THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

ON EXPERTISE IN FINGERPRINT IDENTIFICATION

MATTHEW B. THOMPSON

B.Inf.Tech., B.Sc. (Hons)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2013

School of Psychology

Abstract

Little is known about the nature and development of fingerprint expertise and, therefore, the best way to turn novices into experts. Little is known about the factors that affect matching accuracy and, therefore, what experts can legitimately testify to in court. This thesis explores the factors that affect matching accuracy and the development of expertise in fingerprint identification, in order to inform training, and to provide an empirical basis for expert testimony in the courtroom. The investigation is grounded in exemplar, signal detection, and dual-process theories, and draws from literature on expertise and skill acquisition, and familiar and unfamiliar face recognition.

The thesis comprises four parts. In Part 1—Establishing Expertise—I attempt to find evidence for expert-novice differences in fingerprint matching, and explore where performance differences might lie. In Part 2—Depicting Expertise—I explore alternate methods for presenting signal detection results by depicting the data in a contingency space. In Part 3—Nature of Expertise—I explore the cognitive processes that might account for the superior performance of expert fingerprint examiners, and I explore the limits of rapid expert decision making. In Part 4—Expression of Expertise—I propose a framework for the expression of expert opinion in the courtroom, in order to integrate extra-legal recommendations and emerging research.

Taken together, I find that qualified, court-practicing fingerprint examiners are more accurate and more conservative than novices, and that errors are more likely to occur on prints from large databases, which are highly similar but nonmatching. I find that performance, both in terms of accuracy and response bias, changes as people move from novice, to trainee, to expert. I find that experts can discriminate matching and nonmatching prints that are

artificially noisy, spaced by a short time, briefly flashed on screen, and even when presented in the blink of an eye. These findings indicate that experts make use of non-analytic processing when identifying prints, and they perform accurately when information is sparse—experts can do a lot with a little. Further programs of research, like this one, on the factors that affect fingerprint matching accuracy and performance, will create a foundation for evidence-based training, and serve to increase confidence in the legitimacy of claims made by expert witnesses in the courtroom.

Declaration by Author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the *Copyright Act 1968*.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

0.1 Publications During Candidature

0.1.1 Journal Publications

Edmond, G., **Thompson, M. B.**, & Tangen, J. M. (2013). A guide to interpreting forensic testimony: Scientific approaches to fingerprint evidence. *Law, Probability & Risk*. doi: 10.1093/lpr/mgt011

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Human matching performance of genuine crime scene latent fingerprints. *Law and Human Behavior*. doi: 10.1037/lhb0000051

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

Tangen, J. M., Murphy, S. C., & **Thompson, M. B.** (2011). Flashed face distortion effect: Grotesque faces from relative spaces. *Perception, 40*, 628–630. doi: 10.1068/p6968

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

0.1.2 Conference Publications

Thompson, M. B., Tangen, J. M., & McCarthy D. J. (2013). Decision making, expertise, and non-analytic cognition in human fingerprint matching. *Proceedings of The 54th Annual Meeting of the Psychonomic Society.* Toronto, Canada: 14–18 November, 2013.

Thompson, M. B., Tangen, J. M., & McCarthy D. J. (2013). Expertise, memory, and non-analytic cognition in fingerprint matching: Experts can discriminate prints in noise, spaced in time, and in the blink of an eye. *Proceedings of the 40th Australasian Experimental*

Psychology Conference. Adelaide, Australia: 3–4 April, 2013.

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Evidence for expertise in fingerprint identification and the ramifications for the future study of forensic expertise. *American Academy of Forensic Sciences (AAFS) Annual Meeting.* Washington, DC: 18–23 February, 2013.

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2012). Evidence for expertise and accuracy in fingerprint identification. *Proceedings of the 21st International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society (ANZFSS).* Hobart, Australia: 23–27 September, 2012.

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2012). Evidence for expertise in the matching performance of human fingerprint examiners. *Proceedings of the 6th European Academy of Forensic Science Conference (EAFS).* The Hague, the Netherlands: 20–24 August, 2012.

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2012). Evidence for expertise in fingerprint identification. *Proceedings of the 97th International Educational Conference of the International Association for Identification.* Phoenix, Arizona: 22–28 July, 2012.

