this block with a relatively high degree of accuracy.

In contrast to the preceding three blocks, the latent print matching trials (Block 4) revealed a significant decline in observers' accuracy. This section of the test was designed to challenge observers by requiring them to match latent prints to inked or rolled fingerprints, a task often undertaken by fingerprint examiners. The difficulty of the task was compounded by several factors. Firstly, several of the latent prints were unclear, contained background interference, or were distorted. These stimuli therefore contained fewer areas with high diagnostic value to the novice observer. They may, however, contain sufficient diagnostic information to allow an experienced fingerprint examiner to reach the correct conclusion (Busey et al., 2011). Some stimuli were also inverted, requiring the observer to mentally rotate the image as they had during the palmprint matching trials. The difficulty of the task was further increased by the close similarity of the target to the exemplars in the array. Successful matching of the target to the corresponding image was therefore comparable to the most difficult task undertaken by fingerprint examiners, that of evaluating similar mismatching finger marks (Tangen et al., 2011).

Analyses of the accuracy data for the fingerprint test therefore confirmed that each block varied in difficulty, thus capturing the variation in fingerprint comparisons tasks they were designed to reflect. Converging evidence in support of these differences is provided by the absence of accuracy correlations between some blocks of the test, thereby reflecting the dissociable aspects of the varied fingerprint tasks. Differences in the response times of observers in most blocks of the test also suggest that different cognitive processes may be engaged according to the demands of each block of trials.

At the same time, correlational analyses showed that observers tended to perform consistently across the test, with some showing high levels of accuracy across all four blocks. This indicates the suitability of the test to measure a generalised aptitude for fingerprint comparison. In each block of the fingerprint test, higher accuracy was also associated with higher confidence ratings which supports that observed in previous fingerprint comparison research (Kellman et al., 2014; Thompson et al., 2014), and in other forensic tasks such as unfamiliar face matching (White et al., 2014) and eyewitness identification (Wixted & Wells, 2017). Taken together, these findings show that each block of the test performed a different function in terms of assessing fingerprint comparison aptitude and captured the observers' ability across a range of varied fingerprint stimuli.

Finally, the accuracy data was also examined in relation to errors. In particular, this analysis focused on how frequently 'no match' was erroneously entered as a response during target-present trials, thereby indicating that a target had been missed as opposed to the observer incorrectly selecting another exemplar. Analyses showed that targets were missed more frequently during pattern matching (Block 1) and palmprint matching (Block 3) than in the other two blocks. It is currently difficult to draw conclusions as to whether this reflected a pattern of responding to particular trial stimuli, but the mixture of these responses indicates the potential of the fingerprint test to probe for both types of errors in different observer groups. Comparing the responses of forensic experts and a large sample of novice observers will assist in further identifying the relevance of this finding.

In summary, this chapter detailed the creation of a fingerprint aptitude test and reported the data from a group of novice observers who undertook took the test on two occasions. Analyses showed the test reliably captured aptitude for fingerprint comparison over time and confirmed its suitability to differentiate the abilities of a wider pool of observers. The variation in stimuli used in each block of the test manifested differences in accuracy and response times, thereby capturing different aspects of fingerprint comparison. In the next stage of the research, the fingerprint aptitude test will be used to examine the nature of fingerprint expertise using a larger sample of novice observers, experienced fingerprint examiners, and examiners-in-training.

Chapter 3

Measuring Fingerprint Comparison Expertise

3.1 Introduction

The previous chapter described the creation and reliability testing of a fingerprint aptitude test with data from novice observers who undertook the test on two occasions. Across four different blocks of trials, the test measured observers' performance in pattern matching, fingerprint image matching, palmprint matching, and latent print matching trials. There were differences in performance between each block, thus indicating that the test captured the complexity and format of the different stimuli, with latent print matching the most challenging of the four blocks of the test. The intention of the current experiment is to use this fingerprint test to compare the abilities of a larger group of novice observers, experienced fingerprint examiners, and fingerprint examiners in training. The aim is to identify the level at which professional examiners might excel in fingerprint comparison tasks, and to examine whether this level of performance can be equalled, or surpassed, by observers who are untrained in fingerprint comparison methods. In short, the chapter examiners the nature of fingerprint examiners' *expertise*.

A definition of expertise is that it should reflect training and experience in an area of knowledge that is likely to be outside the domain of the general population (Roberts, 2021). Forensic fingerprint examiners undergo a rigorous programme of training, assessment, and mentorship to achieve such 'expert' status, which qualifies these professionals to provide forensic testimony as to the source of a fingerprint in criminal investigations and judicial proceedings (CPS, 2020). In turn, the need for fingerprint comparison ability is rare in everyday life, with few opportunities for observers who are untrained in fingerprint analysis to practice this skill. Therefore, previous research reports of an accuracy advantage for experienced examiners over untrained observers are not unexpected (Searston & Tangen, 2017a, 2017b; Stevenage & Pitfield, 2016; Tangen et al., 2011; Thompson et al., 2014). However, *some* knowledge of fingerprint comparison methods does not necessarily equate to expert performance. In one study, for example, new trainees with experience of between five weeks and six months were less accurate than untrained observers in the comparison of matching fingerprints (i.e., those from the same source), but were more accurate than novices with non-matching and similar non-matching fingerprints (Thompson et al., 2014).

Such group differences can make it difficult to define clear cut-offs for expert performance and this issue is complicated further by *individual differences* in fingerprint comparison ability. Experienced fingerprint examiners can exhibit inconsistencies in their fingerprint identification accuracy (Ulery et al., 2011, 2012; Wertheim et al., 2006), as well as in the number and type of minutiae that are identified to reach an identification decision (Ulery et al., 2014). These inter-observer differences are not unique to fingerprint comparison but are observed in perceptual tasks such as object recognition (Gauthier, 2018), and in professional roles such as face identity matching in passport officers (White et al., 2014), mammogram interpretation by radiographers, (Hornsby & Love, 2014), and decision making by forensic firearms examiners (Mattijssen et al., 2021). Research in these related domains demonstrates that untrained observers also exhibit similar individual differences in performance (see, e.g., Bindemann et al., 2012; Fysh & Bindemann, 2018). This can produce substantial overlap in identification performance between novices and forensic experts (see, e.g., Phillips et al., 2018). In the research of fingerprint expertise, however, research has so far focused only on differences between *groups* of observers. Consequently, the extent to which the fingerprint identification performance of novices and experts might overlap remains unknown. Considering that a fingerprint expert should possess knowledge and ability that is not found in the population at large (CPS, 2020), this is an important question to resolve.

This issue is complicated further as previous studies of fingerprint comparison have typically employed specific sets of stimuli, for example, comprising solely of simulated crime scene prints (Tangen et al., 2011), or genuine crime scene prints (Thompson et al., 2014), high- and low-quality fingerprints from a database (Stevenage et al., 2017), or cropped rolled impressions (Searston & Tangen, 2017a; 2017b). By examining fingerprint expertise with stimuli in only a single format, it becomes difficult to establish the types of fingerprints with which professional examiners might excel. Yet, identifying those tasks that differentiate the performance of professional fingerprint examiners and non-professional observers is key to understanding the nature of fingerprint comparison expertise.

The current chapter employs the fingerprint aptitude test that was developed in Chapter 2 to address these questions by comparing experienced fingerprint examiners, trainee fingerprint examiners, and novices in fingerprint comparison tasks. The fingerprint aptitude test requires observers to match patterns and fingerprints in the absence of any identification and magnification aids usually deployed by experts. This is designed to capture any differences in perceptual ability rather than permitting experts to rely on their workplace comparison protocols. The analysis first focuses on general quantitative differences between these groups in relation to accuracy, speed of responses, and errors, to identify those aspects of fingerprint comparison that are most associated with expertise. Then the fingerprint comparison accuracy of *individual* novice observers will be compared with that of the professional examiners. This approach will offer insight into this ability in the general population and will identify the extent to which untrained observers can match, or exceed, the performance of forensic fingerprint examiners when comparing a range of fingerprint stimuli.

3.2 Method

Participants

Seventeen qualified fingerprint examiners (‘Experts’) from a UK police service took part in this experiment (mean age = 46.41 years, *SD* = 9.30, range 27-59 years, 6 males), with mean experience in fingerprint examination of 20.47 years (*SD* = 10.80, range 2-36 years). Twenty trainee examiners (‘Trainees’) from the same UK police service (mean age = 27.55 years, *SD* = 6.43, range = 21-48, 1 male) with mean training in fingerprint examination of 16.70 weeks (*SD* =18.76, range 1-52 weeks) also took part. All police employees undertook the experiment online whilst they were working from home during the Covid pandemic.

A further group of ninety-six participants who were untrained in fingerprint examination (mean age = 20.30 years, *SD* = 4.62, range 18-58, 21 males) participated as ‘Novices’, and undertook the experiment online. This group predominantly comprised university students who participated in return for course credits, and the remainder were volunteers who received no remuneration for their participation. All participants reported normal or corrected-to-normal eyesight and provided informed consent to take part. This research was approved by the University of Kent Ethics Committee.

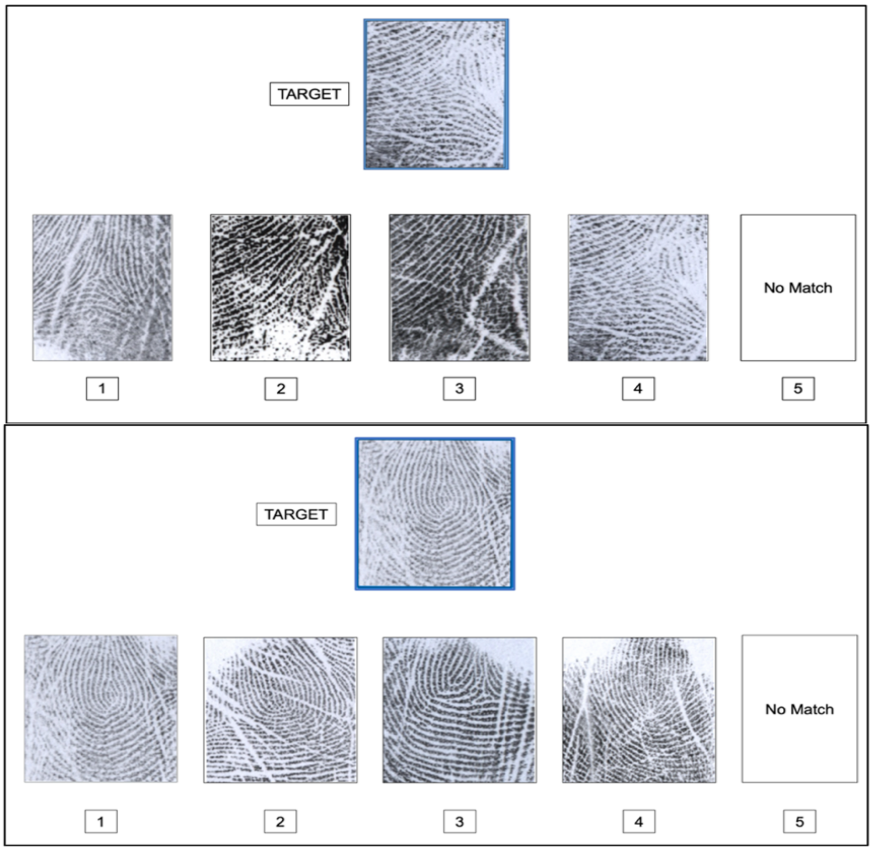
Materials

This experiment used the Online Fingerprint Aptitude Test described in Chapter 2. Thus, the test is only summarised briefly here. The test comprised of four blocks of trials, presented in numerical order, and with the same order of stimuli presentation maintained for all observers. Block 1 measured Visual Pattern Matching and comprised of fifteen trials in which participants were presented with a single target above an array of four exemplars and a box labelled ‘No Match’.

Block 2 measured Fingerprint Image Matching and comprised of twenty fingerprint image matching trials and followed the same format as Block 1 with a single target image presented above an array of four exemplars and a box labelled ‘No Match’. An example of test stimuli is shown in Figure 3.1.

Figure 3.1

Example of Target Present (top) and Target Absent (bottom) Trials from Block 2

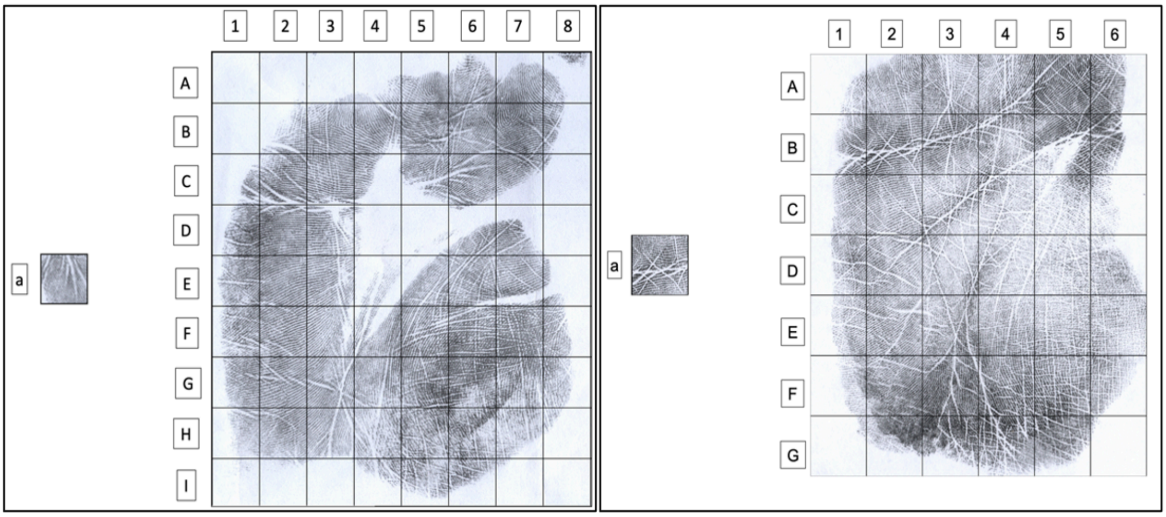


The target was either a rolled fingerprint or a cropped section from a palm print, and in six trials the target was absent from the array. In the first fifteen trials, observers needed to match the target with its identical image, i.e., a plain impression to a plain impression or a rolled impression to a rolled impression. In the remaining trials the target and its corresponding image were not identical, and observers needed to match a plain impression to a rolled impression or vice versa. Each target and array measured 1000(w) x 562(h) pixels at a screen resolution of 127 ppi.

Block 3 measured Palmprint Matching and comprised of twenty palmprint matching trials in which a cropped section from a palm print needed to be located on, or excluded from, a complete palm print. In the first ten trials, an 8 x 9 grid was overlaid onto a palmprint, annotated with numbers and letters to allow identification of the target area. In the final ten trials, a different palmprint was used and was overlaid with a larger 6 x 7 annotated grid. Example stimuli are shown in Figure 3.2.

Figure 3.2

Example of a Target with Grid Overlay on a Complete Palm from Block 3, with Trials 1-10 on the Left and Trials 11-20 on the Right

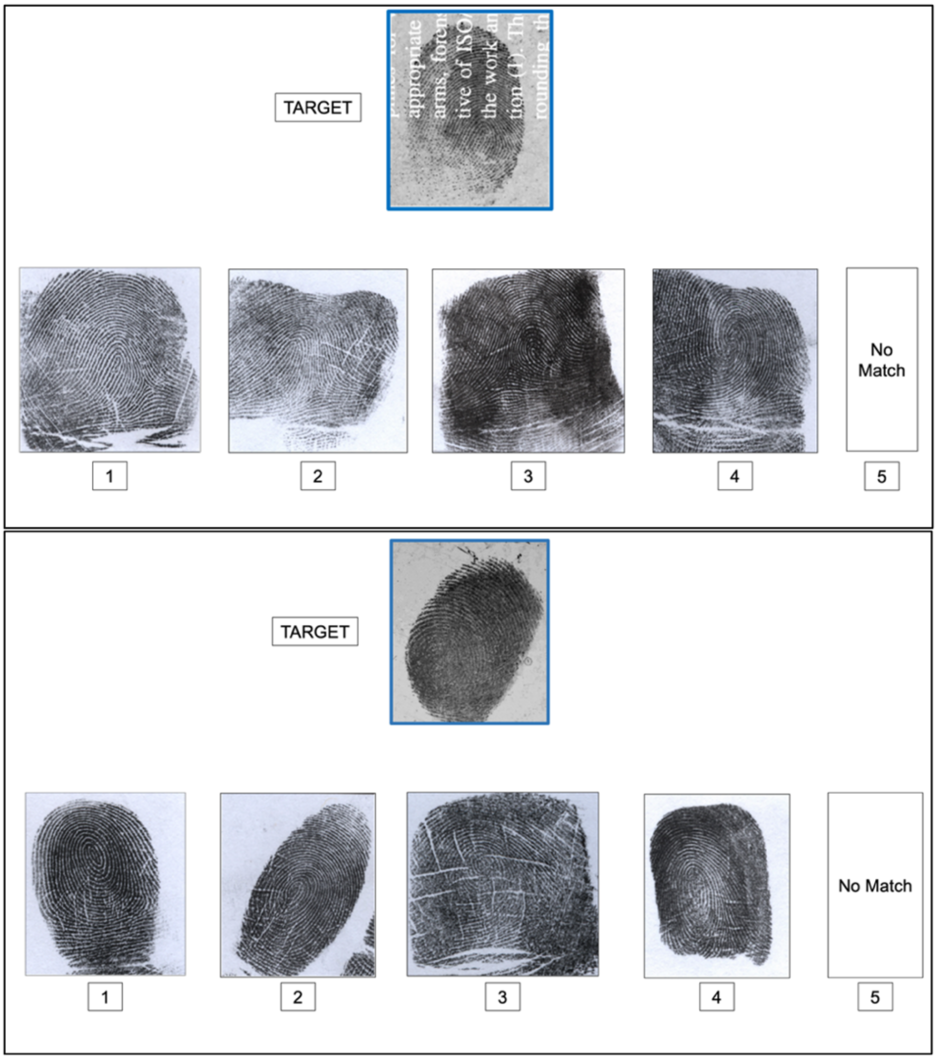


For each palm print, four targets were correctly oriented, two were inverted, one target was oriented ninety degrees to the left, one target oriented ninety degrees to the right, and the target was absent in two trials. The total size of the target and comparison palm print as displayed on screen was 700(w) x 555(h) pixels at a screen resolution of 127 ppi.

Finally, Block 4 measured Latent Print Matching and comprised of twenty fingerprint print matching tasks in which the target print was a latent print, representative of those left at a crime scene. The array was constructed from inked impressions, and example stimuli are shown in Figure 3.3.

Figure 3.3

Example of Target Present (top) and Target Absent (bottom) Trials from Block 4



The test format followed that of Blocks 1 and 2, with the target presented above an array of four exemplars together with a box labelled ‘No Match’. In four trials, the target was absent. Each target and array measured 1000(w) x 563(h) pixels at a screen resolution of 127 ppi.

3.3 Procedure

This experiment was conducted online using Qualtrics software and all participants undertook the Fingerprint Aptitude Test. Prior to commencing the test, participants were provided with instructions to calibrate their computer screen. This was done by placing a credit card against an onscreen template of a standard sized credit card (85.6mm x 53.98mm) and adjusting the browser magnification until the card and template matched. Task specific instructions preceded each block, and an example question with the correct answer was always shown. One question was presented per page and remained visible until a response had been entered. All questions required the observer to match a target to a sample array or grid, or to identify that the target did not match the exemplars.

All responses were entered using a standard computer keyboard and all questions needed to be attempted. Observers were able to change their response to any question currently in view but were unable to amend or review any previous responses. Pre-test information advised that the test needed to be taken in one session and may take up to two hours to complete. Task instructions emphasised accuracy over speed of responses. Onscreen information displayed between each block of the test advised observers to take a short screen break prior to moving on to the next section.

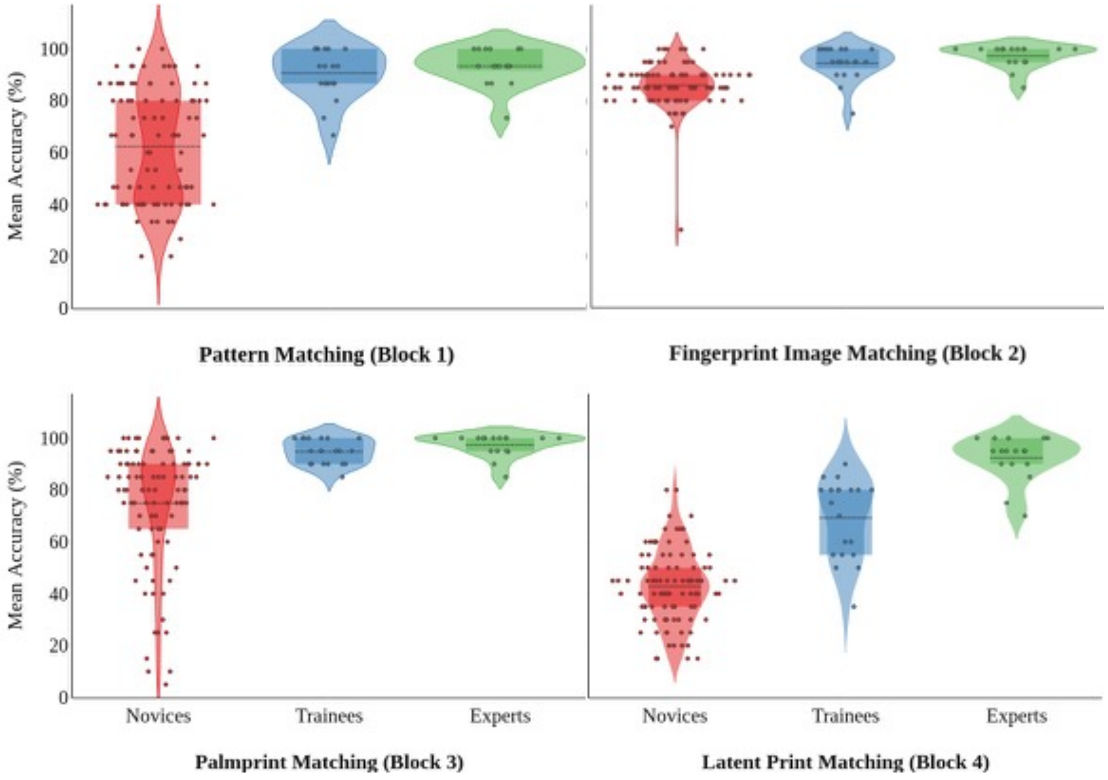
3.4 Results[1]

Analysis of Accuracy Data

Participant Accuracy: To compare performance between novices, trainees and experts, mean percentage accuracy was first calculated separately for each group in each block of the fingerprint test. As shown in Figure 3.4, experts appeared the most accurate across all blocks, with Novices consistently the least accurate of the three groups.

Figure 3.4

Mean Percentage Accuracy Scores for Each Group Within Each Block of the Test



Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

[1] The results focus on differences between Novices, Trainees and Experts. For this reason, only group results are reported. Full details of all analyses, including all main effects and non-group interactions, are included in the appendix.

These differences in accuracy were explored with a 3(Group: Experts, Trainees and Novices) x 4(Block: 1, 2, 3 or 4) mixed-model ANOVA of the data. This revealed a main effect of Group, *F*(2, 131) = 56.79, *p* < .001, partial η^2 = 0.461. A series of independent samples *t* tests with Bonferroni correction showed that Novices (*M* = 66.00, *SE* = 1.23) were less accurate than both the Expert (*M* = 95.10, *SE* = 2.94), *t*(131) = 9.12, *p* < .001, and Trainee groups (*M* = 86.40, *SE* = 2.71), *t(*131) = 6.83, *p* < .001, with no difference in accuracy between Experts and Trainees, *t*(131) = 2.18, *p* = .09.

There was an interaction between Group and Block, *F*(6, 393) = 13.75, *p* < .001, partial η^2 = 0.172. This was explored with a series of independent *t* tests with Bonferroni correction to compare accuracy between the groups across each block of the test. A full summary of post hoc tests is provided in Table 3.1.

Table 3.1

Summary of Post Hoc Analyses Comparing Accuracy Between Groups, With T-Values and Cohen's D for Each Comparison.

Reference and Comparison Group	Block 1	Block 2	Block 3	Block 4
Novices: Experts	*t* = 6.12** (<) *d* = 1.50	*t* = 5.47** (<) *d* = 1.39	*t* = 4.12** (<) *d* = 1.04	*t* = 13.37** (<) *d* = 3.61
Novices: Trainees	*t* = 6.01** (<) *d* = 1.40	*t* = 4.29* (<) *d* = 1.00	*t* = 3.25 *d* = 0.75	*t* = 7.76** (<) *d* = 1.82
Trainees: Experts	*t* = 0.41 *d* = 0.30	*t* = 1.17 *d* = 0.56	*t* = 0.86 *d* = 0.60	*t* = 4.93** (<) *d* = 1.78

Note. Significant comparisons are denoted * *p* < .01, ** *p* < .001. Symbols **(>** and **<)** identify whether the reference group is more or less accurate than the comparison group.

Overall, this comparison showed that Experts were more accurate than Novices in each block. This reflected a large (*d* =1.04) to very large (*d* = 1.50) difference in accuracy in the first three blocks and in Block 4, Cohen's *d* of 3.61 revealed a huge performance gap between these groups. This equates to a 99.5% chance that an observer picked at random from the Novice group will have a score lower than an observer picked at random from the Expert group. Trainees were also more accurate than Novices in all blocks except palmprint matching in Block 3. These differences were very large in Blocks 1 and 4 (*d* = 1.40 and *d* = 1.82 respectively), and large in Block 2 (*d* = 1.00). Finally, differences in accuracy between Trainees and Experts only emerged in Block 4, with Experts more accurate than Trainees. Cohen's *d* of 1.78 reflected a very large difference between these groups, with an 89.6% chance that an observer picked at random from the Trainee group will have a higher score than an observer picked at random from the Expert group.

The interaction was further explored with a series of paired *t* tests with Bonferroni correction to compare differences in Group performance within each block of the test and are shown in Table 3.2. These comparisons show that Expert performance was consistently high throughout the test, with no difference in accuracy in each block. By comparison, although Trainee accuracy was consistent in the first three blocks, accuracy in Block 4 was lower than in the preceding blocks. In this group, Cohen's *d* revealed a very large difference in accuracy between Block 1 and Block 4 and between Block 1 and Block 3 (*d*s > 1.2), and a huge difference in accuracy between Block 2 and Block 4 (*d* > 2.00). Finally, accuracy in the Novice group varied, with differences emerging between each block of the test. The smallest difference in accuracy was between Block 1 and Block 3 (*d* = 0.56), and the largest difference in accuracy was between Block 2 and Block 4 (*d* = 3.65).

Table 3.2

Summary of Post Hoc Analyses Comparing Accuracy Within Groups Across Each Block, With T-Values and Cohen's D for Each Comparison.

Reference and Comparison Block	Novices	Trainees	Experts
Block 1 : Block 2	t = 12.95*(<) d = 1.42	t = 0.89 d = 0.44	t = 0.95 d = 0.70
Block 1 : Block 3	t = 6.15*(<) d = 0.56	t = 0.13 d = 0.05	t = 0.85 d = 0.70
Block 1 : Block 4	t = 9.49*(>) d = 1.05	t = 4.77*(>) d = 1.65	t = 0.18 d = 0.10
Block 2 : Block 3	t = 6.07*(>) d = 0.63	t = 0.75 d = 0.29	t = 0.01 d = 0.00
Block 2 : Block 4	t = 29.77*(>) d = 3.65	t = 7.87*(>) d = 2.10	t = 1.45 d = 0.71
Block 3 : Block 4	t = 15.81*(<) d = 1.66	t = 4.95*(>) d = 1.52	t = 1.04 d = 0.71

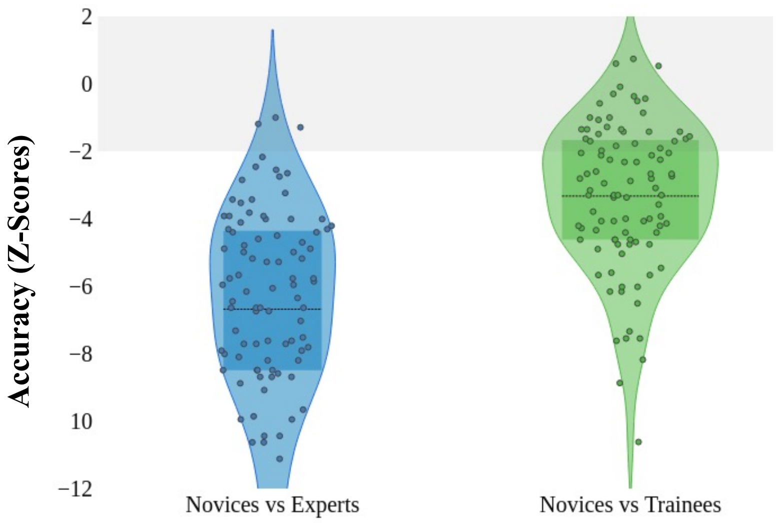
Note. Significant comparisons are denoted * $p < .001$. Symbols (> and <) identify whether accuracy in the reference block is more or less accurate than the comparison block.

Overall, these data therefore show that latent print matching in Block 4 proved the most challenging for Novices and Trainees, whereas Experts demonstrated a clear accuracy advantage in this task. They also maintained their high performance throughout each block of the test.

Individual Differences in Accuracy: Analyses of accuracy data showed that Experts outperformed Novices across all blocks of the test, with Trainees more accurate than Novices in every block except palmprint matching. However, these group differences concealed a range of individual differences, notably within the Novice group. To compare the performance of individual Novices against that of Experts and Trainees, *z*-scores were computed for each Novice observer using the mean accuracy and standard deviation of each police group (see Figure 3.5). In this way, Novice *z*-scores could be compared separately with those in the Expert and Trainee groups.

Figure 3.5

Comparison of Novice Z-Scores for Mean Overall Accuracy Calculated Using the Means and Standard Deviations of the Expert and Trainee Groups



Note. The boxplot reflects the interquartile range of scores, and the black line within each plot represents the mean *z*-score of Novices. The grey shaded area denotes scores within the -1.96 *SD* to 1.96 *SD* range.

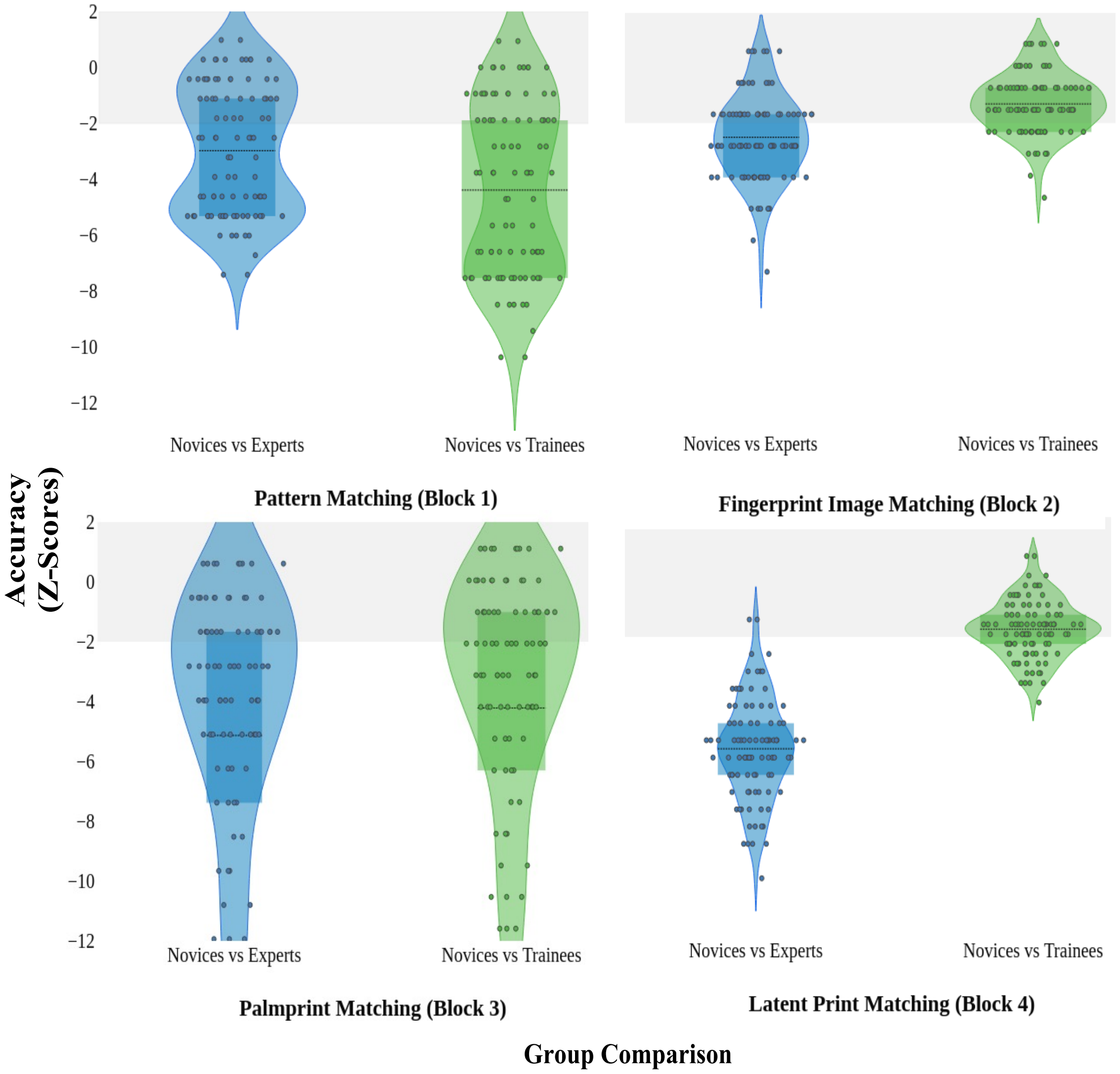
Z-scores of 1.96 and -1.96 standard deviations were applied as a two-sided test of difference in accuracy between Novices and the police groups, with scores outside this range therefore reflecting a difference in accuracy with ninety-five percent confidence ($p < .05$)., None of the Novice *z*-scores were above 1.96 which indicated that they did not outperform the Expert and Trainee groups. Novice *z*-scores below -1.96 reflected accuracy that was below that of the comparison group. In relation to overall accuracy, Figure 3.6 shows that only three Novices were at least as accurate as Experts, and around one third were at least as accurate as Trainees.

To compare accuracy on a block-by-block basis, mean Novice accuracy in each block was again converted to *z*-scores using the mean and standard deviation of the Trainee and Expert groups. Figure 3.6 shows that the majority of Novices were less accurate than Experts in all blocks of the test. The comparison of Novice performance within Trainee accuracy is less clear. Although most Novices were less accurate than Trainees in Block 1, in the remainder of the blocks it appears that approximately fifty percent of Novice scores were different to those of Trainees. The scores of Novices did not exceed 1.96 standard deviations in any block of the fingerprint test.

To provide a more accurate picture of Novice performance in comparison with Experts and Trainees, the percentage of Novices with a *z*-score below -1.96 standard deviations was calculated. This data therefore reflect the percentage of Novices with scores that were lower than the Trainee and Expert groups. None of the Novice observers had *z*-scores above 1.96 in any comparison. The analysis is shown in Table 3.3 and shows the majority of Novices were less accurate than Experts across all blocks of the test. In comparison with Trainees, around two-thirds of Novices were less accurate than Trainees in Blocks 1 and 3 and in terms of overall accuracy. In Blocks 2 and 4, around two-thirds of Novices were at least as accurate as Trainees.

Figure 3.6

By Block Comparisons of Novice Z-Scores for Mean Accuracy, Calculated Using the Means and Standard Deviations of the Expert and Trainee Groups



Note. Plots have been scaled to allow comparison, and data below -12 is therefore not shown in Block 3.The boxplot reflects the interquartile range of scores, and the black line within each plot represents the mean *z*-score of Novices. The grey shaded area denotes scores within the -1.96 *SD* to 1.96 *SD* range.

Table 3.3

Percentage of Novice Observers in Each Block with Accuracy Z-Scores Below -1.96 SD in Comparison Within Expert and Trainee Z-Scores

Comparison	Pattern Matching (Block 1)	Fingerprint Image Matching (Block 2)	Palmprint Matching (Block 3)	Latent Print Matching Block 4	Overall Mean Accuracy
Vs Experts (%)	58.33	57.29	63.54	97.92	96.87
Vs Trainees (%)	64.58	27.08	63.54	33.33	69.79

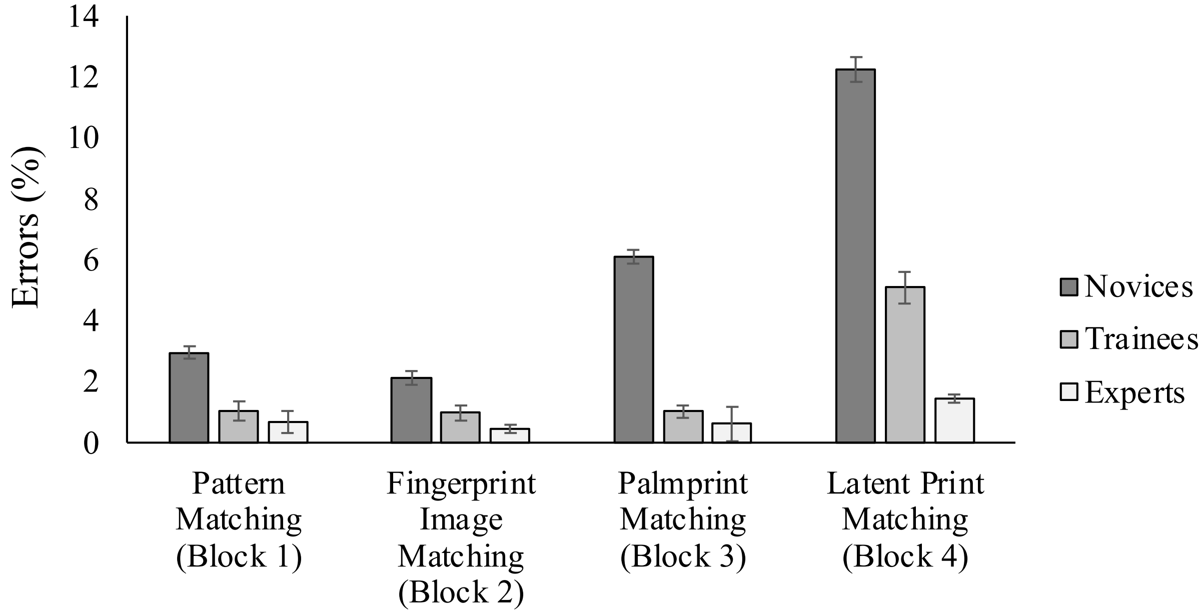
Note. Novices in each block, *n* = 96

In summary, in Blocks 1, 2 and 3 the performance of many Novices fell within the Expert range. In contrast, a clear difference in performance emerged in Block 4 and in overall test accuracy, with Expert scores higher than ninety-seven percent of Novices. Differences between Trainees and Novices were less marked. In Blocks 1 and 3 and in terms of overall accuracy, the performance of around two-thirds of Novices was lower than that of Trainees. In Block 2 and 4 the accuracy of around one-thirds of Novice was lower than that of Trainees.

Errors in Target-Present Trials: During target-present trials observers could enter a ‘no match’ response if they could not locate the target in the array, thus indicating that the target had been missed. To examine whether there was any difference in the percentage of targets missed by Experts, Trainees, and Novices, the number of errors in target-present trials was first counted. This data reflected the number of trials in which the target had been *missed* (‘no match’) as well as those trials in which the target was *misidentified*. The mean percentage of trials in which there were errors was then calculated (Figure 3.7). This data broadly reflects the accuracy data with fewer errors in Blocks 1 and 2, and more errors committed in Block 4.

Figure 3.7

Percentage of Errors (Target Misses and Target Misidentifications) in Target Present Trials by Block and by Group



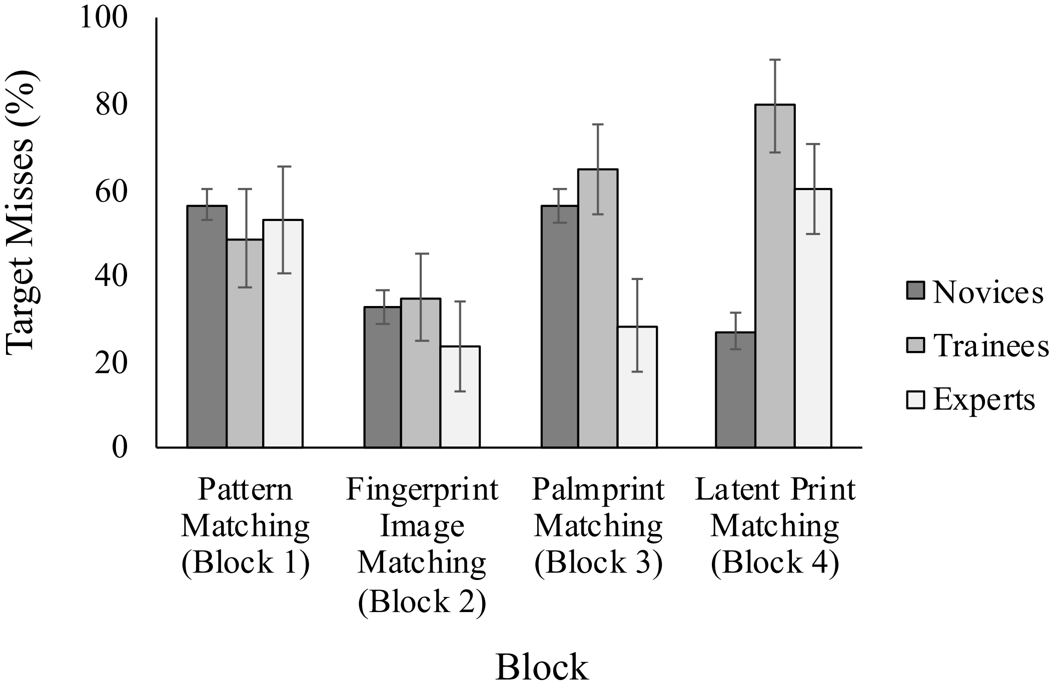
Note. Error bars denote the standard error of the mean

The mean percentage of trials in which the target was missed was then computed for each group in each block of the test. As shown in Figure 3.8, there are between-group differences in the percentage of targets missed, and in the percentage of targets missed in each block. These were analysed with a 3(Group: Novices, Trainees, Experts) x 4 (Block: 1, 2, 3, 4) mixed-model ANOVA of the data. This revealed a main effect of Group, $F(2, 131) = 4.43$, $p < .001$, partial $\eta^2 = 0.06$. A series of independent t tests showed that, collapsing data across all blocks of the test, Novices ($M = 43.1$, $SE = 2.02$) missed targets less than Trainees

(*M* = 57.1, *SE* = 4.45), *t*(131) = 2.86, *p* = .02, with no difference between Novices and Experts (*M* = 41.3, *SE* = 4.82), *t*(131) = 0.35, *p* > .05. The difference between Trainees and Experts was not significant, *t*(131) = 2.42, *p* = .05.

Figure 3.8

Percentage of Target Miss Errors as a Percentage of Total Errors in Target Present Trials by Block and by Group



Note. Error bars denote the standard error of the mean

There was an interaction between Group and Block, *F*(6, 393) = 7.14, $p < .001$, partialη^2 = 0.09, which was explored with a series of independent *t* tests with Bonferroni correction. A full summary of post hoc comparisons between groups is shown in Table 3.4. This shows that differences between the groups only emerged in Block 4, with Novices missing fewer targets than Trainees or Experts.

Table 3.4

Summary of Post Hoc Analyses Comparing the Percentage Targets Missed Between Groups, With T-Values and Cohen's D for Each Comparison

Reference and Comparison Group	Block 1	Block 2	Block 3	Block 4
Novices: Experts	*t* = 0.36 *d* = 0.10	*t* = 0.84 *d* = 0.22	*t* = 2.48 *d* = 0.66	*t* = 5.30*** (<) *d* = 1.33
Novices: Trainees	*t* = 0.83 *d* = 0.21	*t* = 0.24 *d* = 0.06	*t* = 0.85 *d* = 0.21	*t* = 8.99***(<) *d* = 2.83
Trainees: Experts	*t* = 0.33 *d* = 0.08	*t* = 0.85 *d* = 0.25	*t* = 2.60 *d* = 0.77	*t* = 2.46 *d* = 0.53

Note. Significant correlations are denoted by * $p < .05$, ** $p < .01$, *** $p < .001$. Symbols (> and <) identify whether the reference group use 'no match' more or less than the comparison group.

The interaction was further explored with a series of paired *t*-tests to compare the targets missed within each group across the test. A full summary is shown in Table 3.5 which shows that Novices missed more targets in Blocks 1 and 3 than in Block 4 and missed fewer in Block 2 than in Block 3. For both Experts and Trainees, fewer targets were missed in Block 2 than in Block 4, but other comparisons were not significant.

Table 3.5

Summary of Post Hoc Analyses Comparing Percentage 'No Match' Responses Within Groups Across Each Block, With T-Values and Cohen's D for Each Comparison.

Reference and Comparison Block	Novices	Trainees	Experts
Block 1 : Block 2	*t* = 3.93 ** (>) *d* = 0.66	*t* = 1.02 *d* = 0.28	*t* = 2.00 *d* = 0.60
Block 1 : Block 3	*t* = 0.98 *d* = 0.03	*t* = 1.34 *d* = 0.32	*t* = 1.86 *d* = 0.50
Block 1 : Block 4	*t* = 6.70 *** (>) *d* = 1.11	t = 3.18 *d* = 0.77	*t* = 0.70 *d* = 0.15
Block 2 : Block 3	*t* = 4.16 ** (<) *d* = 0.58	*t* = 2.40 *d* = 0.62	*t* = 0.36 *d* = 0.11
Block 2 : Block 4	*t* = 1.25 *d* = 0.18	*t* = 4.66 *** (<) *d* = 1.20	*t* = 3.54 * (<) *d* = 0.79
Block 3 : Block 4	*t* = 5.71 *** (>) *d* = 0.91	*t* = 1.31 *d* = 0.38	*t* = 2.62 *d* = 0.69

Note. Significant comparisons are denoted by * *p* < .05, ** *p* < .01, *** *p* < .001. Symbols (> and <) identify whether targets were missed the reference block more or less than the comparison block.

In summary, across the fingerprint aptitude test targets were missed more frequently by Trainees than Novices, and there was no difference between Trainees and Experts. Accuracy data had previously identified Block 4 as containing the most challenging comparisons, and it was here that differences in target misses emerged between the Experts and Trainees and Novices. Using Sawilowsky's (2009) interpretation of Cohen's *d* effect sizes, this revealed a very large difference between Novices and Experts (*d* = 1.33), with Experts missing targets twice as often as Novices. The difference in target misses between Novices and Trainees was huge (*d* = 2.83), with Trainees missing the targets more than twice as often as Novices. There was no difference between Experts and Trainees in this block.

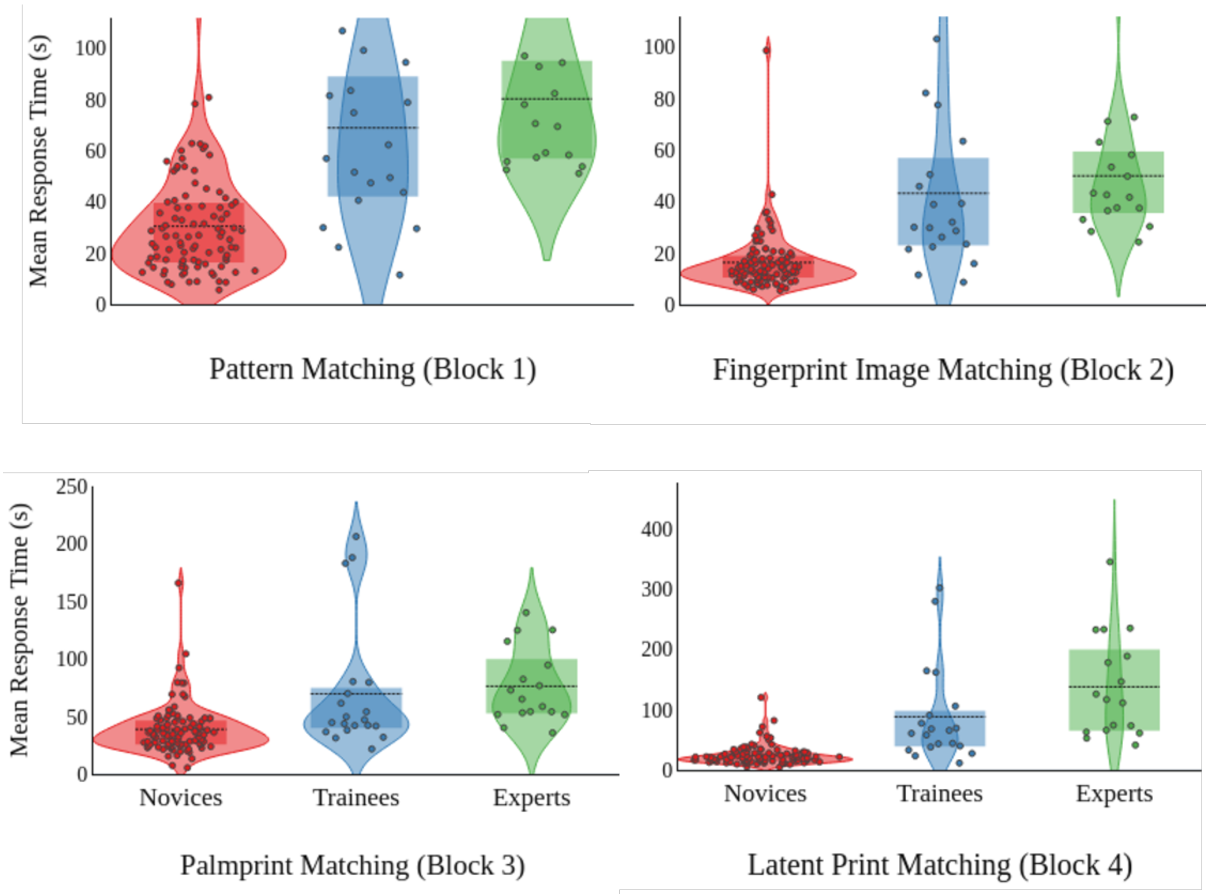
Analysing the targets missed by each group in a block-to-block comparison revealed a different pattern of responding between Novices and the police examiners. Trainees and Experts both missed targets more often during the challenging tasks in Block 4 than they had in the easier tasks in Block 2, with no differences emerging between the other blocks of the test. In the context of Novice test accuracy, Novices missed fewer targets during the easier trials in Block 2 than they did during Block 1 and Block 3. Conversely, when faced with more difficult comparisons in Block 4, Novices missed fewer targets than they had in Blocks 1 and 3. Therefore, the tendency for Experts and Trainees to miss more targets when comparisons were most ambiguous was not observed in Novices.