Tangen, J. M., Murphy, S. C., & **Thompson, M. B.** (2012). When pretty girls turn ugly: The flashed face distortion effect. *12th Annual Meeting of the Vision Science Society (VSS).* Naples, Florida: 14 May, 2012.

Sung, B. C. Y., **Thompson, M. B**. & Tangen, J. M. (2012). When pretty girls turn ugly: The boundary conditions of the flashed face distortion effect. *Proceedings of the 39th Australasian Experimental Psychology Conference.* Sydney, Australia: 12–15 April, 2012.

Tangen, J. M., **Thompson, M. B.**, & McCarthy, D. J. (2011). Forensic reasoning and uncertainty: Identifying fingerprint expertise. *Impressions and Expressions: Expert Opinion Evidence in Reports and Courts, AAFS Conference.* Sydney, Australia: 3–4 December, 2011.

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2011). Accuracy and expertise in human fingerprint identification. *Proceedings of the 38th Australasian Experimental Psychology Conference.* Auckland, New Zealand: 28-30 April, 2011.

Thompson, M. B., Tangen, J. M., & McCarthy, D. (2010). Enhancing performance in human decision making: The role of similarity in forensic identification. *Proceedings of the 20th International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society (ANZFSS).* Sydney, Australia: 5–9 April.

Tangen, J. M., **Thompson, M. B.**, McCarthy, D., & Tear, M. J. (2010). Ground truth: On certainty in forensic decision-making research. *Proceedings of the 20th International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society (ANZFSS).* Sydney, Australia: 5–9 April.

Thompson, M. B., Tangen, J. M., Treloar, R., & Ivison, K. J. (2010). Humans matching fingerprints: Sequence and Size. *Proceedings of the 54th Annual Meeting of the Human Factors and Ergonomics Society.* San Francisco, CA: September 27–October 1.

Thompson, M. B., Tangen, J. M., Ivison, K. J., & Treloar, R. (2010). Expertise in matching fingerprints and faces. *Proceedings of the 37th Australasian Experimental Psychology Conference.* Melbourne, Australia: 8–10 April.

Tear, M. J., **Thompson, M. B.**, & Tangen J. M. (2010). The importance of ground truth: An open-source biometric repository. *Poster presented at the 54th Annual Meeting of the Human Factors and Ergonomics Society.* San Francisco, CA: September 27–October 1.

Thompson, C., Sanderson, P., Watson, M., **Thompson, M. B.**, Muthukrishna, M., & Murphy, S. (2010). Testing auditory alarm effectiveness with three different alarm sets. *Poster presented at the Australian and New Zealand College of Anaesthetists Annual Scientific Meeting (ANZCA ASM).* Christchurch, NZ: 1–5 May 2010.

0.2 Publications Included in this Thesis

0.2.1 Journal Publications

Incorporated as Chapter 2:

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

Contributor	*Statement of contribution*
Matthew B. Thompson (Candidate)	Designed experiment (50%) Collected data (60%) Analysed data (60%) Wrote the manuscript (50%)
Jason M. Tangen	Designed experiment (50%) Collected data (40%) Analysed data (40%) Wrote the manuscript (50%)
Duncan J. McCarthy	Police database search (100%)

This chapter was published in the journal *Psychological Science* and contributions to design, data collection, analysis, writing, etc., were shared equally between Jason Tangen and myself.

Incorporated as Chapter 3:

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

Contributor	*Statement of contribution*
Matthew B. Thompson (Candidate)	Wrote the manuscript (80%)
Jason M. Tangen	Wrote the manuscript (17%)
Duncan J. McCarthy	Wrote the manuscript (3%)

This chapter was published in the *Journal of Forensic Sciences* and is very much my own work, with contributions from Jason Tangen on conception and writing making up around 20% of the final publication.

Incorporated as Chapter 4:

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Human matching performance of genuine crime scene latent fingerprints. *Law and Human Behavior*. doi: 10.1037/lhb0000051

Contributor	*Statement of contribution*
Matthew B. Thompson (Candidate)	Designed experiments (50%) Collected data (60%) Analysed data (60%) Wrote the paper (75%)
Jason M. Tangen	Designed experiments (50%) Collected data (40%) Analysed data (40%) Wrote the paper (25%)
Duncan J. McCarthy	Police database search (100%)

This chapter was published in the journal *Law and Human Behavior*, and the majority of the work is my own, with Jason Tangen contributing 50% to the experimental design, 10% to other areas, and 25% to writing of the final publication.