Analysis of Response Time Data

Participant Responses: Forensic fingerprint comparison is typically perceived as a methodical process that allows close examination of features within fingerprints. To identify whether there were differences in the response times of Novices, Trainees and Experts undertaking the fingerprint aptitude test, the mean response time was first calculated separately for each group in each block. As shown in Figure 3.9, Experts were slowest to respond across all blocks and Novices consistently the fastest of the three groups.

Figure 3.9

Mean Percentage Response Times for Each Group Within Each Block of the Test



Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

These differences in response times were further analysed with a 3(Group: Experts, Trainees, Novices) x 4(Block: 1, 2, 3, or 4) mixed-model ANOVA of the data. This revealed a main effect of Group, $F(2, 131) = 62.84$, $p < .001$, partial $\eta^2 = 0.489$. A series of independent t tests with Bonferroni correction showed that when response time was collapsed across blocks, Novices ($M = 27.2$s, $SE = 2.42$) responded faster than Examiners ($M = 87.2$s, $SE = 5.78$), $t(131) = 9.58$, $p < .001$, and Trainees ($M = 69.8$s, $SE = 5.33$), $t(131) = 7.27$, $p < .001$.

There was no difference in the response times of Examiners and Trainees, $t(131) = 2.22$, $p = .07$.

There was also an interaction between Group and Block, $F(6, 393) = 24.1$, $p < .001$, partial $\eta^2 = 0.269$, which was explored with a series of independent t tests with Bonferroni correction to compare response times between the groups across each block of the test. A full summary of these post hoc tests is shown in Table 3.6. In summary, across all blocks of the test, Novices responded faster than Experts and Novices, with no difference in the response times of Experts and Trainees.

Table 3.6

Summary of Post Hoc Analyses Comparing Response Times Between Groups, with T-Values and Cohen's D for Each Comparison

Reference and Comparison Group	Block 1	Block 2	Block 3	Block 4
Novices: Experts	t = 8.10* (<) d = 2.44	t = 7.73* (<) d = 2.44	t = 4.81* (<) d = 1.59	t = 9.99* (<) d = 3.26
Novices : Trainees	t = 6.98* (<) d = 1.75	t = 6.89* (<) d = 1.80	t = 4.51* (<) d = 1.10	t = 6.17* (<) d = 1.88
Experts: Trainees	t = 1.26 d = 0.29	t = 1.04 d = 0.21	t = 0.47 d = 0.10	t = 3.37 d = 0.60

Note. Significant comparisons are denoted * $p < .001$. Symbols (> and <) identify whether the reference group is faster or slower than the comparison group.

The interaction was further explored with a series of paired *t* tests with Bonferroni correction to compare differences in Group performance within each block of the test. A full summary of post hoc comparisons is shown in Table 3.7.

Table 3.7

Summary of Post Hoc Analyses Comparing Response Times Within Groups Across Blocks, , with T-Values and Cohen's D for Each Comparison

Reference and Comparison Block	Novices	Trainees	Experts
Block 1 : Block 2	*t* = 6.69** (<) *d* = 0.89	*t* = 5.83** (<) *d* = 0.77	*t* = 6.23** (<) *d* = 1.14
Block 1 : Block 3	*t* = 2.86 *d* = 0.41	*t* = 0.03 *d* = 0.00	*t* = 0.74 *d* = 0.16
Block 1 : Block 4	*t* = 1.41 *d* = 0.32	*t* = 2.21 *d* = 0.32	*t* = 6.05** (>) *d* = 0.92
Block 2 : Block 3	*t* = 8.88** (>) *d* = 1.29	*t* = 4.88** (>) *d* = 0.61	*t* = 4.32** (>) *d* = 0.89
Block 2 : Block 4	*t* = 2.36 *d* = 0.58	*t* = 6.17** (>) *d* = 0.78	*t* = 10.96 ** (>) *d* = 1.42
Block 3 : Block 4	*t* = 3.97* (>) *d* = 0.73	*t* = 2.52 *d* = 0.29	*t* = 7.56 ** (<) *d* = 1.01

Note. Significant comparisons are denoted *, p < .01, ** *p* < .001. Symbols (> and <) identify whether response times in the reference block are faster or slower than the comparison block.

In summary, Novices completed all blocks of the test faster than Trainees and Experts, with no differences emerging in the response times of Trainees and Experts. All groups completed fingerprint image matching in Block 2 more quickly any other blocks of the test, and there were no group differences in the response times between pattern matching in Block 1 and palmprint matching in Block 3. Latent print matching in Block 4 was designed as the most challenging section of the fingerprint test. It was here that clear differences in the response times of the three groups emerged, with Experts taking longer to complete these tasks than they had in previous blocks. In contrast, there was no difference in the response times of Trainees between these blocks, and Novices responded more quickly during latent print matching than they had during palmprint matching.

General Discussion

This chapter examined fingerprint expertise by comparing novices, trainees in fingerprint analysis, and experienced fingerprint examiners on the fingerprint test detailed in Chapter 2. The test comprises of four tasks, which allow for the examination of expertise in pattern matching, fingerprint image matching, palmprint matching, and latent print matching. By comparing Novices, Trainees, and Experts with these different types of fingerprints, the aim was to identify which of these tasks specifically differentiates the performance of fingerprint examiners from observers who do not have similar experience and professional status. In contrast to previous work (e.g., Searston & Tangen, 2017a, 2017b; Stevenage & Pitfield, 2016; Tangen et al., 2011; Thompson et al., 2014), this was examined by comparing mean performance for these observer groups *and* individual differences in fingerprint identification ability.

In the pattern matching tasks of Block 1, Trainees and Experts outperformed Novices at a group level, but there was also considerable overlap in the performance between individuals of all groups. This pattern was replicated with fingerprint image matching in Block 2 and palmprint matching in Block 3. Here, the group-level differences between

Trainees and Examiners compared to Novices became more pronounced. However, fingerprint image matching and palmprint matching could not reliably distinguish Trainees from Examiners, both at a group and individual level. Moreover, substantial overlap remained between individuals in the three observer groups, whereby many of the Novices performed at the level of the Trainees and Experts in both tasks (See Figure 3.6). Thus, neither pattern matching, fingerprint image matching nor palmprint matching provide a clear index of fingerprint examiner expertise in this chapter. More broadly, these results show that group-level performance with pattern, fingerprint image, and palmprint matching is an imprecise means to measure expertise. In addition, the use of individual-level metrics are essential to contextualise this type of expertise appropriately. This supports findings from a range of perceptual tasks such as face identification (Lander et al., 2018; White et al., 2014), object recognition (Gauthier, 2018), mammogram interpretation (Hornsby & Love, 2014), and forensic firearm examination (Mattijssen et al., 2021), that focusing on the average performance of groups of observers can conceal a wide range of individual differences in ability.

Importantly, performance in the latent print matching trials in Block 4 produced a markedly different pattern by clearly distinguishing Experts from Trainees. This was evident at a group level but also corroborated by the individual differences analysis which showed that only a very small proportion of Trainees (less than three percent, see Table 3.5) can perform within the Expert range. Unsurprisingly, there was more overlap in the individual performance of Novices and Trainees, all of whom are in the process of acquiring fingerprint knowledge during their training. However, many Trainees still fell outside of the Expert range (see Figure 3.6). Thus, these results demonstrate that the true nature of fingerprint expertise lies in the ability to deal with the most challenging of fingerprints – latent prints of the type that may have been recovered from crime scenes – and this expertise is such that it distinguishes the experienced examiners from almost any novice observer as well as many fingerprint examiners in training.

The fingerprint test also provides some hints as to what might underline the superior examiner performance in latent print matching. The response times show, for example, that Experts were considerably slower than Novices in making their decisions, suggesting a more analytical approach to fingerprint comparison. This may reflect workplace procedures such as ACE which is characterised by a slow, methodical, and iterative process in which pairs of fingerprints are analysed, compared, and evaluated (FSR, 2017; SWGFAST, 2013). However, Trainees also exhibited slower response times than Novices without fully matching the accuracy of the Experts, which indicates that additional factors may be contributing to the performance of Experts.

Previous research has found that experience with fingerprints allows examiners to identify and localise relevant features in a fingerprint (Hicklin et al., 2019), with workplace comparisons providing many opportunities for examiners to gain exposure to a wide range of fingerprint features. Therefore, if familiarity with a class of stimuli underpins the development of proficiency (Thompson et al., 2014), this is likely to be a factor in the accuracy advantage shown by the Experts in latent print matching. In addition, fingerprint examiners have an ability to identify more areas of diagnostic value in a fingerprint than less experienced observers (Busey et al., 2010), which may prove advantageous when viewing latent prints in which features can be unclear. The challenge of identifying these distorted or unclear prints may therefore serve to bring the expertise of the experienced examiners to the fore compared to observers with no, or some, training in fingerprint identification.

In addition, differences between the groups also emerged in the current study in relation to errors committed in target-present comparisons, whereby Experts and Trainees were more likely than Novices to fail to identify a target during latent print matching. This converges with previous work (Busey et al., 2010; Tangen et al., 2011; Thompson et al., 2014; Ulery et al., 2011) and may reflect a more cautious approach to these ambiguous fingerprint comparisons. This may be based on workplace practices and training that emphasises the suitability and quality of a latent print for comparison (SWGFAST, 2013; Vanderkolk, 2011) in an attempt to avoid false positives in identification.

Overall, this chapter demonstrates that, at an *individual level*, many completely untrained and inexperienced observers can demonstrate considerable competence in fingerprint comparison tasks such as pattern, fingerprint image, and palmprint matching. In turn, a *group-level* advantage is also found for Experts and Trainees in these tasks. This could suggest that some of the easier aspects of fingerprint comparison may be underpinned by a domain-general feature comparison ability (Growns et al., 2022). With latent prints, however, Experts demonstrate a clear and consistent accuracy advantage over virtually all Novices, pointing to specific, experienced-based expertise for difficult fingerprint comparisons (Searston et al., 2017b). These insights emerged with a research methodology that combined four different fingerprint tasks with a group-level and individual differences approach to data analysis.

Chapter 4

Understanding Fingerprint Comparison Expertise

4.1 Introduction

The previous chapters outlined the construction of a novel online fingerprint aptitude test and its subsequent use in comparing the performance of police forensic fingerprint examiners ('Experts'), trainee fingerprint examiners ('Trainees'), and untrained observers ('Novices'). The test comprised of four blocks of different tasks relevant to fingerprint comparison: pattern matching, fingerprint image matching, palmprint matching and latent print matching. In terms of overall performance, Experts and Trainees were more accurate than Novices. Comparing accuracy in each block showed that Experts outperformed Novices, and Trainees were more accurate than Novices in every task except palmprint matching. Most importantly, a clear accuracy advantage was demonstrated by Experts over both Novices *and* Trainees during latent print matching. This was designed as the most challenging block of the test, with stimuli representative of those encountered in case work. Finally, comparing the performance of *individual* Novices with that of Experts and Trainees revealed that high accuracy in such fingerprint comparison is likely to be rare within the untrained population. Overall, the differences in the performance of the three groups therefore demonstrated fingerprint comparison expertise, and the development of expertise, within the fingerprint examiners.

Throughout the test the response times of the Experts and Trainees were also considerably slower than those of the Novices, suggesting a more methodical approach to the task. However, although the fingerprint aptitude test captured differences in performance between Experts, Trainees, and observers untrained in fingerprint comparison, it could not provide insight into the underlying cognitive processes and visual perceptual skills that accompany fingerprint comparison expertise. Forensic fingerprint examination is a complex cognitive and perceptual task. Fingerprints are comprised of an infinite range of features such as patterns, ridge endings and bifurcations, pores, and creases (Champod et al., 2004). Examiners compare these features across pairs of fingerprints to identify points of similarity, or disagreement, to determine whether they are from the same or different sources (CPS, 2017). Studies have found that experienced examiners outperform untrained observers in fingerprint comparison tasks under experimental conditions (Searston & Tangen, 2017a; 2017b; Stevenage & Pitfield, 2016; Tangen et al., 2011; Thompson et al., 2014), however, much remains unknown about the cognitive processes that may underpin this ability.

Previous research suggests that memory may be a key cognitive component of fingerprint expertise, with examiners' performance less affected by a delay between target presentation and a subsequent paired matching task than novices (Busey & Vanderkolk, 2005; Thompson & Tangen, 2014). Although examiners typically view prints in a side-by-side format, an ability to retain relevant fingerprint information in short-term memory is undoubtedly needed in order to make the visual comparison between images. Research has also found that compared to novices, examiners tend to focus more on the areas of a fingerprint with high diagnostic value (Busey et al., 2011) and are quicker to identify two corresponding areas in a pair of prints (Hicklin et al., 2019). This suggests that examiners can more readily identify salient features in a fingerprint, which they retain in short-term memory, and use to locate corresponding, or discrepant, features in the comparison image.

Forensic fingerprint examination may also require an ability to mentally rotate images. Although prints are correctly oriented prior to examination, crime scene prints may be distorted or impaired by substrate interference thereby necessitating comparison across different viewing planes. Research has found that examiner accuracy is less affected than novices when viewing inverted prints (Thompson et al., 2014), which suggests that examiners may have enhanced mental rotation abilities.

On the other hand, in the domain of face perception, holistic processing is believed to develop through experience with the class of stimuli and is disrupted when the image is inverted (e.g., Farah, Tanaka & Drain, 1995; Maurer, Le Grand & Mondloch, 2002). The findings from Thompson et al. (2014) therefore tend to suggest that fingerprints are not processed holistically by examiners. However, anecdotally examiners report that they perceive fingerprints holistically rather than by features (Lisa Hall, personal correspondence). This is also partly supported by research which found that fingerprint matching accuracy declined when expert examiners viewed inverted composite images in which half of the image was of the target and half was from a different fingerprint (Vogelsang et al., 2017. These results therefore suggest that the incongruent part of the image disrupted experts' holistic processing of the target part of the image. However, there had been no differences in the accuracy of novices and experts when the images were inverted or misaligned. Therefore, the authors concluded that the experiment only provides weak evidence that holistic processing of fingerprints underlies the expertise of fingerprint examiners.

The absence of strong evidence in support of holistic processing of fingerprints by examiners may in part be due to the feature-to-feature nature of fingerprint comparison. Although fingerprint patterns are broadly categorized as loops, whorls, and arches, they also contain a wealth of minutiae such as ridges, valleys, bifurcations, lines, pores, and scars which are often only visible under magnification (Bigun, 2014, Champod et al., 2004). It is through examination of spatial relations between these features that examiners decide whether two fingerprints are from the same or different sources (Home Office, 2017). Close attention to the detailed structure of a fingerprint is therefore a necessary component of forensic fingerprint examination. This suggests that examiners are more likely than novices to process fingerprints through attention to the local details contained within a print rather than by holistic processing of the whole, or global, image. To date, the relationship between local and holistic processing ability and fingerprint comparison has not been fully explored.

Research has therefore hinted at some of the cognitive processes that underpin fingerprint expertise: (i) the ability to extract relevant features with which to make a comparison, (ii) the retention of this featural information in short-term memory, and (iii) to accurately process incorrectly oriented fingerprints. In addition, holistic processing of fingerprints by examiners may form a facet of expertise but the results are currently inconclusive. These abilities have so far been tested to a limited degree using fingerprint stimuli. And whilst the basic performance data (i.e., accuracy) that is typically reported may reflect cognitive skills within the examiners' domain of experience, it does not shed light on the nature of these abilities and whether they are more enhanced in fingerprint examiners than untrained and inexperienced observers.

In this chapter, the cognitive abilities of experienced forensic fingerprint examiners ('Experts'), fingerprint examiners in training ('Trainees'), and untrained observers ('Novices') will be compared in a series of tests using non fingerprint stimuli to address this issue. The use of test batteries to identify cognitive and perceptual abilities is well established in the field of cognitive psychology (e.g., Bate et al., 2018; Burton et al., 2010; Davis et al., 2016; Wilhem et al., 2014), but has not yet been applied to understand fingerprint identification.

The current test battery is designed to measure some of the key cognitive components of fingerprint comparison: visual short-term memory, feature comparison, visual search, mental rotation, and holistic and local processing of stimuli. Due to the analytical nature of fingerprint examination, with the focus on feature identification and comparison, it could be predicted that performance in these perceptual tasks will be generally higher in forensic examiners. For this reason, a face matching test has also been included in the battery. Fingerprint examiners have previously been found to outperform novices in this task although it is one which is outside of their area of expertise (Phillipe et al., 2018). The results will therefore shed light on the relationship between feature comparison and fingerprint examination and will explore whether a broader ability in feature matching tasks could underpin fingerprint comparison accuracy. Finally, as police staff will be taking the test battery in their occupational role as forensic examiners, they may be expected to have greater motivation to perform well than novice observers who may regard these as lower stakes tests. To examine the relationship between motivation and fingerprint comparison accuracy, observers will therefore also rate their level of intrinsic motivation with a self-report scale at the conclusion of the test battery.

The results from the test battery will be used to predict accuracy in each block of the fingerprint test. If any of the tests in this battery capture difference in cognitive abilities between police forensic examiners and untrained observers, then performance should be positively associated with accuracy in the fingerprint aptitude test. This will provide insight into the cognitive abilities that underpin accuracy in fingerprint comparison.

4.2 Method

Participants

Seventeen qualified fingerprint examiners (‘Experts’) from a UK police service took part in this experiment (mean age = 46.41 years, *SD* = 9.30, range 27-59 years, 6 males), with mean experience in fingerprint examination of 20.47 years (*SD* = 10.80, range 2-36 years). Fifteen trainee examiners (‘Trainees’) from the same UK police service (mean age = 27.06

years, *SD* = 6.87, range = 21-48) with mean training in fingerprint examination of 16.5 weeks (*SD* =16.11, range 1-52 weeks) also took part. All police employees undertook the experiment online whilst they were working from home. A further group of ninety-five participants who were untrained in fingerprint examination (mean age = 20.30 years, *SD* = 4.62, range 18-58, 21 males) participated as 'Novices', and undertook the experiment online. This group predominantly comprised university students who participated in return for course credits, and the remainder were volunteers who received no remuneration for their participation. All participants reported normal or corrected-to-normal eyesight and provided informed consent to take part. This research was approved by the University of Kent Ethics Committee.

Procedure

This experiment was conducted in one online test session using Inquisit software. Participants were able to take screen breaks between each test if needed. Prior to commencing each test, participants were provided with instructions to calibrate their computer screen. This was done by placing a credit card against an onscreen template of a standard sized credit card (85.6mm x 53.98mm = 323.5 x 204.01 pixels at a screen resolution of 127 ppi) and adjusting the browser magnification until the card and template matched. All participants took the tests in the same order, and responses were entered using the keyboard.

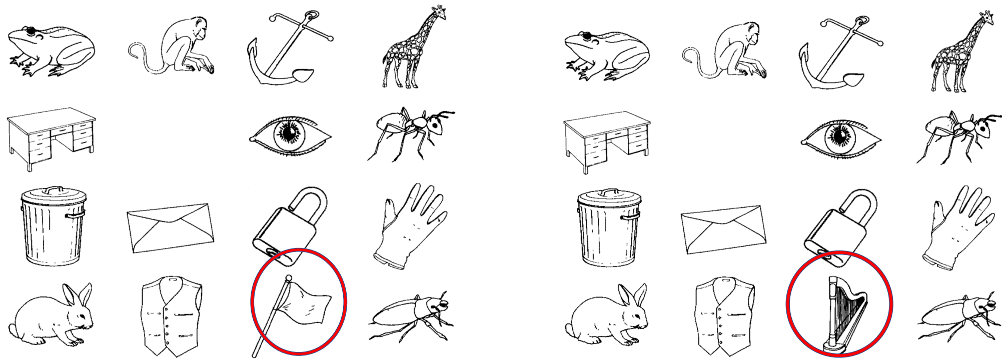
Test Battery

Visual Short-Term Memory (VSTM) Test: This test was obtained from the Millisecond test library (Borchert, 2020) and is based on the paradigm described in research by Beck and Levin (2003). The test comprised stimulus arrays of everyday objects from the Snodgrass and Vanderwort (1980) image set, such as animals, clothing, furniture, and household objects. These were displayed in arrays of 3 to 16 items. Each target and array measured 1133(w) x 944(h) pixels at a screen resolution of 127 ppi.

During the test, participants were presented with a 'pre-change' array (2000ms) followed by a 'post-change' array (2000ms) in which one of the objects has been replaced. An example from the test is shown in Figure 4.1. Onscreen images shifted to either the left or right by 10% between pre- and post-change arrays. The same arrays were then presented separately, with unlimited viewing times, and observers needed to identify the pre-change and post-change objects using the computer mouse to select the correct items on the screen. The experiment comprised 10 practice trials with feedback using arrays of 4 objects. This was followed by arrays of 3 to 16 objects with 4 repetitions of each array size, presented in random order and with 56 trials in total. Short term memory accuracy was calculated in each array size for those trials in which both the pre-change *and* post-change item was correctly identified.

Figure 4.1

Example of a 15-Object Array from the Visual Short-Term Memory (VSTM) Test Showing Pre-change (left) and Post-change (right) Screens

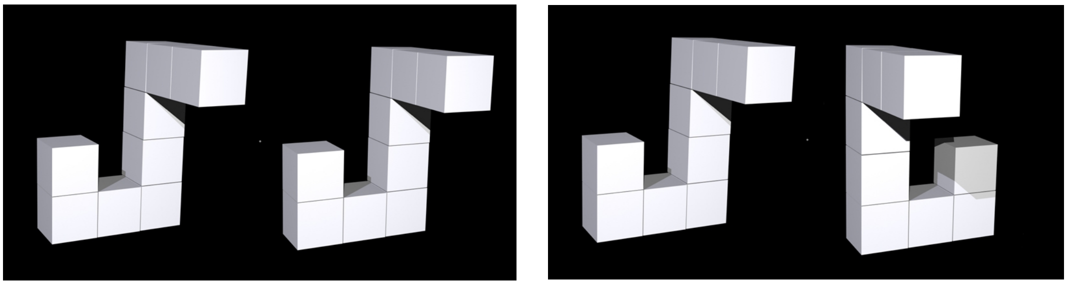


Note. Pre- and post-change array items are encircled.

Mental Rotation Task: This task was obtained from the Millisecond test library (Borchert, 2020) and used stimuli and test procedure from research by Ganis and Kievit (2015). Participants were presented with two 3D cube-based objects which they needed to identify as the same or different. The stimuli comprised twelve different cube objects, with the right-hand object displayed in four different degrees of rotation from the vertical plane (0^0, 50^0, 100^0, 150^0), and equal numbers of same or different pairs. Each stimuli pair measured 585(w) x 245(h) pixels at a screen resolution of 127 ppi. An example of test stimuli is shown in Figure 4.2. Stimuli were displayed onscreen for a maximum duration of 7500ms, with a 'no response detected' warning if this was exceeded without a response. Using the computer keyboard, participants entered 'S' to indicate that both pairs were the same, and 'D' to indicate that both pairs were different. The experiment comprised 16 practice trials with feedback followed by 96 test trials, and a short break after 48 trials.

Figure 4.2

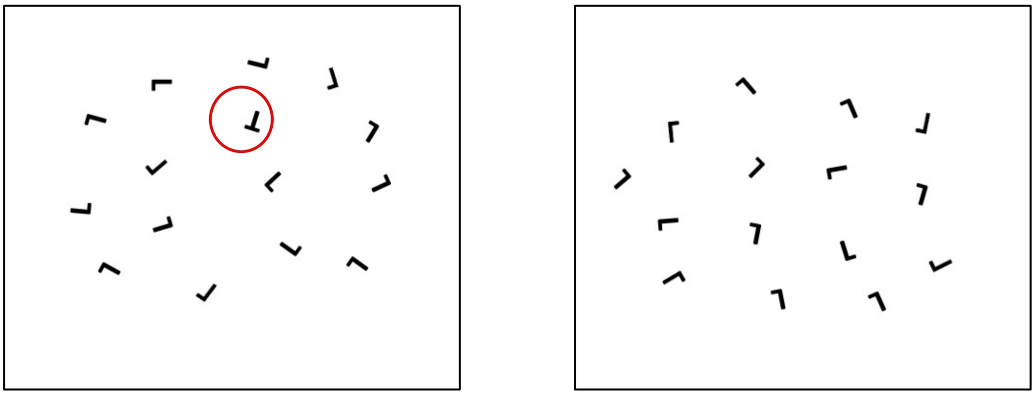
Example of Stimuli from the Mental Rotation Task Showing a 'Same' Pair of 3D Cube Objects (left) and a 'Different' Pair of 3D Objects with Zero Degrees of Rotation (right)



Visual Search Task: In this task the observer needed to identify whether there was a target letter 'T' within an array constructed from the letter 'L' and was based on a design used by Duncan and Humphreys (1989). Both the target and array letters could be presented upright, inverted or at displayed an angle. Stimuli comprising black letters on a white background were created for this experiment, and arrays contained 1, 5, 15 or 30 letters. The maximum size of an array was 566(w) x 264(h) at a screen resolution of 127 ppi. Arrays were presented in a random order, with 20 trials per array size and equal numbers of target absent and target present trials, resulting in 160 trials. Example stimuli are shown in Figure 4.3. Using the computer keyboard, participants entered 'P' to indicate the target was present, and 'A' to indicate the target was absent. Viewing and response times were self-paced.

Figure 4.3

Examples of 15-Item Arrays from the Visual Search Task Showing a Target Present (left) and a Target Absent (right) Trial



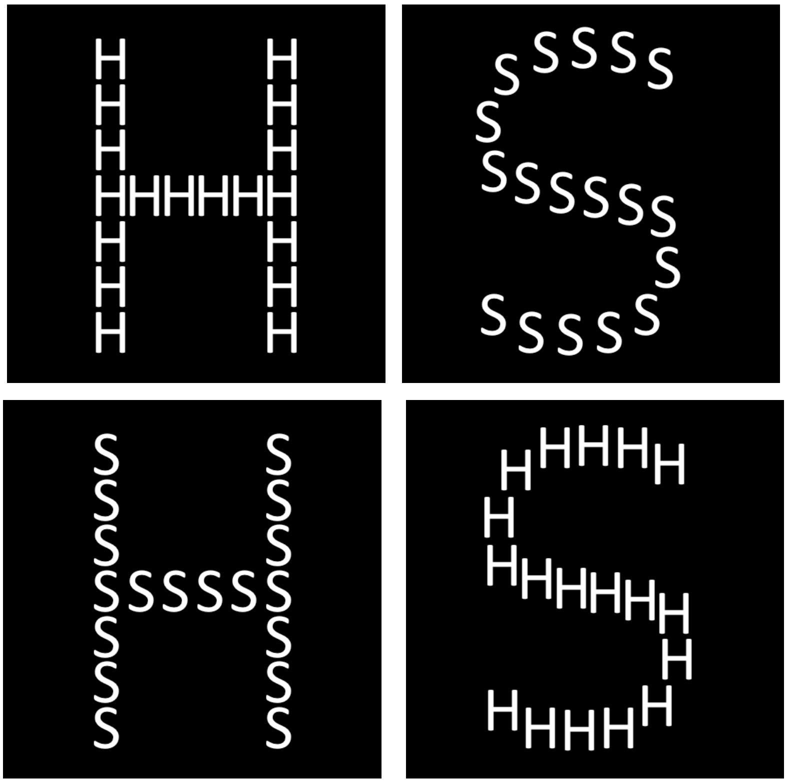
<u>Navon Letters Test:</u> This test was obtained from the Millisecond test library (Borchert, 2020) with the procedure based on that reported in Navon (1977). Stimuli comprised large letter shapes (H or S) constructed from smaller letters (H or S). Each letter shape measured 245(w) x 245(h) pixels at a screen resolution of 127 ppi. Stimuli were either congruent in which the large and small letters were the same, or incongruent whereby large

and small letters were different, thereby allowing for four different combinations of stimuli (see Figure 4.4).

A letter shape was presented to participants for 100ms and appeared in random order in one of the four screen quadrants to avoid repeated fixation on the same area of the screen. The test comprised four blocks of 48 trials equal numbers of congruent and incongruent trials. In two blocks participants needed to identify the large letter (global condition) and in the other two blocks they needed to identify the small letters (local condition). Participants entered their responses using the keyboard to indicate whether they saw the letter ‘H’ or ‘S’ in each condition.

Figure 4.4

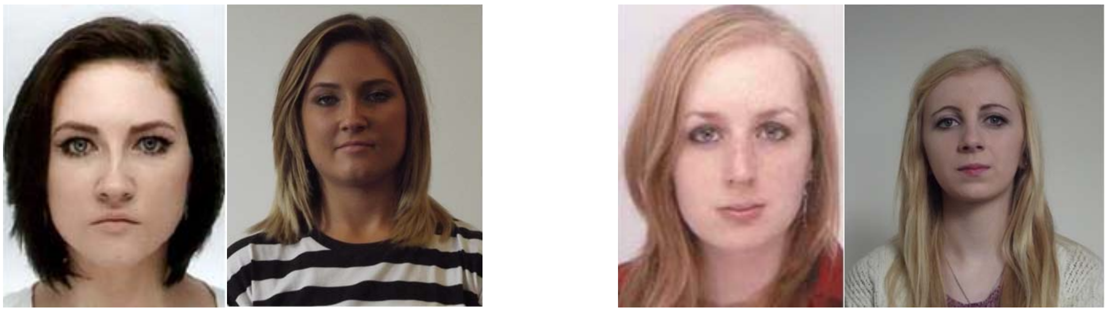
Examples of Navon letters Depicting Congruent (top) and Incongruent Stimuli (bottom)



Kent Face Matching Test: This test used forty face pairs from the KFMT (Fysh & Bindemann, 2018). In each face pair, the image displayed on the right is of the subject with a neutral facial expression, taken under controlled conditions against a plain background and with even illumination (scaled to 302 (w) x 359 (h) pixels at a screen resolution of 127 ppi). The image displayed on the left comprised an unconstrained image from a student ID photograph (scaled to 207(w) x 151(h) pixels at a screen resolution of 127 ppi). The test comprised of twenty pairs of images in which the left and right images were of the same person (match), and twenty pairs of images in which the left and right images were of different people (mismatch). Example stimuli are shown in Figure 4.5. Each face pair was displayed onscreen in a random order and participants needed to identify whether they depicted the same person or two different people by pressing ‘S’ or ‘D’ on the keyboard. Viewing and response times were unrestricted in this test.

Figure 4.5

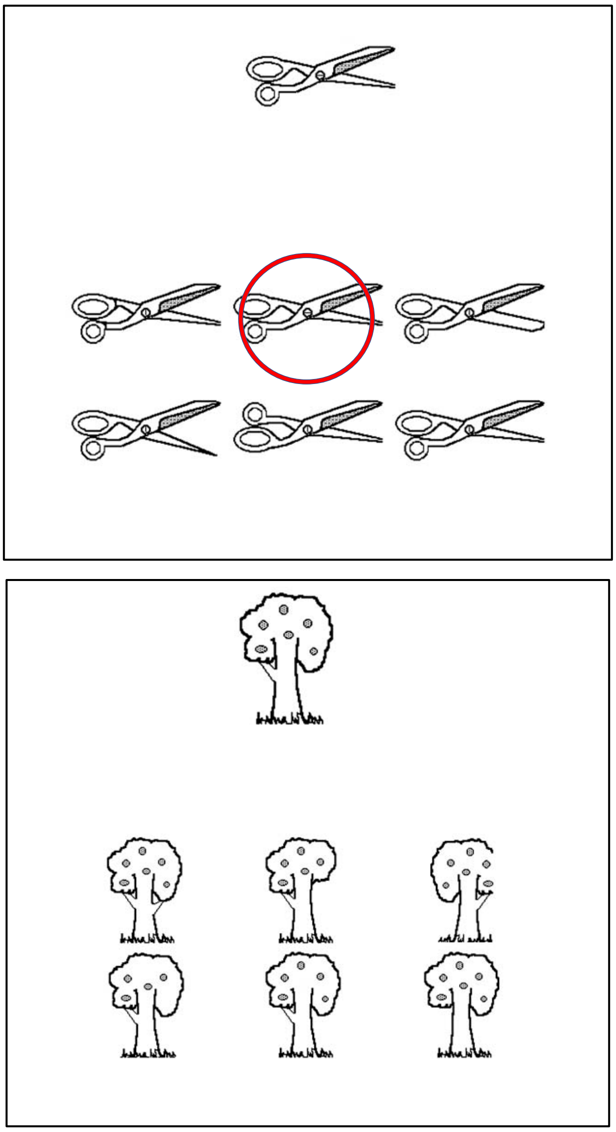
Examples of a Match (left) and Mismatch (right) Face Pair from the KFMT



Matching Familiar Figures Test: The MFFT uses black on white line drawings of familiar objects from Kagan (1965). In each trial participants needed to identify whether the target object displayed at the top of the screen was among the six exemplars in the array below. Each target and array measured 567(w) x 567(h) pixels at a screen resolution of 127 ppi. If the target was present participants entered the corresponding array number using the computer keyboard and entered the letter ‘A’ if the target was absent. There were 48 trials in total of equal numbers of present and absent trials, with order maintained for each participant. Example stimuli are shown in Figure 4.6. Viewing and response times were unrestricted in this test.

Figure 4.6

Example of Target Present (top) and Target Absent (bottom) Stimuli from the MFFT



The Intrinsic Motivation Inventory (IMI): This part of the test battery was conducted using Qualtrics software. The IMI[2] (Deci et al., 1994) is a 22-item inventory to measure subjective motivation to perform a particular task, in this case the preceding battery of tests. It contains four sub-scales: interest/enjoyment, perceived competence, perceived choice, pressure/tension. All items comprising the IMI are shown in Table 1. Observers entered their responses on the keyboard using a Likert scale from 1 (not true at all) to 7 (very true) and could miss any questions they did not wish to answer.

Table 4.1

The Four Sub-scales and Question Items Comprising the IMI

Interest/Enjoyment	Perceived Competence	Perceived Choice	Pressure/Tension
While I was working on the tests, I was thinking about how much I enjoyed them.	I think I am pretty good at these tests.	I felt that it was my choice to do the tests.	**I did not feel at all nervous about doing the tests.**
I found the tests very interesting.	I think I did pretty well at this activity compared to other participants.	**I didn't really have a choice about doing the tests.**	I felt tense while doing the tests.
Doing the tests was fun.	I am satisfied with my performance in these tests.	I felt like I was doing what I wanted to do while I was working on the tests.	**I felt relaxed while doing the tests.**
I enjoyed doing the tests very much.	I felt pretty skilled at these tests.	**I felt like I had to do the tests.**	I was anxious while doing the tests.
I thought the tests were very boring.	After working at these tests for a while I felt pretty competent.	**I did the tests because I had no choice**	I felt pressured while doing the tests.
I thought the tests were very interesting.			
I would describe the tests as very enjoyable.			

Note. Items in bold font are reverse scored.

[2] Available from https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

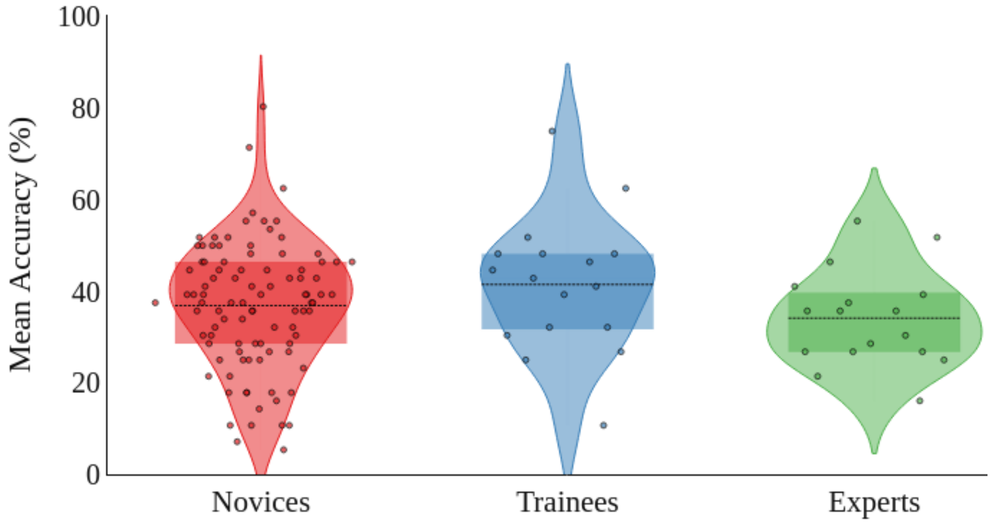
4.3 Results

The analyses focused on examining differences between Novices, Trainees and Experts. For this reason, only group differences are shown here, and main effects and interactions of factors other than by group are reported in the Appendix.

Visual Short-Term Memory (VSTM) Test: In this test, observers needed to correctly identify both the pre-change object *and* the post-change object within an array in each trial. Therefore, mean accuracy reflected the correct identification of both objects and was calculated by collapsing data across all sizes of array. As can be seen from Figure 4.7, Trainees appeared to be the most accurate group, with similar performance by Novices and Experts.

Figure 4.7

Mean Accuracy (%) of Novices, Trainees and Experts in the VSTM Test

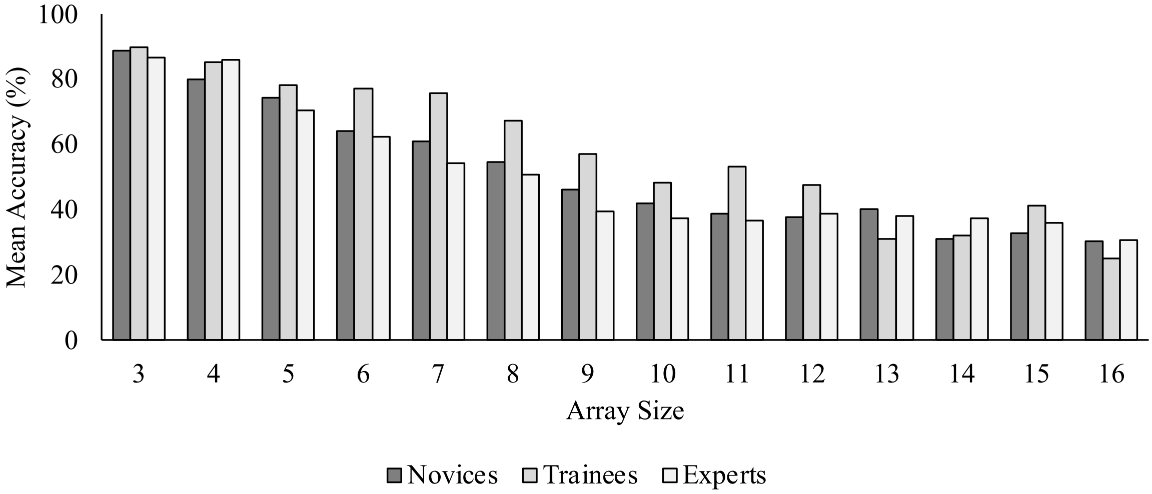


Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

A one-way ANOVA to compare accuracy between Novices, Trainees and Experts revealed no differences between the groups, $F(2,124) = 2.35$, $p = 0.10$, partial $\eta^2 = .021$. To examine whether differences in accuracy emerged between the groups across different array sizes of the VSTM task, the mean overall accuracy for each array size was calculated for each group. This is shown in Figure 4.8 and reflects a general decline in accuracy for all groups as the number of items in the array increased.

Figure 4.8

Mean Accuracy of Novices, Trainees, and Experts by Array Size in the VSTM Test

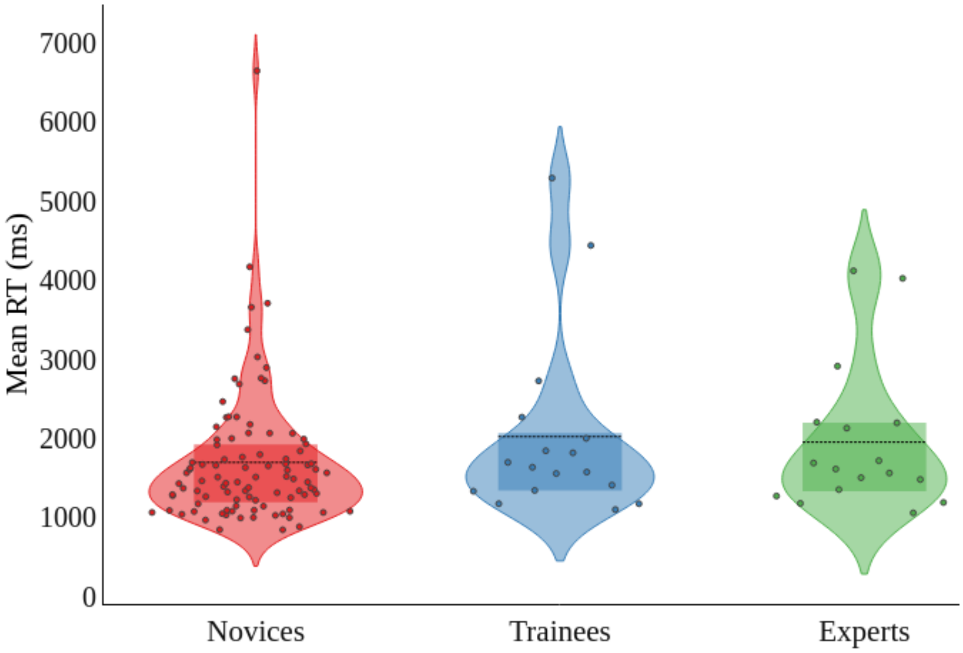


Group mean accuracy data was then split to allow comparisons of performance between smaller (3 to 9) and larger (10 to 16) array sizes. This data was analysed with a 3 (Group: Novices, Trainees, Experts) x 2 (Array Size: 3 to 9, 10 to 16) mixed-model ANOVA to compare performance between the groups by array size. This revealed a main effect of Array Size, $F(1, 18) = 139.69$, $p < .001$, partial $\eta^2 = 0.88$, due to higher accuracy in smaller arrays ($M = 69.1$, $SE = 3.23$) than in larger arrays ($M = 37.5$, $SE = 1.49$). There was no difference in accuracy between Groups, $F(2, 18) = 1.15$, $p = .34$, partial $\eta^2 = 0.11$, and no interaction between Group and Array Size, $F(2, 18) = 0.78$, $p = .47$, partial $\eta^2 = 0.08$.

The mean response times of observers were then calculated, with data collapsed across all array sizes. Figure 4.9 shows that Novices appear to be quicker to respond than Experts and Trainees, with no apparent difference between the fingerprint examiner groups. However, a one-way ANOVA revealed no reliable differences in response times between all three groups $F(2,124) = 1.57$, $p = 0.21$.

Figure 4.9

A Comparison of Mean Response Times of Novices, Trainees and Experts in the VSTM Test



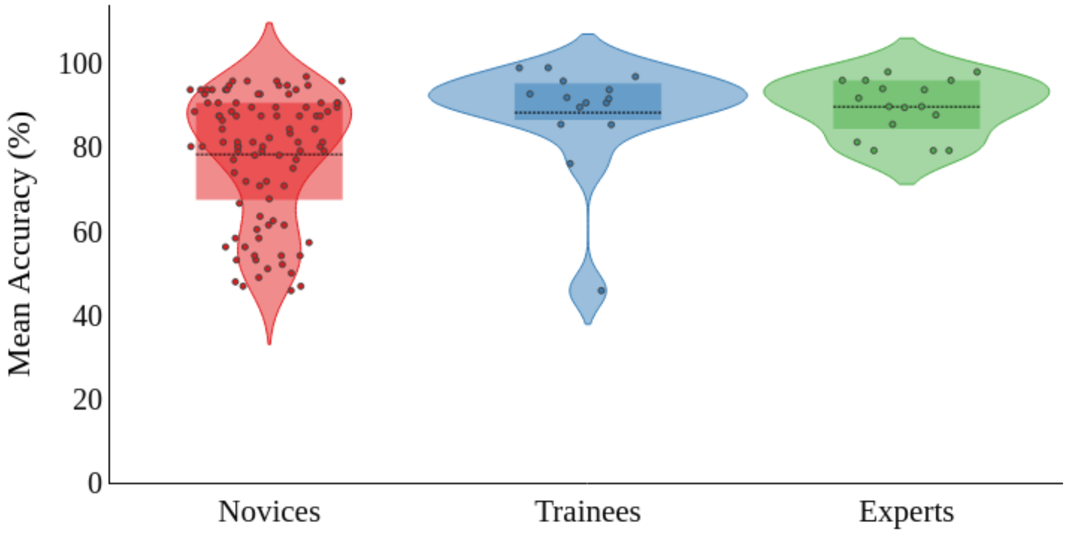
Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

In summary, this test had been designed to capture the observer's ability to recall two objects: the object that was replaced in the first array and the object it had been replaced with in the subsequent array. The size of the arrays varied between three and sixteen objects, and accuracy in all groups declined as the size of the array increased. Analyses revealed no differences in either the speed of responses or accuracy between the three groups. Overall mean accuracy in each group was around forty percent, thereby suggesting that most observers found this to be a challenging test.

Mental Rotation Test (MRT): In this test, observers viewed pairs of 3D images that were either the same or different, with the right-hand image displayed across four degrees of rotation from the vertical plane (0^0, 50^0, 100^0, 150^0).

Figure 4.10

A Comparison of Mean Accuracy of Novices, Trainees and Experts in the MRT



Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

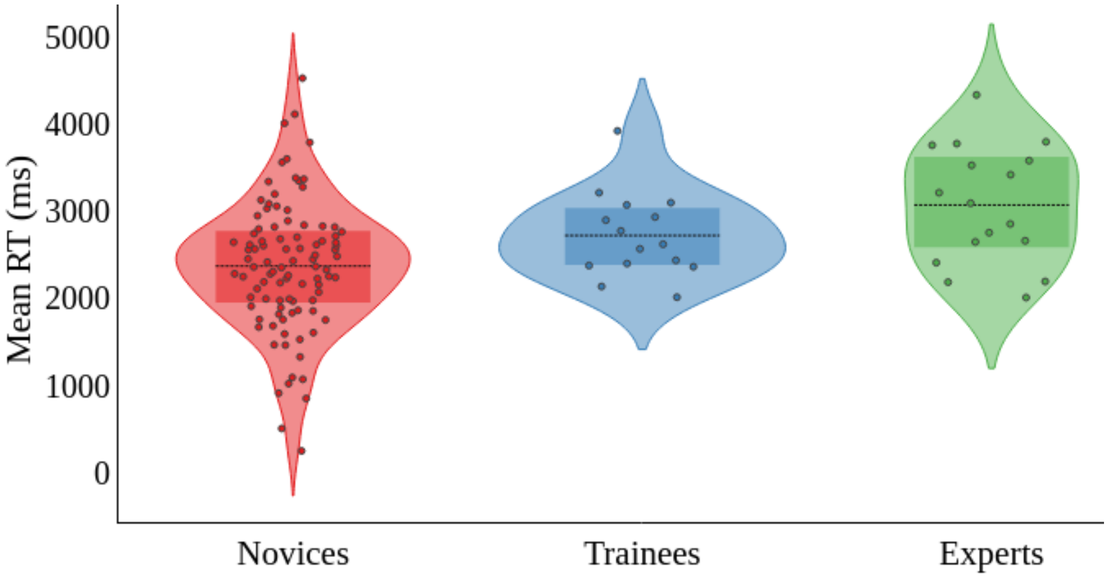
To compare accuracy between Novices, Trainees and Experts, the mean percentage accuracy of each group was first calculated for same and different trials across each of the four rotations. This data was then analysed with a 3 (Group: Novices, Trainees, Experts) x 2 (Trial Type: Same or Different) x 4 (Rotation: 0^0, 50^0, 100^0, 150^0) mixed-model ANOVA. This revealed a main effect of Group, $F(2,124) = 7.07$ $p < .001$, partial $\eta^2 = 0.102$, which is shown in Figure 4.10. This data was further explored with a series of independent *t*-tests with Bonferroni correction. These showed that Novices ($M = 78.5$, $SE = 1.43$) were less accurate than Experts ($M = 90.0$, $SE = 3.43$), $t(124) = 3.12$, $p = .007$, and Trainees ($M = 88.3$, $SE = 3.66$), $t(124) = 2.53$, $p = .01$. There was no difference in the accuracy of Experts and Trainees, $t(124) = 0.34$, $p = 1.00$, and no interaction between Group and Trial Type, $F(2, 124) = 0.21$, $p = .81$, partial $\eta^2 = 0.003$, Group and Rotation, $F(6, 372) = 0.80$, $p = 0.57$, partial $\eta^2 = 0.01$, or Group by Trial Type and Rotation, $F(6, 372) = 1.84$, $p = 0.09$, partial $\eta^2 = 0.03$.

Mean response times for each group were then calculated for same and different trials across the four rotations. This data was analysed with a 3 (Group: Novices, Trainees, Experts) x 2 (Trial Type: Same or Different) x 4 (Rotation: 0^0, 50^0, 100^0, 150^0) mixed-model ANOVA and revealed a main effect of Group, $F(2,123) = 7.78$, $p < .001$, partial $\eta^2 = 0.112$, which is shown in Figure 4.11. This was further explored with a series of independent *t*-tests with Bonferroni correction. These showed that Novices ($M = 2367$, $SE = 73.5$) were faster than Examiners ($M = 3069$, $SE = 173.2$), $t(123) = 3.74$, $p < .001$, but not Trainees ($M = 2719$, $SE = 174.4$), $t(123) = 1.78$, $p = .23$. There was no difference in the response times of Experts and Trainees, $t(123) = 1.39$,$p = .51$, and there were no interactions between Group and Trial Type, $F(2, 123) = 0.80$, $p = .45$, partial $\eta^2 = 0.01$, Group and Rotation, $F(6, 369) = 2.03$, $p =$

.06, partial η^2 = 0.03, or Group by Trial Type by Rotation, F(6, 369) = 2.15, p = .05, partial η^2 = 0.03.

Figure 4.11

A Comparison of Mean Response Times of Novices, Trainees and Experts in the MRT



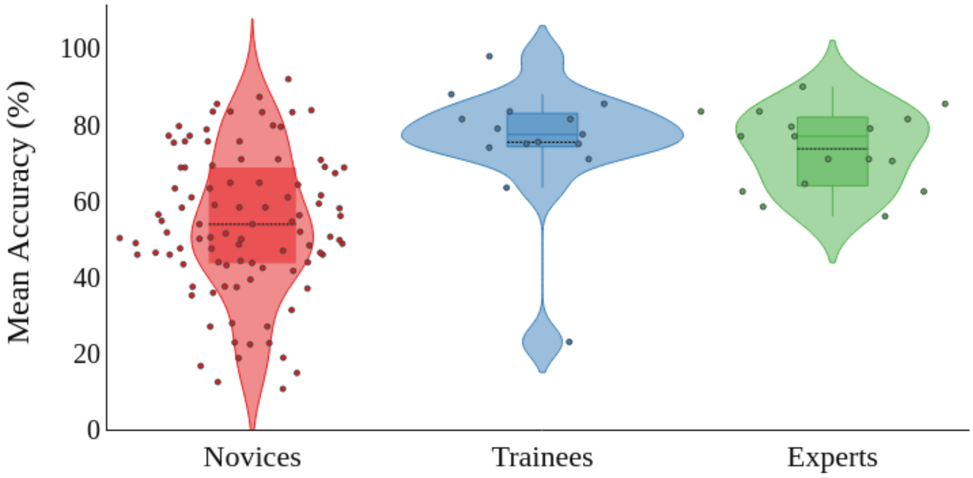
Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

In summary, Experts and Trainees were more accurate than Novices across the test, with performance not related to the type of trial or the degrees of rotation. Novices were faster to respond than Experts, but not Trainees. There was no difference between the performance of Experts and Trainees in this test. Mean accuracy in all groups was high (> 78%), thereby indicating that observers as a group did not find the test to be overly difficult. However, several Novices and one Trainee scored below 50%, which suggests that some individuals found this to be a challenging task.

Matching Familiar Figures Test (MFFT): In this test observers viewed equal numbers of target-present and target-absent arrays in which they needed to identify the target figure. To compare the accuracy of Novices, Trainees and Experts, mean percentage accuracy was calculated for each group by trial type and analysed with a 3 (Group: Novices, Trainees, Experts) x 2 (Trial Type: Present or Absent) mixed-model ANOVA. This revealed a main effect of Group, $F(2, 124) = 16.7$, $p < .001$, partial $\eta^2 = 0.212$ (shown in Figure 4.12). This data was further analysed with a series of independent t-tests with Bonferroni correction, which identified that Novices ($M = 53.5$, $SE = 1.83$) were less accurate than Experts ($M = 73.7$, $SE = 4.32$), $t(124) = 4.31$, $p < .001$, and Trainees ($M = 75.4$, $SE = 4.60$), $t(124) = 4.44$, $p < .001$. There was no difference in the accuracy of Experts and Trainees, $p > .05$.

Figure 4.12

Comparison of the Mean Percentage Accuracy of Novices, Trainees and Experts in the MFFT



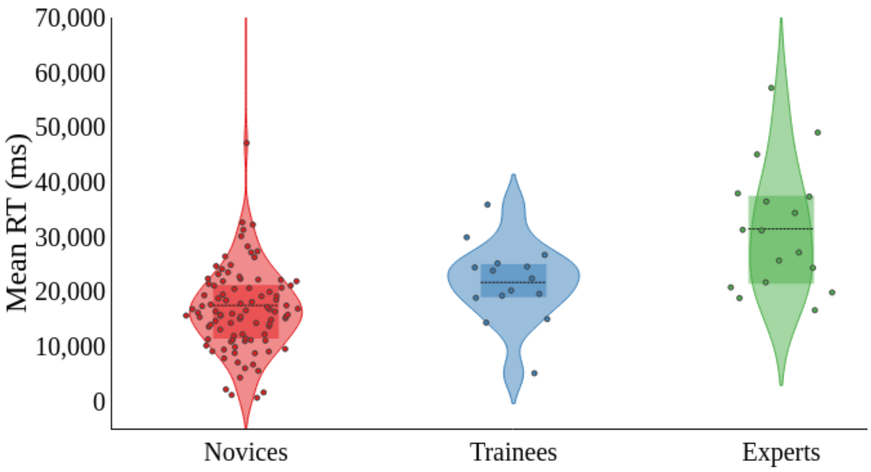
Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

An interaction between Group and Trial Type was also observed, *F*(2,124) = 16.81, *p* < .001, partial η^2 = 0.213, was explored with a series of independent *t*-tests with Bonferroni correction. This showed that on target-present trials, Novices (*M* = 48.0, *SE* = 2.10) were less accurate than Experts (*M* = 82.9, *SE* = 4.96), *t(*124) = 6.48, *p* < .001, and Trainees (*M* = 84.1, *SE* = 5.28), *t*(124) = 6.35, *p* < .001, with no difference in the accuracy of Experts and Trainees, *t*(124) = 0.16, *p* = 1.00. On target-absent trials there was no difference in the accuracy of Novices (*M* = 59.0, *SE* = 2.32) and Trainees (*M* = 66.7, *SE* = 5.85), *t*(124) = 1.24, *p* = 1.00, Novices and Experts (*M* = 64.5, *SE* = 5.49), *t*(124) = 0.93, *p* = 1.00, or Trainees and Experts, *t*(124) = 0.28, *p* = 1.00.

The mean response times of the groups were then calculated by trial type and used to compare performance with a 3 (Group: Novices, Trainees, Experts) x 2 (Trial Type: Target Present or Target Absent) mixed model ANOVA. This revealed a main effect of Group, *F*(2, 124) = 12.2, *p* < .001, partial η^2 = 0.165 (shown in Figure 4.13)

Figure 4.13

A Comparison of the Mean Response Times of Novices, Trainees and Experts in the MFFT



Note. Boxplot denotes the interquartile range of scores, and the black line the mean score.

This was further explored with a series of independent *t*-tests with Bonferroni correction which identified that Novices (*M* = 17.5s, *SE* = 1.10) were faster than Experts (*M* = 31.3s, *SE* = 2.62), *t*(124) = 4.89 *p* < .001, but not Trainees (*M* = 21.67s, *SE* = 2.79), *t*(124) = 1.41, *p* = .50, whereas Experts were slower to respond than Trainees, *t*(124) = 2.53, *p* = 0.40. There was no interaction between Group and Trial Type, *F*(2, 124) = 0.23, *p* = .80, partial η^2 = 0.004.

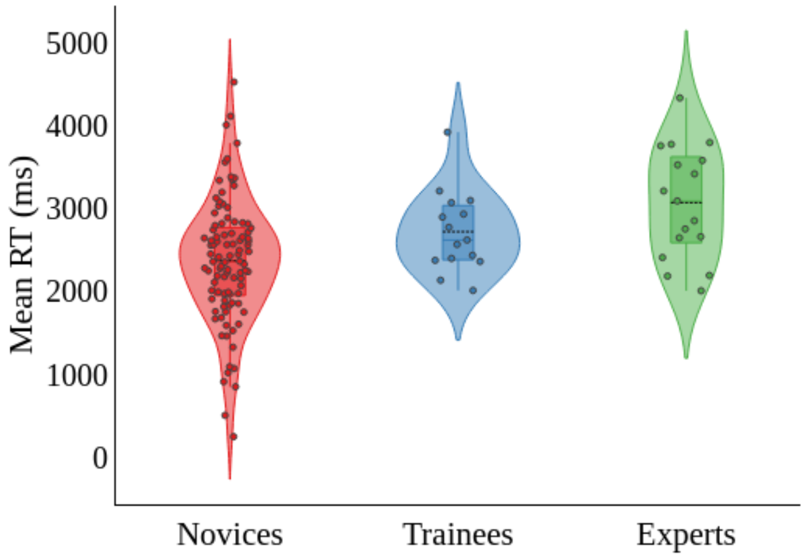
In summary, Experts and Trainees were both more accurate than Novices, with no difference between the fingerprint examiner groups. When the target was present, Novices were around half as accurate as Experts and Trainees and, again, there was no difference in the performance of the fingerprint examiners. When the target was absent, there was no difference in accuracy between the three groups. Overall, Experts were slower to respond than Novices and Trainees, and there was no difference in the response times of Novices and Trainees.

Visual Search Test: In this test, observers needed to identify whether a letter ‘T’ was present or absent in arrays of 1, 5, 15 and 30, constructed from the letter ‘L’. As a first stage, mean overall accuracy was computed for Novices, Trainees and Experts. This identified that many observers were at ceiling and group means were high: Novices (*M* = 94.5, *SE* = 0.57), Trainees (*M* = 97.9, *SE* = 1.44), Experts (*M* = 97.8, *SE* = 1.35). Therefore, in line with previous studies (Müller & Krummenacher, 2006; Wolfe, 2010), the analyses focused on the speed of responses as a measure of performance. Mean responses times were calculated for correct trials by trial type and array size, and analysed with a 3 (Group: Novices, Trainees, Experts) x 2 (Trial Type: Present or Absent) x 4 (Array: 1, 5, 15, 30) mixed-model ANOVA. This revealed a main effect of Group, *F*(2, 124) = 20.0, *p* < .001, partial η^2 = 0.244, and is shown in Figure 4.14. This was further analysed with a series of independent *t*-tests with

Bonferroni correction which found that Novices (*M* = 1.43s, *SE* = 0.05) were faster than Experts (*M* = 2.14s, *SE* = 0.10), *t*(124) = 6.32, *p* < .001, but not Trainees (*M* = 1.54s, *SE* = 0.11), *t*(124) = 0.75, *p* = .99. In contrast, experts were slower to respond than Trainees, *t*(124) = 4.11, *p* < .001.

Figure 4.14

A Comparison of Mean Response Times of Novices, Trainees and Experts in the Visual Search Test



Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

There was also an interaction between Group and Trial Type, *F*(2,124) = 4.90, *p* = .009, partial η^2 = 0.073. This was further explored with a series of independent *t*-tests with Bonferroni correction, which revealed Novices (*M* = 1.43s, *SE* = 0.44) responded faster on target-present trials than Experts (*M* = 2.14s, *SE* = 1.12), *t*(124) = 5.72, p < .001, but not Trainees (*M* = 1.41s, *SE* 1.20), *t*(124) = 0.19, *p* > .05, and Trainees responded faster than

Experts, *t*(124) = 4.40, *p* < .001. On target-absent trials, Novices (*M* = 1.42s, *SE* = 0.45) were also faster than Experts (*M* = 2.15s, *SE* = 1.06), *t*(124) = 6.39, *p* < .001, but not Trainees (*M* = 1.62s, *SE* = 1.13), *t*(124) = 1.68, *p* > .05, and Trainees were faster than Experts, *t*(124) = 3.43, *p* = .01. There was no interaction between Group, Trial Type and Array, *F*(6, 372) = 0.46, *p* = .84, partial η^2 = 0.007.

In summary, Novices and Trainees responded faster than Experts in correct target-present and target-absent trials. There was no difference in the response times of Novices and Trainees. Therefore, as a faster response time reflected higher performance in this test, the Novice and Trainee groups both outperformed the Expert group.

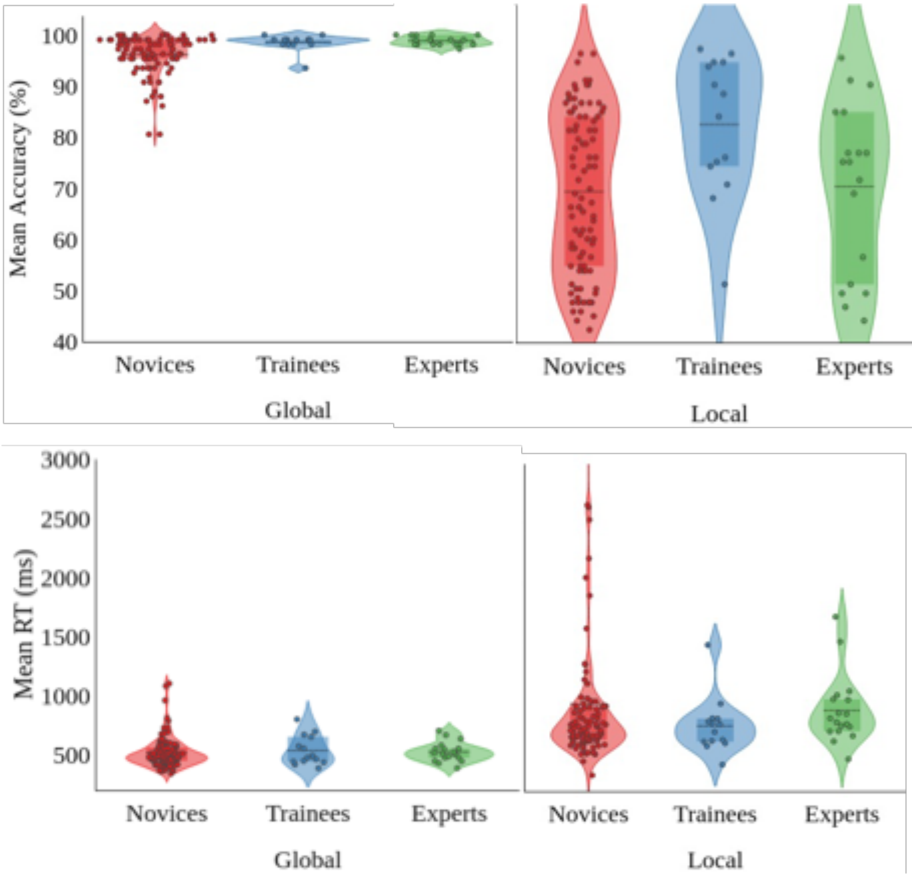
Navon Letters Test: In this test observers viewed Navon letters in local and global trials in two conditions: either the letter shape was constructed from the same letter (consistent) or from another letter (conflicting). To compare performance between the groups, mean percentage accuracy was calculated for each group by trial type and by condition. The data was used to compute a 3 (Group: Experts, Trainees, Novices) x 2 (Trial Type: Global or Local) x 2 (Condition: Consistent or conflicting) mixed-model ANOVA. This revealed a main effect of Group, *F*(2,124) = 7.71, *p* < .001, partial η^2 = 0.111, which was further analysed with a series of independent *t*-tests with Bonferroni correction. These showed that Trainees (*M* = 92.2, *SE* = 2.21) were more accurate than Experts (*M* = 84.3, *SE* = 2.08), *t*(124) = 2.62, *p* = .03, and Novices (*M* = 82.9, *SE* = 0.88), *t*(124) = 3.94, *p* < .001. There was no difference between Experts and Novices, *t*(124) = 0.62, *p* > .05.

There was also an interaction between Group and Trial Type, *F*(2, 124) = 5.37, *p* = .006, partial η^2 = 0.080 (Figure 4.15). This was further explored with a series of independent *t*-tests with Bonferroni correction. In global trials, there was no difference in accuracy between Experts (*M* = 84.3, *SE* = 2.08) and Novices (*M* = 96.2, *SE* = 0.88), *t*(124) = 2.59, *p* =

.16, or between Experts and Trainees (*M* = 98.6, *SE* = 0.90), *t*(124) = 0.00, *p* > .05 or Trainees and Novices, *t*(124) = 2.46, *p* = .23. In local trials, there was no difference in accuracy between Novices (*M* = 69.5, *SE* = 1.66) and Experts (*M* = 69.6, *SE* = 3.92), *t*(124) = 0.09, *p* > .05, or between Trainees (*M* = 85.8, *SE* = 4.18) and Experts, *t*(124) = 2.77, *p* > .05. However, in these trials Trainees were more accurate than Novices, *t*(124) = 3.63, *p* = .006.

Figure 4.15.

Comparison of Mean Accuracy (Top) and Mean Response Times in Global and Local Trials (Bottom) for Experts, Trainees, and Novices



Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

To further compare performance in this test, mean response times for all groups were calculated by trial type and condition (see Figure 4.17). The data was then used to compute a 3 (Group: Experts, Trainees, Novices) x 2 (Trial Type: Global or Local) x 2 (Condition: Consistent or conflicting) mixed-model ANOVA. This revealed no differences in response times between groups, $F(2, 123) = 0.43$, $p = .65$, partial $\eta^2 = 0.007$, and no interaction of Group and Trial Type, $F(2, 123) = 0.64$, $p = .53$, partial $\eta^2 = 0.01$, Group and Condition, $F(4, 246) = 0.30$, $p = .88$, partial $\eta^2 = 0.005$, or Group by Trial Type and Condition, $F(4, 246) = 0.33$, $p = .86$, partial $\eta^2 = 0.005$.

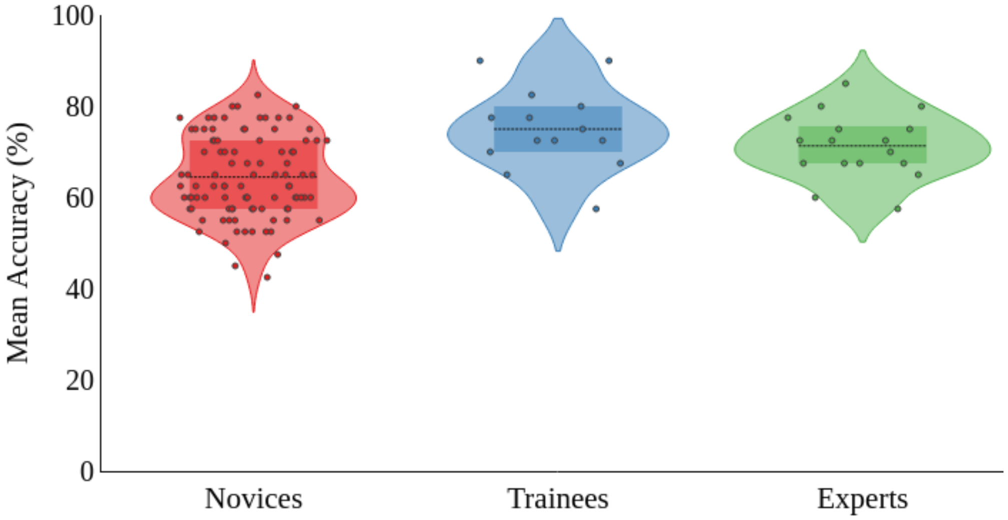
To conclude, when data was collapsed across the Navon letters test, Trainees were more accurate than Novices and Experts, and there was no difference in the accuracy of Experts and Trainees. During trials in which the stimuli were presented globally, no differences in accuracy emerged between the groups. However, when the stimuli were presented locally, Trainees were more accurate than Novices, but not Experts, and there was no difference in the accuracy of Experts and Novices. There were no differences in the response times of the three groups across the test.

Kent Face Matching Test (KFMT): In this test observers viewed pairs of faces which they needed to identify as the same or different. To compare accuracy between Novices, Trainees and Experts, mean percentage accuracy was first calculated for each group by trial type. This data was then analysed with a 3 (Group: Novices, Trainees, Experts) x 2 (Trial Type: Same or Different) mixed-model ANOVA. This revealed a main effect of Group, $F(2, 124) = 12.5$, $p < .001$, partial $\eta^2 = 0.168$, which is shown in Figure 4.12. This was further explored with a series of independent *t*-tests with Bonferroni correction which found that Novices ($M = 64.5$, $SE = 0.90$) were less accurate than Experts ($M = 71.3$, $SE = 2.12$), $t(124)$

= 2.94, *p* = .01, and Trainees (*M* = 75.3, *SE* = 2.26), *t*(124) = 4.43, *p* < .001. There was no difference in the accuracy of Experts and Trainees, *t*(124) = 1.29, *p* = .60. The interaction between Group and Trial Type did not reach significance, *F*(2, 124) = .096, p = .37, partial η^2 = 0.02.

Figure 4.16

A Comparison of Mean Percentage Accuracy of Novices, Trainees and Experts in the KFMT



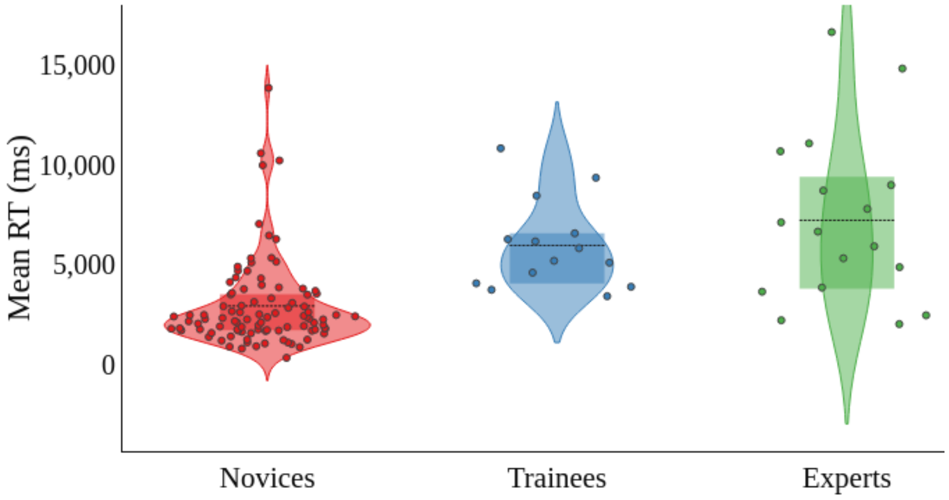
Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

The mean response times of each group were then calculated by trial type and used to compare performance with a 3 (Group: Novices, Trainees and Experts) x 2 (Trial Type: Same or Different) mixed-model ANOVA. This revealed a main effect of Group, *F*(2, 124) = 22.4, *p* < .001, partial η^2 = 0.266, and is shown in Figure 4.13. This was further explored with a series of independent *t*-tests with Bonferroni correction which found that Novices (*M* = 2.96s, *SE* = 0.33) were faster to respond than Experts (*M* = 7.23s, *SE* = 0.77), *t*(124) = 5.10, *p* <

.001, and Trainees (*M* = 7.46s, *SE* = 0.77), *t*(124) = 5.01, *p* < .001. There was no difference in the response time of Experts and Trainees, *t*(124) = 0.20, *p* > .05. Finally, there was no interaction between Group and Trial Type, F(2, 124) = 0.22, *p* = .81, partial η^2 = 0.003.

Figure 4.17

A Comparison of Mean Response Times of Novices, Trainees and Experts in the KFMT



Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

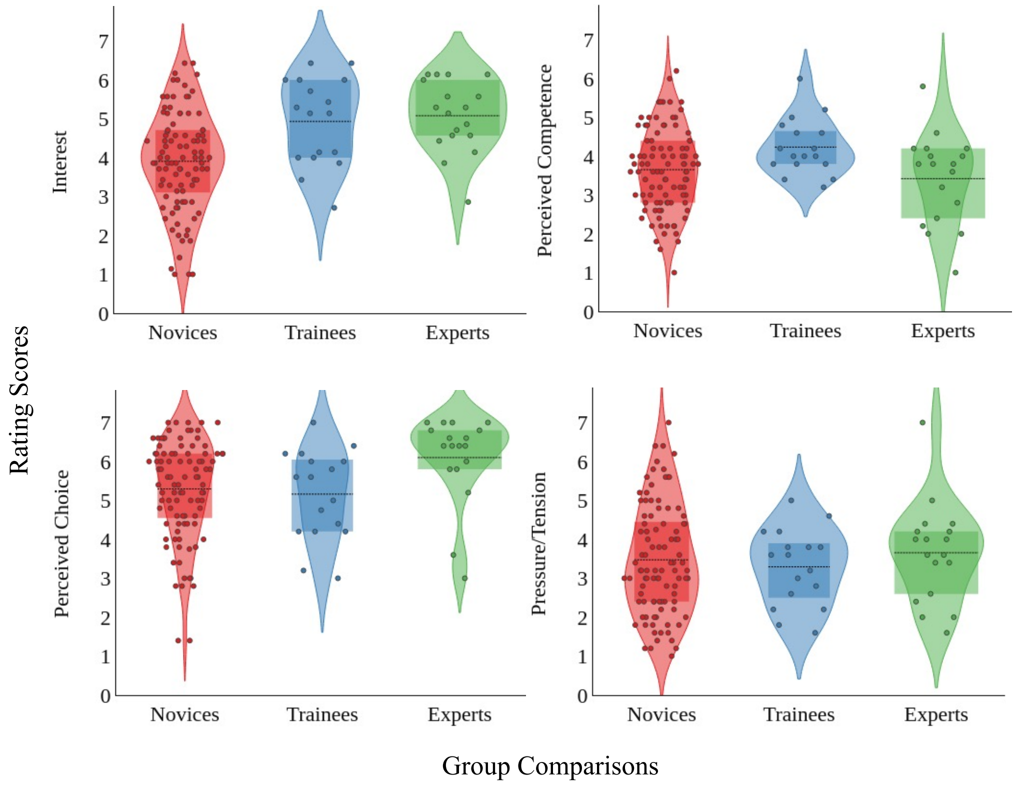
In summary, Experts and Trainees were both more accurate than Novices, and took at least twice as long to provide the correct response. There was no difference in the accuracy, or the response times, of the fingerprint examiner groups, who both performed above the normative mean of 66% accuracy for the KFMT.

<u>Intrinsic Motivation Inventory (IMI):</u> The next stage of the analyses was to compare the ratings of Experts, Trainees and Novices in the IMI. This survey was taken after

completing the test battery and related to test taking motivation. Observers' responses to the twenty-two items in the scale were categorised according to the subscales: interest, competence, choice, and pressure. Items which were reverse scored were re-coded, and the mean score within each subscale was calculated for each observer and then for each group (see Figure 4.18).

Figure 4.18

Comparison of the Mean Ratings of the IMI Subscales by Novices, Trainees and Experts



Note. Mean rating score is denoted by the black line.

A 3 (Group: Experts, Trainees, Novices) x 4 (IMI Subscale: Interest, Competence, Choice, Pressure) mixed-model ANOVA was then used to compare ratings. This revealed a main effect of Group, *F(*2, 124) = 4.30, *p* = .02, partial η^2 = 0.065, which was further explored with a series of independent *t*-tests with Bonferroni correction. These showed the overall mean ratings of Novices (*M* = 4.09, *SE* = 0.67) were lower than Experts (*M* = 4.54, *SE* = 0.16), *t*(124) = 2.67, *p* = .03, but not Trainees (*M* = 4.37, *SE* = 0.17), *t*(124) = 1.60, *p* = .34. There was no difference in the ratings of Experts and Trainees, *t*(124) = 0.74, *p* > .05.

There was an interaction between Group and IMI Scale, *F*(6,372) = 3.90, *p* < .001, partial η^2 = 0.060 (Figure 4.18). Novice interest in the tests appears lower than for the fingerprint examiner groups, and Trainees feel more competent than Novices and Experts. Expert scores appear to reflect a higher sense of test-taking choice than Trainees and Novices, and scores in relation to pressure/tension appear similar between the groups. This data was further explored with a series of independent *t*-tests with Bonferroni correction. Within the interest sub-scale, Experts (*M* = 5.07, *SE* = 0.29) showed more interest than Novices (*M* = 3.89, *SE* = 0.12), *t*(124) = 3.67, *p* = .03, but not Trainees (*M* = 4.92, *SE* = 0.31), *t*(124) = 0.33, *p* > .05. There was no difference in the ratings of interest between Novices and Trainees, *t*(124) = 3.05, *p* = .18. For comparisons in relation to the sub-scales of Competence, Choice, and Pressure, all *p*s > .05. Within-group comparisons across the sub-scales are not reported here and full analyses are included in Appendices.

Test Battery Summary: The test battery was designed to compare the performance of Experts, Trainees and Novices in a battery of cognitive tests related to fingerprint comparison. The results revealed differences in the performance of the three groups, which are summarised in Table 4.2.

Table 4.2

Summary of Analyses to Compare the Performance of Novices, Trainees and Experts in the Test Battery

	Reference and Comparison Groups		
Test	Novices : Experts	Novices : Trainees	Trainees: Experts
VSTM Overall Accuracy	=	=	=
VSTM RT	=	=	=
MRT Overall Accuracy	<	<	=
MRT Overall RT	<	=	=
KFMT Overall Accuracy	<	<	=
KFMT Overall RT	<	<	=
MFFT Target Present Accuracy	<	<	=
MFFT Target Absent Accuracy	=	=	=
MFFT Overall RT	<	=	<
Visual Search Target Present RT	>	=	>
Visual Search Target Absent RT	>	=	>
Navon Global Accuracy	=	=	=
Navon Local Accuracy	=	<	=
Navon Overall RT	=	=	=
IMI Interest	<	=	=

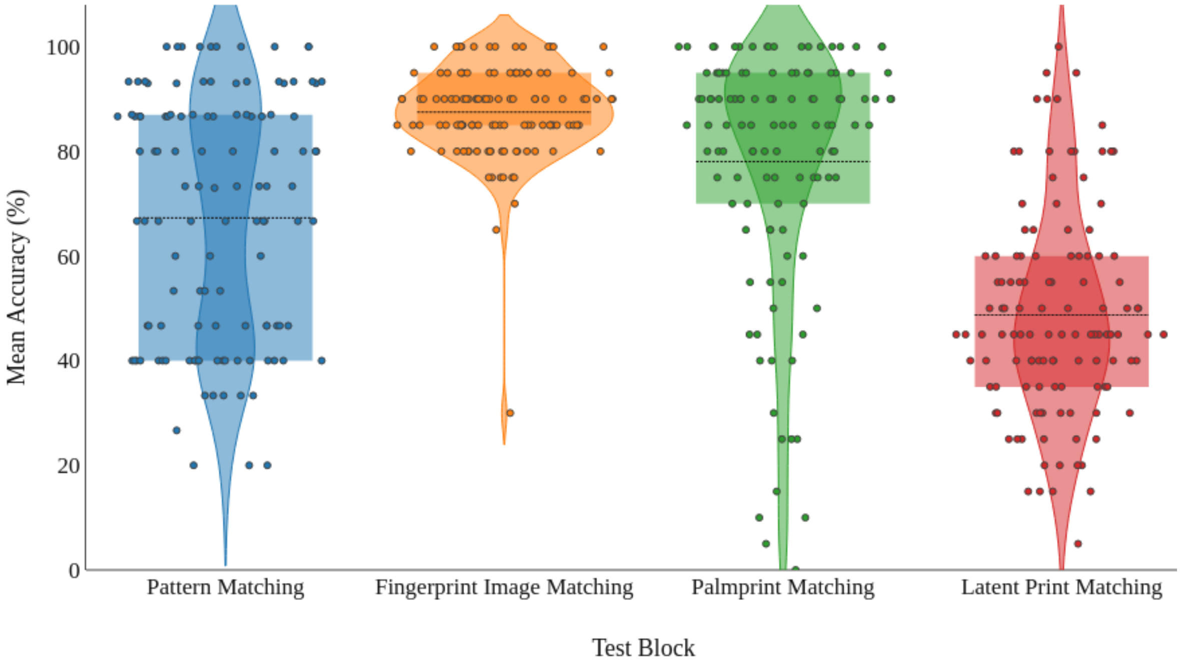
Note. Symbols (< and >) denote whether scores for the reference group are superior or inferior to the comparison group, or there is no difference in scores (=). Abbreviations: VSTM = Visual Short-Term Memory, MRT = Mental Rotation Test, KFMT = Kent Face Matching Test, MFFT = Matching Familiar Figures Test, IMI = Intrinsic Motivation Inventory.

The analyses showed that Experts and Trainees were more accurate than Novices in visual comparison tasks with unrestricted viewing time, namely figure matching (MFFT), face matching (KFMT), and mental rotation (MRT). Although Novices were faster to respond in these tasks, this was accompanied by lower accuracy than the Experts and Trainees. However, during the Visual Search test, Novices and Trainees responded more quickly than Experts. In this task, speed of response in correct trials was taken as a measure of performance, and Experts were therefore outperformed by less experienced observers In relation to global and local responses to stimuli, differences were only revealed during local trials in which Trainees were more accurate than Novices. At the conclusion of the test battery, observers completed a measure of their test taking motivation (IMI). Differences only emerged in relation to the interest subscale, with Experts expressing higher interest in the test battery tasks than Novices.

Multiple Regression Analyses: In the final stage of the analyses, the relationship between performance in the test battery and accuracy in fingerprint comparison was explored using data from observers who had taken part both in the fingerprint test reported in Chapter 3 (see Figure 4.19 for a summary of the accuracy data) *and* the test battery. This resulted in test scores from Experts (*n* = 6), Trainees (*n* = 15), and Novices (*n* = 95) which were combined (*n* = 116) for entry into regression analyses. The predictor variables used mean accuracy and mean response time data from the test battery, together with IMI ratings. The outcome variables reflected mean accuracy in each block of the fingerprint aptitude test: Pattern Matching (Block 1), Fingerprint Image Matching (Block 2), Palmprint Matching (Block 3), and Latent Print Matching (Block 4).

Figure 4.19

Comparison of Mean Percentage Accuracy for Each Block of the Fingerprint Aptitude Test



Note. Boxplot denotes the interquartile range of scores, and the black line represents the mean score.

As each block of the fingerprint aptitude test reflected a different element of fingerprint comparison ability, predictor variables were entered into a separate regression analysis for each block of the test. Jamovi software (The Jamovi Project, 2021) was used for the analyses, and all predictors were initially entered into the regression. A backward elimination process was used to identify predictors that were significantly associated with fingerprint accuracy, and the lowest AIC (Akaike Information Criterion; Akaike, 1974) value was used to determine which of the predictors to retain.

Multiple Regression Analyses Summary: Table 4.3 shows the predictors that were *retained* in the final model of the regression analysis for each block of the fingerprint aptitude test.

Table 4.3

Model Fit, R^2, T-Values and Beta for Variables Retained from Multiple Regression Analyses Predicting Accuracy in Each Block of the Fingerprint Aptitude Test

	Outcome Variables (DV)			
	Pattern Matching (Block 1)	**Fingerprint Image Matching (Block 2)**	**Palmprint Matching (Block 3)**	**Latent Print Matching (Block 4)**
Model Fit				
F (df) R^2	27.5 (3,112) ** .42	21.6 (3,112) ** .37	25.1 (4,111) ** .48	19.8 (5,110) ** .47
Predictor Variables (IV)	***t*, β**	***t*, β**	***t*, β**	***t*, β**
MRT Overall Accuracy	3.66, 0.27 **		4.88, 0.38 **	4.17, 0.31 **
MRT Overall RT		4.20, 0.32 **	2.83, 0.21 *	
KFMT Overall RT				3.55, 0.26 **
MFFT Target Present Accuracy	4.38, 0.34 **	3.40, 0.27 **		2.97, 0.24 *
Global Accuracy (Navon Letters)				2.19, 0.16 *
Local Accuracy (Navon Letters)	3.90, 0.30 **		3.45, 0.26 **	
Interest (IMI)		3.67, 0.29 **		2.46, 0.24 *
Choice (IMI)			2.38, 0.17 *	

Note. ** $p < .001$, * $p < .05$. Abbreviations: Mental Rotation Test (MRT), Kent Face Matching Test (KFMT), Matching Familiar Figures Test (MFFT), Intrinsic Motivation Inventory (IMI).

As the predictor variables used different scales of measurement, Beta is reported to reflect the standardized coefficients for comparison. These analyses identified clear differences in the perceptual and cognitive abilities associated with each block of the fingerprint aptitude test, as well as identifying those abilities that were shared across the test. Higher accuracy in mental rotation (MRT) and in matching target-present familiar figures (MFFT) predicted fingerprint accuracy in three of the four blocks of the fingerprint test. Slower response times in mental rotation (MRT) also predicted better performance in fingerprint image matching (Block 2) and palmprint matching (Block 3). Similarly, slower response times in the KFMT predicted higher accuracy in latent print matching (Block 4). In relation to holistic and featural processing of stimuli, higher accuracy in matching locally presented Navon letters predicted accuracy in pattern matching (Block 1) and palmprint matching (Block 3), while higher accuracy in globally presented Navon letters only predicted accuracy in latent print matching (Block 4).

Greater test taking motivation, as measured by the interest subscale of the IMI, significantly predicted accuracy in fingerprint image matching (Block 2) and latent print matching (Block 4), and a greater sense of freedom (choice) to participate predicted higher accuracy in palmprint matching tasks (Block 3). Of all the outcome variables, the largest number of predictors (five) were associated with accuracy in latent print matching (Block 4).

In summary, comparing the performance of experienced fingerprint examiners with that of trainee fingerprint examiners and novices in a series of cognitive and perceptual tests identified several key differences in ability. Experts and Trainees were both more accurate than Novices in the feature comparison tasks of face matching (KFMT), matching familiar figures (MFFT) and mental rotation (MRT), and this accuracy advantage was largely accompanied by slower response times in these tests. In fact, during the Visual Search Test, where faster response times in correct trials equated to better performance, the Novice and

Trainee observers outperformed the Expert group. Conversely, other tests in the battery revealed no, or few, differences between the groups. In the visual short-term memory (VSTM) test there was no difference in the performance of Experts and Trainees and Novices, and differences in the Navon Letters test were confined to those between Trainees and Novices in local trials. In terms of test taking motivation, differences only emerged between Experts and Novices, with Experts reporting higher levels of interest in the tests.

Using multiple regression to explore the relationship between cognitive skills and fingerprint comparison provided converging evidence to support the role of feature comparison ability in fingerprint identification accuracy. In three blocks of the fingerprint test, better performance in these comparison tasks (MRT, MFFT) predicted higher accuracy. The relationship between longer response times and higher accuracy that was observed in the between-group test battery data was also supported by findings from the regression analyses. Here, *slower* responses in the MRT and the KFMT predicted higher fingerprint comparison accuracy. The role of accuracy when viewing locally or globally presented stimuli was also predictive of fingerprint test accuracy, although few differences in this ability had been revealed in the between-group analyses. Similarly, the use of the IMI to measure test taking motivation had only identified differences in interest between Experts and Novices, however, higher motivation predicted higher accuracy in three of the four blocks of the fingerprint test.

General Discussion

This chapter compared the performance of experienced forensic fingerprint examiners (‘Experts’), trainee forensic finger examiners (‘Trainees’), and a control group (‘Novices’) with no previous experience in fingerprint comparison, in a battery of cognitive and perceptual tests. These were designed to reflect some of the cognitive processes likely to be required for accurate fingerprint comparison decisions. The aim of the experiment was to identify the cognitive abilities that may underly fingerprint comparison proficiency. As such, some variation between the performance of forensic examiners and untrained observers was anticipated, with experience in forensic comparison perhaps predictive of an accuracy advantage in cognitive tasks.

Analysis of the accuracy data from the test battery revealed clear differences in the performance of the fingerprint examiners and Novices, with Experts and Trainees more accurate than Novices in the three feature comparison tasks. In the Kent Face Matching Test (Fysh & Bindemann, 2018), Expert and Trainee accuracy was higher than the normative score of 66% for this test. This signals that they not only outperformed the novice group but also demonstrated a high level of accuracy in the test. Experts and Trainees were also more accurate than Novices in the Matching Familiar Figures Test, with scores that were higher than the ~66% reported in previous studies (e.g., Burton et al, 2010; Megreya & Burton, 2006). This comparison again reflects a generally high level of accuracy in this task, rather than performance that is merely better than the control group. The Mental Rotation Test is designed to measure spatial ability, with observers required to mentally rotate two 3D stimuli to decide whether they are the same pair or different (Ganis & Kievett, 2015; Shepard & Metzler, 1988). Although not solely a test of feature comparison, to compare the shapes and patterns of the blocks comprising the stimuli this skill is undoubtedly required in addition to the ability to mentally rotate stimuli. Normative data for this test usually compares accuracy and response times by degrees of rotation. However, in this experiment there was no interaction between rotation and accuracy, or by trial type. Nonetheless, the Experts and Trainees again convincingly outperformed the Novices, with an accuracy advantage in excess of ten percent.

Given the superior performance of the Experts and Trainees in these tests, the role of a *general* feature comparison ability in fingerprint comparison should be given consideration.

Fingerprint comparison is an analytical process, usually following ACE (FSR, 2017; SWGFAST, 2013) methodology. The ‘comparison’ element of ACE requires examiners to locate features across pairs of fingerprints to identify those in common or those that are discrepant. A similar comparative process could therefore account for high accuracy with non-fingerprint stimuli in tests such as the MRT, KFMT, and MFFT. Stimuli within these tests could only be determined as a match to the target, or different to the target, through identification, localization, and comparison of the constituent features.

The findings from this experiment contrast with those observed in previous research. In a study to compare the performance of experienced fingerprint examiners and novices in a face-matching task, no differences in accuracy were reported (Searston & Tangen, 2017b). This led to the conclusion that fingerprint comparison accuracy may be a domain-specific, rather than a domain-general, ability. However, the results from the current study tend to suggest that feature comparison ability in forensic examiners may exist across a range of perceptual matching tasks rather than being confined to those depicting fingerprint stimuli. Of relevance here is the finding that Experts and Trainees *both* outperformed Novices in these feature matching tasks, with no difference emerging in the accuracy of the fingerprint examiner groups. Some Trainees were only a few weeks into their training at the time of the tests and unlikely to have developed expertise across a range of fingerprint stimuli. They were, however, recruited to the role based on their high performance during an in-house fingerprint test. Therefore, it seems plausible that some observers may have a greater *aptitude* for feature comparison tasks more broadly, rather than a specific aptitude for fingerprint comparison.

Importantly, the observed relationship between performance in feature comparison tests and accuracy in the fingerprint aptitude test also tends to suggest a shared ability in these tasks. The regression analyses show that higher accuracy in the MFFT and MRT predicted better performance in most blocks of the fingerprint test. Observers were predominantly untrained novices, and while they may not have outperformed the Experts and Trainees observers as a group, the regression analyses revealed a positive relationship between higher accuracy in feature matching tasks and higher accuracy in fingerprint comparison. This again may reflect a feature comparison ability that generalises to more than one class of stimuli. Performance in other tasks, such as visual search and a test of visual short-term memory, did not reveal a similarly positive relationship to fingerprint comparison and differences in accuracy between the Experts, Trainees, and Novices were not observed. Converging evidence in support of a domain-general feature comparison ability was provided in a study by Growns et al. (2022). In tests to measure matching ability with fingerprints, faces, and firearms, a positive relationship in the performance of novice observers was observed across experiments. Conversely, this relationship was not found in relation to other perceptual tasks such as visual search and statistical learning.

In addition to considering accuracy in cognitive tasks, the response times of observers were also compared. During the test battery, Novices were quicker to respond during the KFMT, MRT and MFFT, and in the visual search task their faster response times reflected better performance than the Experts and Trainees. Previous research (e.g., Stevenage & Pitfield, 2016), as well as data from the experiment described in Chapter 3, has shown that fingerprint examiners take longer to reach fingerprint identification decisions than novice observers. Whilst a measured and analytical approach to comparisons may account for these longer response times, they may also reflect a tendency to respond more conservatively when faced with increasingly ambiguous stimuli (Busey et al, 2011). In workplace comparisons, this may reflect a desire to avoid false positive identifications. Whether this response pattern persists in lower stakes comparisons, such as the test battery, is difficult to determine.

A further consideration with regards to longer response times in Experts and Trainees is the influence of the ACE-V method of forensic fingerprint comparison. This is an iterative process of checking the reference, or target, print and exemplars. The response times of Experts and Trainees during the test battery may therefore also reflect the application of this comparison method to non-fingerprint stimuli. The longer response times for Experts and Trainees certainly suggest a more analytical and deliberate approach to *all* feature comparison decisions. It is also of note that Expert and Trainee accuracy was generally the same as Novices when response times were restricted during the VSTM and Navon Letters tasks, thereby limiting opportunities for analytical comparison.

Motivation to perform well may also have encouraged slower response times in Experts and Trainees. Although they undertook these tests whilst they were working from home and were aware that individual results would not be shared, their participation arose from their workplace experience. Therefore, longer response times may reflect a more cautious approach to the tests to ensure that accuracy reflected their professional standing. Differences in motivation, as captured by the IMI, only revealed differences in the subscale of 'interest' between Experts and Novices. It is therefore difficult to determine the influence of motivation on test performance, with a relationship between the subscales of interest, choice, and fingerprint accuracy only observed in two blocks of the fingerprint test (fingerprint image matching and palmprint matching). The role of motivation on performance in fingerprint comparison tasks is therefore inconsistent, which tends to support evidence from a recent study in which individual differences in feature comparison tasks were not accounted for by task motivation (Growns et al., 2022)

In addition, the test battery data revealed differences in global and local processing, with Trainees more accurate than Novices in local tasks. Local accuracy also predicted performance in pattern matching and palmprint matching in the fingerprint aptitude test. In tests using Navon letters, the global shape of the letter typically interferes with local processing (Navon, 1977). Global processing is believed to be more automatic, whereas local processing requires deliberate attention to stimuli. Therefore, observers who can focus their attention on the features within the stimuli, rather than the global image, are likely to demonstrate higher accuracy in pattern matching and palmprint matching.

Conversely, higher global accuracy was associated with performance gains in latent print matching. The difference here could be explained by the importance of *contextual information* during latent print matching in guiding the observer to the relevant target area. Research has previously found that accuracy in localizing the target area from one print onto a corresponding exemplar is impaired when only a cropped area of the target print is visible, but faster and more accurate when the target area and surrounding print are visible (Hicklin et al., 2019). The results from the current study tend to suggest the importance of both local and global accuracy during fingerprint examination, and an ability to adapt these visual processes to suit the nature of the comparison.

Finally, while the test battery was designed to reflect some of the perceptual processes related to forensic fingerprint comparison, it did not reveal between-group differences in a test of visual short-term memory and this ability did not predict fingerprint test accuracy. An important caveat in relation to this outcome may relate to the type of short-term memory test deployed in this battery. The test was essentially in two parts: observers needed to identify an item that was replaced in an array and the item it was replaced with. On reflection, the complexity of this task may not have accurately captured the memorial processes required for fingerprint comparison. Researchers have indicated that short-term memory must be a component of fingerprint examination by virtue of the requirement to remember relevant features in a target print with which to compare with an exemplar (Busey et al., 2011: Hicklin et al., 2019). Differences in this ability may therefore have been captured if a less complex test had been used here, perhaps with a single stimulus change and a shorter delay between presentation and test. The absence of a significant finding in this experiment should therefore not be regarded as conclusive evidence that visual short-term memory does not have a role to play in fingerprint comparison ability. This important cognitive process should be further explored in future research.

In conclusion, the test battery identified differences in cognitive ability between experienced and trainee forensic fingerprint examiners and untrained observers. In feature comparison tasks, such as those requiring observers to compare faces, familiar figures, and 3D objects, Experts and Trainees outperformed the Novices, and there were no differences in accuracy between experienced fingerprint examiners and those undergoing training in fingerprint comparison. These findings suggest that during feature-comparison tests, higher accuracy may reflect a domain-general ability for these tasks, rather than performance that is based on expertise and experience. Slower response times for the Experts and Trainees also suggested a more methodical approach to these tasks. These may reflect workplace forensic comparative processes as well as a desire for high accuracy to be commensurate with their professional standing.

Converging evidence from multiple regression supported the importance of feature comparison skill for fingerprint comparison accuracy. The ability to compare features across stimuli was positively associated with performance across different fingerprint tasks. Slower response times in these cognitive tests also suggested a careful and more analytical approach is likely to benefit accuracy. In addition, the ability to adapt to stimulus demands was demonstrated by the relationship between global and local processing in different blocks of the fingerprint test.

Forensic fingerprint examiners undergo a rigorous training, mentoring, and testing regime within the workplace. The results from this experiment are not intended to imply that this should, in future, incorporate feature-comparison training with non-fingerprint stimuli. The experiment does, however, raise some important considerations for the recruitment of suitable staff to the role. In previous research, fingerprint examiners were more accurate than novice observers in a test of face matching ability, which suggests an advantage in feature-comparison tasks outside of the area of experience (Phillips et al., 2018). The current experiment has demonstrated a clear relationship between feature-comparison performance and fingerprint aptitude in experiences and trainee fingerprint examiners. This relationship was also apparent in a group of predominantly untrained observers, with high performance in feature-matching tests predictive of accuracy gains during fingerprint matching tasks. The inclusion of a battery of tests, to measure feature-comparison ability and perceptual adaptability in the global and local processing of stimuli, could therefore assist in the selection of those applicants with the highest aptitude for forensic fingerprint comparison.

Chapter 5

Summary, Discussion, and Future Directions

5.1 Summary of Main Findings

This thesis investigated expertise in forensic fingerprint comparison with the creation of a fingerprint aptitude test that incorporated stimuli of varying difficulty. This test was subsequently used to explore the expertise of forensic fingerprint examiners, and the relationship between fingerprint comparison and cognitive ability in a battery of perceptual tests. The first chapter provided a comprehensive review of research into forensic fingerprint examination and the nature of forensic expertise to date. The importance of expertise in forensic comparison has been highlighted in recent years following comprehensive reviews undertaken by the National Academy of Sciences (NAS, 2009) and the President's Council of Advisors on Science and Technology (PCAST, 2016). These found considerable variation in performance and accreditation measures between laboratories, and a lack of foundational validity in forensic comparison in general. Whether forensic comparison decisions were accurate, repeatable, and reliable was a key consideration.

In relation to fingerprint comparison, the false positive rate at which fingerprints were wrongly attributed to the same source was higher than expected, with the true accuracy of live casework difficult to determine (PCAST, 2016). Inter and intra-examiner variability in fingerprint comparison decisions has been reported (Ulery et al., 2011; Ulery et al., 2012), along with variation in the number of minutiae identified (Dror et al., 2011; Ulery et al., 2016)), and the number of minutiae required for identification (Ulery et al., 2014). The risk of contextual bias (Dror et al., 2005; Dror & Charlton, 2005), and the effect of confirmation bias (Fraser Mackenzie et al., 2013) on examination decisions are further factors that may impact upon the reliability of forensic fingerprint comparison outcomes.

The absence of a sound scientific basis for forensic fingerprint comparison is particularly pertinent in light of the ‘expert’ status afforded to testimony from fingerprint examiners. Given the highly probative influence of forensic examiner testimony on jury decision making (Ribeiro et al., 2019; Schweitzer & Saks, 2007), and the risk of harm to victims (Poyser & Milne, 2010) and the wrongly accused (Gould & Leo, 2010) from flawed evidence, the foundational validity of forensic fingerprint expertise warrants greater understanding. Of particular relevance is the fact that an identification decision is based solely on an examiner’s subjective determination rather than a quantitative measure of comparison (Hicklin et al., 2019). It therefore follows that in order to provide information which is likely to be outside of the knowledge of a court (CPS, 2020), fingerprint experts should possess experience and skills that are unlikely to be found in the general population (Roberts, 2021).

To identify whether there are clear differences in the fingerprint identification abilities of forensic fingerprint examiners and untrained observers, fingerprint comparison tasks have been created using simulated crime scene prints (Tangen et al., 2011), genuine crime scene marks (Thompson et al., 2014), good and poorer quality fingerprints from a database (Stevenage et al., 2017), and cropped rolled fingerprint impressions (Searston & Tangen, 2017a; 2017b). These studies have shown that examiners demonstrate superior accuracy to novices, and a domain-specific ability in fingerprint comparison is assumed which is deemed unlikely to generalize to other classes of stimuli (Searston & Tangen, 2017b). However, although different types of fingerprint images have been incorporated into previous research, there was currently no test available that contains varied stimuli within a single test. Such a test would assist in determining the level at which experienced examiners excel in fingerprint comparison and provide a clearer comparison of this ability with untrained observers.

The purpose of Chapter 2 was to address current limitations in the availability of a suitable fingerprint aptitude test through the creation of a novel test. This was designed in collaboration with senior forensic fingerprint examiners from a large UK police service and incorporated bespoke fingerprint stimuli created for this research. Fingerprint images were graded to ensure suitability for viewing by a range of observers. Using prints from volunteer donors ensured the ground truth as to the source of the fingerprints was known. The decision to create this test for online participation was largely driven by restrictions imposed during the COVID pandemic.

The fingerprint test comprised of four blocks of different trials, designed to capture observers' abilities in a range of tasks related to fingerprint comparison. The first block of the test measured visual pattern matching ability using a series of black on white line drawings. These were created to be analogous to comparing features across fingerprints but did not resemble fingerprint stimuli in appearance. The second block of the test measured fingerprint image matching and was the first introduction to fingerprint stimuli. Clear images were selected and observers needed to identify the identical image of a target in an array or determine that the target image was not present. The third block of the test comprised of palmprint stimuli, overlaid with grids of smaller squares. Observers viewed a cropped section of the palmprint which they needed to locate on, or exclude from, the image of the entire palm. In the final block of the test, observers were required to identify whether a latent print matched an array of inked fingerprints. The latent prints were representative of those recovered from a crime scene and contained interference and distortions, thereby making this the most challenging block of the test. In a reflection of live casework fingerprint comparisons, some of the target images were also rotated or inverted in each block.

Chapter 2 reported the performance of novice observers who undertook the online fingerprint aptitude test on two occasions, separated by a mean interval of eleven days. The aim was to determine the reliability and internal consistency of the test. Differences in accuracy and response times between the blocks of the test showed that variations in task difficulty had been effectively captured. Observers were also required to rate the confidence of their response to each test item, and this revealed a strong association between accuracy and confidence in each block. Importantly, there was no difference in this performance data between the first and second test sessions. In addition, there was a strong positive correlation between subject accuracy across sessions and at the item level between corresponding blocks in each test session. These findings show that the online fingerprint aptitude test is a reliable test of fingerprint comparison ability and can be used to measure the performance of observers with a range of fingerprint experience.

In Chapter 3, the fingerprint aptitude test was used as a means of measuring expertise in experienced forensic fingerprint examiners ('Experts'), trainee fingerprint examiners ('Trainees'), and untrained observers ('Novices'). The aim was to identify those fingerprint comparison tasks that could differentiate the performance of fingerprint examiners from untrained observers. In contrast to previous research, this was explored by comparing group mean performance and individual differences in fingerprint comparison ability. To achieve this, individual Novice accuracy was directly compared with that of the Expert and Trainee groups. This measured the level of Novice performance in each fingerprint comparison task, which assisted in identifying whether certain tasks differentiated the abilities of Experts, Trainees, and Novices.

During pattern matching trials in Block 1, Novices were outperformed by both Trainees and Experts at a group level. There was, however, considerable overlap in the accuracy of all groups. This pattern of performance continued during fingerprint image matching in Block 2, and palmprint matching in Block 3. Although these two tasks revealed some group-level differences, they did not differentiate the performance of Experts and Trainees. In addition, there was a substantial overlap in the accuracy of individuals in each observer group in both of these tasks. Therefore, neither pattern matching, fingerprint image matching, or palmprint matching were able to provide a clear index of fingerprint examiner expertise in this chapter. Thus, these types of fingerprint identification tasks show that comparing group-level performance does not accurately measure fingerprint expertise. The inclusion of individual-level metrics are therefore essential to contextualise this type of expertise.

In contrast to the preceding blocks of the test, accuracy in the latent fingerprint matching task of Block 4 clearly differentiated the performance of Experts, Trainees, and Novices. This was evidenced at group-level, with an accuracy advantage for Experts over the other groups. Crucially, this was also corroborated by analysis at the individual level, which revealed that only a very small proportion of Novices – of less than three percent – could perform within the Expert range. There was more overlap in the individual performance of Novices with Trainees, yet many Trainees also fell within the Expert range, which is likely to reflect the on-going accumulation of fingerprint knowledge during their training. These results therefore demonstrate that the true nature of fingerprint expertise relies on the ability to deal with the most challenging fingerprints, namely latent prints of the type that have been recovered from crime scenes. Moreover, the extent of this expertise distinguishes Experts from almost any Novice as well as many fingerprint examiners in training.

The results from the fingerprint test also hinted at other factors that may underlie the superior performance of Experts during latent print matching. For example, their response times were considerably slower than Novices, thereby suggesting a longer decision-making process and a more analytical approach to fingerprint comparison. This may reflect workplace procedures such as ACE, which is typically a slow, methodical, and iterative process in which pairs of fingerprints are analysed, compared, and evaluated (FSR, 2017; SWGFAST, 2013). However, as Trainees were also slower than Novices, this suggests that additional factors may underlie Expert performance. Research has previously shown that experience with fingerprints allows an examiner to identify and localise relevant features in an exemplar (Hicklin et al., 2019). Therefore, familiarity with a wide range of fingerprints during workplace comparisons may underpin the development of their proficiency (Thompson et al., 2014), leading to an accuracy advantage with latent print trials. In addition, fingerprint examiners can also identify more areas of diagnostic value within a fingerprint than less experienced observers (Busey et al., 2011). Challenging comparisons with unclear or distorted prints may therefore be reliant on the *experience* of fingerprint examiners in contrast to those observers with none or only limited experience with fingerprints.

Other between-group differences also emerged in the current study. In relation to errors committed during target-present comparisons, Trainees and Experts were more likely than Novices to fail to identify a target during latent print matching than to falsely identify a foil print as the target. This converges with previous work (Busey et al., 2011; Tangen et al., 2011; Thompson et al., 2014; Ulery et al., 2011) and may reflect a more cautious approach to these ambiguous fingerprint comparisons, based on workplace practices and training that emphasises assessment of the suitability and quality of a latent print for comparison (SWGFAST, 2013; Vanderkolk, 2011), to avoid false positives in identification.

Overall, this chapter demonstrated that, at an individual level, many untrained and inexperienced observers show considerable competence in fingerprint comparison tasks, such as pattern matching, fingerprint image and palmprint matching. Nonetheless, Experts and Trainees still demonstrated a group-level advantage in these tasks, which may suggest that some of these less challenging comparisons may be underpinned by a domain-general feature comparison ability (Growns et al., 2022). In latent print comparison, however, the Experts demonstrated a clear and consistent accuracy advantage over virtually all Novices. This suggests the development of a specific, experience-based, expertise for these difficult comparisons (Searston et al., 2017b).

Although Experts demonstrated clear expertise during the latent print comparison tasks, the fingerprint test was not designed to provide insight into the nature of any cognitive or perceptual abilities that might underpin their performance. The aim of Chapter 4 was to explore this relationship by comparing the abilities of Experts, Trainees, and Novices in a battery of perceptual tests. The use of test batteries is well established in the field of cognitive psychology (e.g., Bate et al., 2018; Burton et al., 2010; Davis et al., 2016; Wilhem et al., 2014), but has not so far been applied to the understanding of fingerprint expertise. The first stage of this study was to identify the cognitive process that are likely to be engaged during forensic fingerprint examination, and to create suitable tests with which to capture these abilities in a range of observers.

Forensic fingerprint examiners typically view fingerprints in a side-by-side format and, compared to novices, can readily focus on areas of the print with higher diagnostic value (Busey et al., 2011) to quickly locate corresponding features and details in a pair of prints (Hicklin et al., 2019). Therefore, the ability to retain information with which to make a comparison between fingerprints suggests a role for short-term memory in this process. In the creation of the test battery, this component was reflected in a test of Visual Short-Term Memory (VSTM) featuring arrays of three to sixteen familiar objects (from Beck & Levin, 2003; Borchert, 2020). A pre-change array was briefly presented (2000ms), followed by a post-change array (200ms) in which one object had been replaced. Observers needed to correctly identify both the pre-change and post-change objects, and accuracy and response times in this task were taken as a measure of performance. The ability to rapidly identify salient information within an image was also tested in a Visual Search task (based on Duncan & Humphreys, 1989). Here, observers needed to identify a target letter 'T' within arrays of one, five, fifteen, or thirty letters. Speed of correct responses was used to measure ability, with faster responses reflecting better performance.

Fingerprint examination also requires an ability to mentally rotate fingerprints across different viewing places, particularly if the print is distorted or impaired by substrate interference. In forensic comparisons, fingerprints are correctly oriented prior to examination which, of course, requires an understanding of the features comprising the upper and lower components of the print. To capture this ability, a test of mental rotation (MRT) was added to the test battery (Borchert, 2020) in which observers were presented with pairs of 3D cube-based objects with one image rotated from the vertical place by zero, fifty, one hundred or one hundred and fifty degrees (from Ganis & Kievet, 2015). Again, accuracy and response times in this task were used to measure performance.

The role of global and local processing in fingerprint comparison has not yet been widely researched. Global, or holistic, processing occurs when fragments of an image are combined into a whole percept (Wong & Gauthier, 2010; Richler et al., 2012) and is associated with experience within a discrete class of stimuli (Farah et al., 1998; Maurer et al., 2002). Anecdotally, examiners report that they process fingerprints holistically, which is partly supported by research (Vogelsang et al., 2017). However, fingerprints are comprised of a wealth of minutiae, such as lines, creases, pores, ridges, and valleys, which are often only visible with magnification. This is in addition to the broad categories of fingerprint patterns, such as loops, whorls, and arches, which can be viewed with the naked eye. It is the examination of spatial relations between fingerprint features that allows examiners to decide on the source of a fingerprint (Home Office, 2017). This therefore also hints at a feature-to-feature comparative process, with forensic examiners likely to attend to these ‘local’ details within a fingerprint rather than the whole image.

For this reason, a measure of local and global processing was added to the test battery in the form of a Navon letter test (Borchert, 2020; Navon, 1977). The stimuli comprised of briefly presented (100ms) large letter shapes (H or S) constructed from smaller letters, in which large and small letters were the same (congruent) or different (incongruent). Observers were directed to attend to the large letter to capture global processing or the small letter to measure local processing. Local processing is typically disrupted by the global letter shape and response times are typically longer (Navon, 1977). Therefore, accuracy and response times by trial type (global or local) were used as measures of performance in this test.

The final two perceptual tests in this battery measured performance in feature comparison tasks. Fingerprint examination primarily requires the observer to locate detail within one image with which to compare with another image. To explore whether skill in fingerprint comparison extends to comparing features across non-fingerprint stimuli, an unfamiliar face matching test, (KFMT; Fysh & Bindemann, 2018) and a Matching Familiar Figures Test (MFFT; Kagan, 1965) were added to the battery of tests.

The test battery was undertaken by the same Novices and Trainees who had completed the fingerprint aptitude test described in Chapter 3, along with a group of Experts, several of whom had also completed the fingerprint aptitude test. To explore any differences in the performance of Experts, Trainees, and Novices, the initial data analyses focused on group differences in relation to accuracy and response times for each test in the battery.

A clear difference in accuracy emerged in relation to the feature comparison tasks of mental rotation (MRT), unfamiliar face matching (KFMT), and matching familiar figures (MFFT). Here, both Experts and Trainees were more accurate than Novices, and there was no difference in accuracy between these police groups. Perhaps surprisingly, Experts and

Trainees did not outperform Novices in the test of short-term memory (VSTM) and during the Visual Search Test, Experts were outperformed by Trainees and Novices. In the latter, response times in correct trials were used to measure of ability, and slower response times for Experts therefore equated to poorer performance. Data from the Navon letters test also did not identify any differences in global or local processing between Experts and the other groups, and the only difference emerged in relation to local accuracy with Trainees being more accurate than Novices. Across the test battery, Novices were faster than Experts in the MRT, KFMT, and the MFFT, and faster than Trainees in the KFMT. The only difference to emerge between Experts and Trainees reflected faster responses by Trainees in the MFFT.

At the conclusion of the test battery, all observers undertook the Intrinsic Motivation Inventory (IMI; Ryan & Deci, 2000). This is a 22-item inventory that was used to measure subjective motivation in relation to observers' interest, perceived competence, perceived choice, and pressure in undertaking the test battery. Although Experts expressed higher interest than Novices, there were no other differences in motivation between the three groups.

The findings from the test battery data reveal several key points for consideration. The first of these relates to the superior accuracy of Experts and Trainees during the three feature comparison tasks. During the KFMT and the MFFT these police examiners not only outperformed the Novice group but demonstrated higher accuracy than the normative data published for these tests, thereby reflecting a high level of performance in these tasks. Although normative data is not available for the MRT, the Experts and Trainees again convincingly outperformed the Novices.

This raises the question of why police fingerprint examiners and fingerprint trainees performed better in tasks comprised of stimuli outside of their domain of experience or expertise. It is important to consider the role of forensic fingerprint comparison methods deployed in the workplace, notably ACE (FSR, 2017; SWGFAST, 2013), and whether the comparative nature of this methodology translates to performance with non-fingerprint stimuli. However, several of the Trainees were newly appointed to the role, with few opportunities to gain experience in fingerprint comparison. This suggests accuracy in non-fingerprint feature comparison tasks cannot be solely attributed to training input or workplace experience. Of note is the slower response times of Experts and Trainees, thereby allowing time for a more analytical comparison of stimuli which may have attributed to improved accuracy. Nonetheless, the police examiners still outperformed Novices when stimulus presentation was restricted to 7500ms in the MRT, thereby indicating that accuracy was not necessarily dependent on a longer viewing time.

The test battery data was further explored with a series of multiple regression analyses to examine whether performance in visual perceptual tasks predicted accuracy in fingerprint comparison. These incorporated data from observers who had taken part in both the fingerprint aptitude test in Chapter 3 and the test battery. The results showed that accuracy in the MRT and the MFFT, and slower response times in the MRT and the KFMT, predicted performance in fingerprint comparison. Of note is that these tasks required the observer to compare features across stimuli in order to locate the correct response. Conversely, performance in test battery tasks that did not require feature comparison ability, namely visual short-term memory (VSTM) and visual search, failed to predict accuracy in fingerprint comparison. In addition, the analyses showed that a relationship between global or local processing was task specific: local accuracy predicted performance in pattern and palmprint matching (Blocks 1 and 3), whereas global accuracy predicted performance in latent print matching (Block 4).

The results from the test battery certainly suggest a broad feature comparison ability may underpin accuracy in fingerprint comparison. Experts and Trainees outperformed Novices in the KFMT, the MFFT, and the MRT, despite the absence of fingerprint stimuli in these tests. Evidence of this relationship between feature comparison ability and fingerprint comparison accuracy was also provided by regression analyses, with performance in the MFFT and the MRT predictive of accuracy in three out of four blocks of the fingerprint test. Findings from Chapter 3 also provided converging evidence that a domain general ability in feature comparison tasks underpins fingerprint comparison accuracy. Here, Experts and Trainees both outperformed Novices in the pattern matching test (of Block 1), which again involved non-fingerprint stimuli. And Trainees were more accurate than Novices in three out of four blocks of the tests (pattern, fingerprint image and latent print matching) despite their relative inexperience with fingerprint stimuli.

Recent research has reported a strong relationship across feature comparison tasks and fingerprint matching in novice observers (Growns et al., 2022). Similarly, forensic fingerprint examiners have previously been more accurate than novices in face matching tasks, which further suggests an advantage in feature comparison tasks that extends beyond experience with fingerprint stimuli (Phillips et al., 2018). Taken together, these findings and those from this chapter suggest that performance in fingerprint comparison may be underpinned by a domain-general ability in feature comparison tasks.

Also emerging from Chapter 4 was a likely role for global and local processing during fingerprint examination. This aspect of visual processing has not been widely explored, and holistic processing of fingerprints is only partly supported by research (Vogelsang et al., 2017). Here, accuracy in global perceptual tasks predicted accuracy in latent print matching (Block 4) of the fingerprint test, perhaps due to contextual information contained within the whole image serving as a guide to the relevant target area. Previous research lends weight to this theory, with fingerprint examiners faster and more accurate in locating a target area on a corresponding exemplar if the original target and surrounding print were visible (Hicklin et al., 2019).

Chapter 4 also revealed a strong relationship between local visual processing and accuracy in pattern matching (Block 1) and palmprint matching (Block 3) of the fingerprint test. The global shape of the Navon letter typically disrupts local processing (i.e., correct identification of the smaller letter comprising the global shape). Therefore, those observers who can ignore the overall shape/pattern of an image, and focus on the features comprising the image, are more likely to excel in certain fingerprint comparison tasks.

Differences in response times was another factor that emerged during Chapter 4, with Novices responding more quickly than police observers in most of the feature matching tasks. However, untrained observers were less accurate than Trainees and Experts across all of these tasks (KFMT, MRT, MFFT), and slower responses in the KFMT and the MRT were associated with performance gains in the fingerprint test. This suggests that a slower, more analytical, approach to feature comparison tasks is therefore likely to benefit accuracy. Conversely, in the visual search test, with speed of correct responses used as a measure of ability, Novices outperformed the police examiners. Slower responses by Experts and Trainees may represent a preference for a more careful or methodological approach to the task.

Finally, as the Experts and Trainees undertook the test battery in their role as forensic examiners, the importance of test taking motivation was also considered with task engagement perhaps higher in these observers in an effort to maintain their professional standing. However, the link between intrinsic motivation (as measured by the IMI), was not robust, with only Experts reporting higher interest in the test battery than Novices. The role of motivation on test performance was largely inconsistent across the test battery, which tends to support findings from recent research in which performance in feature comparison tasks was not explained by task motivation (Growns et al., 2022).

5.2 Theoretical Implications

Several novel theoretical insights emerge from the findings of this thesis. Firstly, incorporating complex latent prints into the fingerprint test clearly differentiated the abilities of experienced examiners, trainee examiners, and novices, with high performance in this task likely to be a characteristic marker of expertise. This demonstrates that it is possible to measure expertise in fingerprint examiners, on the proviso that any tests of competence are suitably challenging. This is an important consideration as familiarity with fingerprints in workplace comparisons means that experienced examiners are likely to outperform less experienced, or inexperienced, observers in *any* test of fingerprint comparison. This finding has important theoretical implications: Any theory of fingerprint comparison expertise must *primarily* explain high accuracy in challenging latent print comparisons, such as those incorporated into the fingerprint test in this thesis.

In parallel, such a theory of fingerprint expertise must also capture *individual differences* in ability, the importance of which have been highlighted in this thesis. Previous research has tended to focus on the differences in fingerprint comparison accuracy between novices and examiners as a measure of expertise (Searston & Tangen, 2017a, 2017b ; Stevenage & Pitfield, 2016; Tangen et al., 2011; Thompson et al., 2014). This overlooks the body of evidence which shows that a range of ability is typically observed in perceptual tasks (e.g., see Gauthier, 2018; White et al., 2010). Consequently, reliance on between group comparisons may actually conceal any high performers within an untrained population.

This role of individual difference in ability has particular significance, given that reviews such as PCAST (2016) and NAS (2009) have highlighted concerns about the validity of forensic examiners' subjective interpretations of the fingerprint evidence. This must raise an important question of how we can be sure that an examiner's expertise reflects knowledge and ability that is not found in the general population (Roberts, 2021) if we do not measure how individuals within the general population perform in these tasks. This thesis has shown that comparing the accuracy of individual novice observers with that of experienced and trainee examiners has identified a level of performance in latent print comparison that virtually no novice was able to match, let alone exceed. If fingerprint examiner testimony is to retain its probative value within the criminal justice system (Ribeiro et al., 2019; Schweitzer & Saks, 2007) then fingerprint tests which clearly separate the performance of examiners and untrained observers should be the operational definition of expertise. In turn, any theory of fingerprint expertise must capture a clear separation in the performance of experienced examiners and observers with limited, or no, training at the level of the *individual*. Such a theory would therefore preclude any overlap in performance between these groups: expertise is demonstrated through fingerprint examiners high levels of performance in fingerprint tasks that 'ordinary' members of the public are unable to attain.

Finally, this thesis has revealed a benefit to using a battery of tests to understand fingerprint expertise. From these test results, a *cognitive* theory of fingerprint expertise can be proposed in which high performance in fingerprint comparison is underpinned by an ability to compare features in non-fingerprint stimuli, and an ability to mental rotate pairs of objects, other than fingerprints, in order to match them. In addition, the test battery has identified a relationship between global processing and latent print comparison. Perhaps surprisingly, it also revealed that other perceptual tasks, such as a visual search ability and visual short-term memory, were not associated with performance in fingerprint comparison.

5.3 Limitations and Future Directions

The research in this thesis was subject to some limitations. One important consideration relates to the accuracy of the fingerprint examiners and trainees, all of whom made errors at some point during the fingerprint test. It is therefore essential to reiterate that this research was designed to measure and understand expertise in these professional examiners, rather than being a study of the nature of errors in fingerprint comparison ability. A relevant factor here is that these observers were unable to access magnification and other tools that they use in forensic comparisons and had no means of using ACE notation on any of the stimuli. If the tasks contained within the fingerprint test were undertaken as workplace comparisons, these may reveal even higher levels of accuracy for the professional examiners.

Although the test battery identified a clear relationship between fingerprint comparison accuracy and perceptual ability, this outcome also reflects some limitations in this experiment. The first of these relates to the incorporation of data from a mostly non-expert sample into the regression analyses in Chapter 4. As there were few Experts who had taken part in both the fingerprint test *and* the test battery, it was decided to combine data from all observers rather than trying to compare differences between groups. Although this approach reflected the inclusion of high scores from the fingerprint test, data from a larger sample of experts and trainees would allow the relationship between fingerprint accuracy and perceptual ability to be explored in more depth. Moreover, by comparing performance between groups, this may reveal differences in the relationship between perceptual and fingerprint accuracy, and further our understanding of the nature of fingerprint expertise.

An additional consideration is whether the absence of a relationship between visual short-term memory (VSTM) and fingerprint comparison that was revealed in this thesis would be replicated if an alternative method of measuring VSTM was included in the test battery. Research suggests that memory is a key component of fingerprint comparison (Busey et al., 2010; Hicklin et al., 2019). However, because observers needed to correctly identify two different objects in the VSTM test of Chapter 4 this may, in hindsight, have been too complex to accurately capture differences between the groups and any relationship with fingerprint accuracy. A more robust examination of the role of VSTM in fingerprint comparison may therefore be revealed with a simpler measure, such as a delayed matching-to-sample test with non-fingerprint stimuli.

Finally, several future directions for research emerge from the current work. For example, this thesis has demonstrated that accuracy in complex latent print comparisons differentiates the performance of experienced examiners, examiners-in-training, and untrained observers. But comparing the performance of experienced examiners and those with less, or no experience, will not further our understanding of expertise unless the to-be-compared fingerprints are sufficiently challenging. In turn, the question arises of whether examiner expertise with such complex fingerprints is completed unrivalled or whether this can be matched or exceeded by some alternatives. Fingerprint matching algorithms, for example, are already in use by police and forensic services to create shortlists of possible fingerprint matches for fingerprint examiners to inspect. In future, it will be interesting to see how examiner accuracy compares with such algorithms with latent fingerprints in a systematic scientific study. Fingerprint matching algorithms are subject to constant review and benchmarking (see Fingerprint Verification Competition, FVC-onGoing, 2022), and a wide range of datasets and fingerprint stimuli are now available. A *direct* comparison of the best algorithms with examiner performance should provide important context for *human* expertise in this domain.

This thesis also illustrates that a battery comprised of fingerprint tests, feature matching tasks with non-fingerprint stimuli, and tests of mental rotation could assist in the recruitment of applicants to fingerprint examiner roles. Newly appointed examiners undergo a lengthy, and costly, programme of training and mentorship prior to becoming court-practicing fingerprint examiners. However, the selection of trainees for fingerprint comparison can also capture individuals who ultimately do not have the ability to qualify as fingerprint examiners (Hall, personal communication). The incorporation of this fingerprint test and cognitive battery into the selection process may therefore assist in identifying those applicants with a higher aptitude for fingerprint comparison. Future experiments on trainee selection with a set of tests that also measure the cognitive facets of fingerprint expertise that were identified in this thesis will reveal whether this can lead to tangible improvements in this domain.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*(1), 3–27. https://doi.org/10.1037/0033-2909.102.1.3

Anthonioz, A., Egli, N., Champod, C., Neumann, C., Puch-Solis, R., & Bromage-Griffiths, A. (2008). Level 3 details and their role in fingerprint identification: a survey among practitioners. *Journal of Forensic Identification*, *58*(5), 562. https://www.ojp.gov/ncjrs/virtual-library/abstracts/level-3-details-and-their-role-fingerprint-identification-survey

Ashbaugh, D. R. (1999). Quantitative-qualitative friction ridge analysis : an introduction to basic and advanced ridgeology. Boca Raton, Fla.: CRC Press. http://doi.org//10.1201/9781420048810

Barnes, J. G., Drazen, J. M., Rennard, S., & Thomson, N. C. (2011). History. In A. McRoberts(Ed.), *The Fingerprint Sourcebook* , (pp. 1–22). National Justice Institute. https://www.ojp.gov/pdffiles1/nij/225320.pdf

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., ... & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive research: principles and implications*, *3*(1), 1-19. http://doi.org/10.1186/s41235-018-0116-5

Bécue, A., Eldridge, H., & Champod, C. (2020). Interpol review of fingermarks and other body impressions 2016–2019. *Forensic Science International: Synergy*. http://doi.org/10.1016/j.fsisyn.2020.01.013

Byrd, J. S., & Bertram, D. (2003). Form-blindness. *Journal of Forensic Identification*, *53*(3), 315. https://www.proquest.com/scholarly-journals/form-blindness/docview/194793498/se-2

Bigun, J. (2014). Encyclopedia of Biometrics. *Encyclopedia of Biometrics*, 1–13. http://doi.org/10.1007/978-3-642-27733-7

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*(3), 277. http://doi.org/10.1037/a0029635

Borchert, K. (2020). *Inquisit Test Library*. [Computer software]. https://www.millisecond.com/download/library

Bruce, V., Bindemann, M., & Lander, K. (2018). Individual differences in face perception and person recognition. *Cognitive Research: Principles and Implications*, *3*(1), 1-3. http://doi.org/10.1186/s41235-018-0109-4

Bobak, A. K., Hancock, P. J. B., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology*, *30*(1), 81–91. http://doi.org/10.1002/acp.3170

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior research methods*, *42*(1), 286-291. http://doi.org/10.3758/BRM.42.1.286

Busey, T. A., & Parada, F. J. (2010). The nature of expertise in fingerprint examiners. *Psychonomic Bulletin and Review*, *17*(2), 155–160. http://doi.org/10.3758/PBR.17.2.155

Busey, T. A., & Vanderkolk, J. R. (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research*, *45*(4), 431–448. http://doi.org/10.1016/j.visres.2004.08.021

Champod, C., Lennard, C. J., Margot, P., & Stoilovic, M. (2004). *Fingerprints and other ridge skin impressions*. CRC press. http://doi.org/10.1201/9780203485040

Cole, S. A. (2008). The ‘opinionization’ of fingerprint evidence. *BioSocieties*, *3*(1), 105–113. http://doi.org/10.1017/s1745855208006030

Cooper, G. S., & Meterko, V. (2019). Cognitive bias research in forensic science: A systematic review. *Forensic Science International*, *297*, 35–46. http://doi.org/10.1016/j.forsciint.2019.01.016

Crown Prosecution Service (CPS) (2019). *Expert Evidence: The code for crown prosecutors*. Retrieved from https://www.cps.gov.uk/legal-guidance/expert-evidence

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology, 30*(6), 827-840. http://doi.org/10.1002/acp.3260

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a ‘target’ comparison. *Forensic Science International*, *208*(1–3), 10–17. https://doi.org/10.1016/j.forsciint.2010.10.013

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification*, *56*(4), 600–616. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5c6e40959915ff0a97c789d7606fcd57d53dc342

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, *156*(1), 74–78. https://doi/10.1016/j.forsciint.2005.10.017

Dror, I. E., Péron, A. E., Hind, S. L., & Charlton, D. (2005). When emotions get the better of us: The effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology*, *19*(6), 799–809. https://doi/10.1002/acp.1130

Edmond, G., Towler, A., Growns, B., Ribeiro, G., Found, B., White, D., ... & Martire, K. (2017). Thinking forensics: Cognitive science for forensic practitioners. *Science & Justice*, *57*(2), 144-154. http://doi/10.1016/j.scijus.2016.11.005

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is “special” about face perception? *Psychological Review*, *105*(3), 482–498. http://doi/10.1037/0033-295X.105.3.482

Forensic Science Regulator (FSR). (2017). Codes of Practice and Conduct Fingerprint Comparison, (2), 44. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/638254/128_FSR_fingerprint_appendix__Issue2.pdf

Fraser-MacKenzie, P. A. F., Dror, I. E., & Wertheim, K. (2013). Cognitive and contextual influences in determination of latent fingerprint suitability for identification judgments. *Science & Justice*, *53(2)*. https://doi.org/10.1016/j.scijus.2012.12.002

FVC-onGoing (2022). On-line evaluation of fingerprint recognition algorithms. https://biolab.csr.unibo.it/fvcongoing/UI/Form/Home.aspx

Fysh, M. C., & Bindemann, M. (2018). The Kent face matching test. *British journal of psychology*, *109*(2), 219-231.https://doi.org/10.1111/bjop.12260

Galton, F. (1892). Finger Prints. London: MacMillan and Co.

Ganis, G., & Kievit, R. A New Set of Three-Dimensional Shapes for Investigating Mental Rotation Processes: Validation Data and Stimulus Set. *Journal of Open Psychology Data, 3.* http://doi.org/10.5334/jopd.ai

Gauthier, I., & Tarr, M. (1997). Becoming a “Greeble” expert: exploring mechanisms for face recognition. Vision Research, 37, 1673-1682 http://doi.org//10.1016/S0042-6989(96)00286-6

Gould, J. B., & Leo, R. A. (2010). One hundred years later: Wrongful convictions after a century of research. *Journal of Criminal Law and Criminology*, *100*(3), 825–868. https://www.jstor.org/stable/25766110

Growns, B., Dunn, J. D., Mattijssen, E. J., Quigley-McBride, A., & Towler, A. (2022). Match me if you can: Evidence for a domain-general visual comparison ability. *Psychonomic Bulletin & Review*, 1-16. http://doi.org/10.3758/s13423-021-02044-2

Haber, L., & Haber, R. N. (2007). Scientific validation of fingerprint evidence under Daubert. *Law, Probability and Risk*, *7*(2), 87–109. http://doi.org/10.1093/lpr/mgm020

Hall, L. J., & Player, E. (2008). Will the introduction of an emotional context affect fingerprint analysis and decision-making? *Forensic Science International*, *181*(1-3), 36-39. https://doi.org/10.1016/j.forsciint.2008.08.008

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in cognitive sciences*, *7*(11), 498-504. https://doi.org/10.1016/j.tics.2003.09.006

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition, 33*(1), 98-106. https://doi.org/10.3758/BF03195300

Hicklin, R. A., Ulery, B. T., Busey, T. A., Roberts, M. A., & Buscaglia, J. (2019). Gaze behavior and cognitive states during fingerprint target group localization. *Cognitive Research: Principles and Implications*, *4*(1). https://doi.org/10.1186/s41235-019-0160-9

Home Office. (2019). *National DNA Database Strategy Board Annual Report 2017/18*. https://assets.publishing.service.gov.uk/government/uploads/system//uploads/attachment_data/file/778064/National_DNA_database_annual_report__2017-18_web.pdf

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. Journal of Applied Research in Memory and Cognition, 3, 72–76. https://doi.org/10.1016/j.jarmac.2014.04.009

Huber, R. (1959). Expert Witness. *The Criminal Law Quarterly 1959-1960*, *2*, 276-295.

Jain, A. K., Prabhakar, S., & Pankanti, S. (2002). On the similarity of identical twin fingerprints. *Pattern Recognition*, *35*(11), 2653–2663. https://doi.org/10.1016/S0031-3203(01)00218-7

Jain, A., Chen, Y., & Demirkus, M. (2006). Pores and ridges: Fingerprint matching using level 3 features. In *18th International Conference on Pattern Recognition (ICPR'06)* (Vol. 4, pp. 477-480). IEEE. https://doi.org/10.1109/ICPR.2006.938

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1671-1683. https://doi.org/10.1098/rstb.2010.0379

Kadane, J. B. (2018). Fingerprint science. *Annals of Applied Statistics*, *12*(2), 771–787. https://doi.org/10.1214/18-AOAS1140

Kagan, J. (1966) Reflection-impulsivity: The generality and dynamics of conceptual tempo. *Journal of Abnormal Psychology, 71(*1), 17-24. https://doi.org/10.1037/h0022886

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kellman, P. J., Mnookin, J. L., Erlikhman, G., Garrigan, P., Ghose, T., Mettler, E., ... & Dror, I. E. (2014). Forensic comparison and matching of fingerprints: using quantitative image measures for estimating error rates through understanding and predicting difficulty. *PloS one*, *9*(5), e94617. https://doi.org/10.1371/journal.pone.0094617

Kücken, M., & Champod, C. (2013). Merkel cells and the individuality of friction ridge skin. *Journal of Theoretical Biology*, (317), 229–237. https://doi.org/10.1016/j.jtbi.2012.10.009

Lander, K., Bruce, V., & Bindemann, M. (2018). Use-inspired basic research on individual differences in face identification: Implications for criminal investigation and security. *Cognitive Research: Principles and Implications*, *3*(1), 1-13. https://doi.org/10.1186/s41235-018-0115-6

Langenburg, G., Champod, C., & Genessay, T. (2012). Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools. *Forensic science international*, *219*(1-3), 183-198. https://doi.org/10.1016/j.forsciint.2011.12.017

Mattijssen, E. J., Witteman, C. L., Berger, C. E., Zheng, X. A., Soons, J. A., & Stoel, R. D. (2021). Firearm examination: Examiner judgments and computer-based comparisons. *Journal of forensic sciences*, *66*(1), 96-111. https://doi.org/10.1111/1556-4029.14557

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences* (Vol. 6). https://doi.org/10.1016/S1364-6613(02)01903-4

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & cognition*, *34*(4), 865-876. https://doi.org/10.3758/BF03193433

Meuwly, D. (2014). Forensic use of fingermarks and fingerprints. In *Encyclopedia of Biometrics* (pp. 1-15). Springer. https://do.org/10.1007/978-3-642-27733-7

Mnookin, J. L. (2007). The validity of latent fingerprint identification: confessions of a fingerprinting moderate. *Law, Probability and Risk*, *7*(2), 127–141. https://doi.org/10.1093/lpr/mgm022

National Academy of Science (NAS) (2009). Strengthening forensic science in the United States. *The Committee on Identifying the Needs of Forensic Sciences Community*, 1–328. https://doi.org/10.17226/12589

National Academies of Sciences (NAS), (2017). *Personnel selection in the pattern evidence domain of forensic science. National Academies Press*. https://doi.org/10.17226/23681

Neumann, C., Champod, C., Yoo, M., Genessay, T., & Langenburg, G. (2014). Improving the understanding and the reliability of the concept of" sufficiency" in friction ridge examination. *NIJ Publication Update,* 1-97. https://www.ojp.gov/library/publications/improving-understanding-and-reliability-concept-sufficiency-friction-ridge

Noyes, E., Phillips, P. J., & O’Toole, A. J. (2017). What is a super-recogniser? In *Face processing: Systems, disorders and cultural differences* (pp. 173-201). Nova Science Publishers Inc.

OIG (2006). Review of the FBI’s handling of the Brandon Mayfield case. *Office of the Inspector General, Oversight and Review Division, US Department of Justice,* 1-330. https://oig.justice.gov/sites/default/files/archive/special/s0601/final.pdf

Osborn, A. S. (1939). Form Blindness and proof (sight defects in relation to the administration of justice). *Journal of Criminal Law and Criminology (1931-1951)*, *30*(2), 243. https://doi.org/10.2307/1137080

Pacheco, Cerchiai & Stoiloff (2014). Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations. https://www.ojp.gov/ncjrs/virtual-library/abstracts/miami-dade-research-study-reliability-ace-v-process-accuracy

Pankanti, S., Prabhakar, S., & Jain, A. K. (2002). On the individuality of fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(8), 1010–1025. https://doi.org/10.1109/TPAMI.2002.1023799

Pauli, R. (1993). *Learning to read mammograms: Complex skill acquisition, training, and individual differences*. University of Surrey (United Kingdom). https://openresearch.surrey.ac.uk/esploro/outputs/doctoral/Learning-to-read-mammograms-Complex-skill-acquisition-training-and-individual-differences/99516115102346

Poyser, A.; Milne, R. (2010). No grounds for complacency and plenty for continued vigilance: miscarriages of justice as drivers for research on reforming the investigative interviewing process. *Article for British Journal of Forensic Practice 2010*. https://doi.org/10.1017/CBO9781107415324.004

President's Council of Advisors on Science and Technology (PCAST) (US). (2016). Report to the president, forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. *Executive Office of the President of the United States, President's Council.* Retrieved from www.whitehouse.gov/ostp/pcast

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372. https://doi.org/10.1037/0033-2909.124.3.372

Ribeiro, G., Tangen, J. M., & McKimmie, B. M. (2019). Beliefs about error rates and human judgment in forensic science. *Forensic Science International*, *297*, 138–147. https://doi.org/10.1016/j.forsciint.2019.01.034

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin*. https://doi.org/10.1037/a0037004

Roberts, A. (2021). A Legal Perspective. In M. Bindemann (Ed.), *Forensic Face Matching: Research and Practice.* Oxford University Press. https://doi.org/10.1093/oso/9780198837749.001.0001

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin and Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54-67. https://doi.org/10.1006/ceps.1999.1020

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892–895. https://doi.org/10.1126/science.1111565

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of modern applied statistical methods*, *8*(2), 26. https://doi.org/10.22237/jmasm/1257035100

Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST). (2013). Standards for examining friction ridge impressions and resulting conclusions (Latent/Tenprint). http://www.clpex.com/swgfast/documentsconclusions/130427_Examinations-Conclusions_2.0.pdf

Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST). (2013a). Standard terminology of friction ridge examination (Latent/Tenprint). http://www.clpex.com/swgfast/documents/terminology/121124_Standard-Terminology_4.0.pdf

Schiffer, B., & Champod, C. (2007). The potential (negative) influence of observational biases at the analysis stage of fingermark individualisation. *Forensic Science International*, *167*(2–3), 116–120. https://doi.org/10.1016/j.forsciint.2006.06.036

Schweitzer, N. J., & Saks, M. J. (2007). The CSI effect: Popular fiction about forensic science affects the public's expectations about real forensic science. *Jurimetrics*, *47*, 357–364. http://www.jstor.org/stable/29762978

Searston, R. A., & Tangen, J. M. (2017a). Expertise with unfamiliar objects is flexible to changes in task but not changes in class. *PLoS ONE*, *12*(6), 1–14. https://doi.org/10.1371/journal.pone.0178403

Searston, R. A., & Tangen, J. M. (2017b). The style of a stranger: Identification expertise generalizes to coarser level categories. *Psychonomic Bulletin and Review*, *24*(4), 1324–1329. https://doi.org/10.3758/s13423-016-1211-6

Searston, R. A., & Tangen, J. M. (2017c). The Emergence of Perceptual Expertise with Fingerprints Over Time. *Journal of Applied Research in Memory and Cognition*, *6*(4), 442–451. https://doi.org/10.1016/j.jarmac.2017.08.006

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703. https://doi.org/10.1126/science.171.3972.701

Sita, J., Found, B., & Rogers, D. K. (2002). Forensic handwriting examiners' expertise for signature comparison. *Journal of Forensic Sciences*, *47*(5), 1117-1124. https://doi.org/10.1520/jfs15521j

Stammers, S., & Bunn, S. (2015). Unintentional bias in forensic investigation. *POST brief, In: Parliamentary Office of Science and Technology. London*, 1-6.

Stevenage, S. V, & Bennett, A. (2017). A biased opinion: Demonstration of cognitive bias on a fingerprint matching task through knowledge of DNA test results. *Forensic Science International*, *276*, 93–106. https://doi.org/10.1016/j.forsciint.2017.04.009

Stevenage, S. V., & Pitfield, C. (2016a). Fact or friction: Examination of the transparency, reliability and sufficiency of the ACE-V method of fingerprint analysis. *Forensic Science International*. https://doi.org/10.1016/j.forsciint.2016.08.026

Stevenage, S. V., & Pitfield, C. (2016b). Data from a fingerprint matching task with experts, trained students and untrained novices. *Data in Brief*, *9*, 621–624. https://doi.org/10.1016/j.dib.2016.09.022

Tangen, J. M., Kent, K. M., & Searston, R. A. (2020). Collective intelligence in fingerprint analysis. *Cognitive Research: Principles and Implications*, *5*(1). https://doi.org/10.1186/s41235-020-00223-8

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science*, *22*(8), 995–997. https://doi.org/10.1177/0956797611414729

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Expertise in fingerprint identification. *Journal of Forensic Sciences, 58*(6), 1519-1530. https://doi.org/10.1111/1556-4029.12203

Thompson, M. B., & Tangen, J. M. (2014). The nature of expertise in fingerprint matching: Experts can do a lot with a little. *PLoS ONE*, *9*(12), 1–23. https://doi.org/10.1371/journal.pone.0114759

Thompson, W. C., Grady, R. H., Lai, E., & Stern, H. S. (2018). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, *17*(2), 133–155. https://doi.org/10.1093/lpr/mgy012

Towler, A., White, D., Ballantyne, K., Searston, R. A., Martire, K. A., & Kemp, R. I. (2018). Are forensic scientists experts? *Journal of Applied Research in Memory and Cognition*, *7*(2), 199-208. https://doi.org/10.1016/j.jarmac.2018.03.010

Tully, G. (2021). *Forensic Science Regulator Annual Report 2020*. https://www.gov.uk/government/publications/forensic-science-regulator-annual-report-2020

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. *PLoS ONE*, *9*(11), 110179. https://doi.org/10.1371/journal.pone.0110179

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2010). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7733–7738. https://doi.org/10.1073/pnas.1018707108

Ulery, B. T., Hicklin, R. A., Buscaglia, J. A., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE*, *7*(3). https://doi.org/10.1371/journal.pone.0032800

Ulery, B. T., Hicklin, R. A., Kiebuzinski, G. I., Roberts, M. A., & Buscaglia, J. A. (2013). Understanding the sufficiency of information for latent fingerprint value determinations. *Forensic Science International*, *230*(1–3), 99–106. https://doi.org/10.1016/j.forsciint.2013.01.012

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. A. (2015). Changes in latent fingerprint examiners' markup between analysis and comparison. *Forensic Science International*, *247*(1), 54–61. https://doi.org/10.1016/j.forsciint.2014.11.021

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. A. (2016). Data on the interexaminer variation of minutia markup on latent fingerprints. *Data in Brief*, *8*, 158–190. https://doi.org/10.1016/j.dib.2016.04.068

Vanderkolk, J. R. (2011). Examination Process. The Fingerprint Sourcebook. *US Department of Justice*. https://doi.org/10.1016/j.jflm.2011.12.018

Vogelsang, M. D., Palmeri, T. J., & Busey, T. A. (2017). Holistic processing of fingerprints by expert forensic examiners. *Cognitive Research: Principles and Implications*. https://doi.org/10.1186/s41235-017-0051-x

Vokey, J. R., Tangen, J. M., & Cole, S. A. (2009). On the preliminary psychophysics of fingerprint identification. *Quarterly Journal of Experimental Psychology*, *62*(5), 1023-1040. https://doi.org/10.1080/17470210802372987

Wertheim, K., Langenburg, G., & Moenssens, A. (2006). A report of latent print examiner accuracy during comparison training exercises. *Journal of forensic identification*, *56*(1), 55.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS one*, *9*(8), e103510. https://doi.org/:10.1371/journal.pone.0103510

Wilhelm, O., Hildebrandt, A., Manske, K., Schacht, A., & Sommer, W. (2014). Test battery for measuring the perception and recognition of facial expressions of emotion. *Frontiers in psychology*, *5*, 404. https://doi.org/10.3389/fpsyg.2014.00404

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., … Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, *107*(11), 5238–5241. https://doi.org/10.1073/pnas.0913053107

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10-65. https://doi.org10.1177/1529100616686966

Wong, Y. K., & Gauthier, I. (2010). Holistic processing of musical notation: Dissociating failures of selective attention in experts and novices. *Cognitive, Affective and Behavioral Neuroscience*, *10*(4), 541–551. https://doi.org/10.3758/CABN.10.4.541

Yu, C., Busey, T., & Vanderkolk, J. (2011). Discovering correspondences between fingerprints based on the temporal dynamics of eye movements from experts. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6540 LNCS*(May 2014), 160–172. https://doi.org/10.1007/978-3-642-19376-7_14

Appendix

Summary of non-group main effects and interactions from mixed-model ANOVA analyses in Chapters 3 and 4.

Analysis		F	*p*	partial η^2
Chapter 3				
ANOVA	Accuracy by Block	49.49	<.001	0.281
	RT by Block	60.10	<.001	0.316
	Target Misses by Block	7.46	<.001	0.054
Chapter 4				
ANOVA	Mental Rotation Test (MRT)			
	Accuracy by Trial Type (same or different)	39.12	<.001	0.241
	Accuracy by Rotation (0, 50, 100, 150)	120.46	<.001	0.495
	Accuracy by Trial Type x Rotation	56.97	<.001	0.317
	RT by Trial Type (same or different)	39.12	<.001	0.241
	RT by Rotation (0, 50, 100, 150)	120.46	<.001	0.495
	RT by Trial Type x Rotation	56.97	<.001	0.317
	Matching Familiar Figures Test (MFFT)			
	Accuracy by Trial Type (same or different)	7.69	.01	0.058
	RT by Trial Type (same or different)	1.48	.23	0.012
	Visual Search			
	Accuracy by Trial Type (present or absent)	1.77	.19	0.014
	Accuracy by Array Size (1, 5, 15, 30)	0.45	.72	0.004
	Accuracy by Trial Type x Array Size	1.88	.13	0.015
	RT by Trial Type	3.72	.06	0.029
	RT by Array Size (1, 5, 15, 30)	1.26	.29	0.010
	RT by Trial Type x Array Size	0.70	.55	0.006
	Navon Letters			
	Accuracy by Trial Type (global or local)	134.26	<.001	0.520
	Accuracy by Condition (consistent or conflicting)	8.85	.00	0.067
	Accuracy by Trial Type x Condition	7.28	.01	0.055
	RT by Trial Type (global or local)	40.52	<.001	0.248
	RT by Condition (consistent or conflicting)	0.90	.35	0.007
	RT by Trial Type x Condition	0.25	.62	0.002
	Kent Face Matching Test (KFMT)			
	Accuracy by Trial Type	2.03	.16	0.016
	RT by Trial Type	0.94	.66	0.002
	Intrinsic Motivation Inventory (IMI)			
	Score by Subscale (Interest, Competence, Choice, Pressure)	42.01	<.001	0.253