Incorporated as Chapter 8:

Edmond, G., **Thompson, M. B.**, & Tangen, J. M. (2013). A guide to interpreting forensic testimony: Scientific approaches to fingerprint evidence. *Law, Probability & Risk*. doi: 10.1093/lpr/mgt011

Contributor	*Statement of contribution*
Matthew B. Thompson (Candidate)	Wrote the included sections (40%)
Gary Edmond	Wrote the included sections (40%)
Jason M. Tangen	Wrote the included sections (20%)

The published article in the journal *Law, Probability & Risk* included several sections, but I have included in this thesis only from the abstract to section four, which I contributed to heavily. I am entirely responsible for the language in *The Guide* itself and I am very much responsible for the *Insights from Medicine: The Diagnostic Model* section, with Jason Tangen contributing 30% to the section's conception and writing.

0.3 Contributions by Others to the Thesis

Jason Tangen contributed to the conception and design of all studies presented in this thesis, and made comments on the written work.

0.4 Statement of Parts of the Thesis Submitted to Qualify for the Award of Another Degree

None.

Acknowledgements

That mine is the only name on this thesis must be an error. There is no earthly way I could have made it through without friends, family, colleagues, and taxpayers.

To Jason Tangen: You are a gentleman and a scholar. Your wisdom, guidance, and infectious enthusiasm has instilled in me a passion for discovery, teaching, and sharing. You taught me to be methodical, open, thoughtful, and impactful. The best I can do in gratitude is to be as generous toward my own students.

To Penelope Sanderson: This is the seventh year that I've had the pleasure of being your student. You have influenced so many aspects of my life. From day one you have guided me towards becoming a better academic and a better person. When you first met my brother at my first graduation, he said he was surprised not to see you with a shining halo. I still gloat about how lucky I am to have you as a mentor, and I will persevere in striving to emulate you.

To William Thompson and Gary Edmond: You have been my legal guardian angels—steering me well and putting me in my place. Your tolerance and tutelage has given me the knowledge and confidence to push my research outside the lab.

To those who supported my Fulbright Scholarship year at UCLA and UC Irvine—Jennifer Mnookin, William Thompson, Itiel Dror, Simon Cole, Mark Darby, and Elizabeth Loftus—thank you for opening doors and making it an incredible year.

Thank you to my police and professional colleagues—Duncan McCarthy, Bruce Comber, Teneille Evans, Jenny Scott, and Alastair Ross—for helping me understand your world and for being our champions.

Thank you to my lab and office mates for making the days fun and enlightening: Wen Wu, Rachel Searston, Ruben Laukkonen, Elise Jones, Alice Towler, Rene Treloar, Kathleen Ivison, Bridie James, Merryn Constable, Katherine Woodward, Sean Murphy, Jacqueline Seah, Elizabeth Whitehouse, Cindy Arnita Theresiana, Billy Sung, Jane Sexton, Hayley Thomason, Joyce Vromen, Itsik Nadler, Tobias Grundgeiger, Stacey Parker, and Tania Xiao.

Thank you to the institutions that have supported me financially: National ICT Australia, the Australian-American Fulbright Commission, the Queensland Government, the Australian Government and The University of Queensland.

To William Harrison, James Retell, Morgan Tear, and David Liu: You have made these the best years of my life.

To Peter, Helen, Dad, Krystal, and Mike: You were patient beyond belief and I'm lucky to have you.

To Mum: You weren't here physically, but I couldn't have done it without you.

To my dearest Jess: In the Acknowledgements of my first thesis, I wrote, "I hope one day to repay you for the unwavering support you have given me. I love you." That day is yet to come, but I am now well and truly indebted to you, and I appreciate everything you do for me and for us. Although, one thing has changed since then: *I love you more than ever.*

Keywords and Classifications

0.5 Keywords

psychology, cognition, expertise, decision making, expert evidence, forensic science

0.6 Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 170202, Decision Making, 70%

ANZSRC code: 180110, Criminal Law and Procedure, 20%

ANZSRC code: 160205, Police Administration, Procedures and Practice, 10%

0.7 Fields of Research (FoR) Classification

FoR code: 1702, Cognitive Sciences, 70%

FoR code: 1801, Law, 20%

FoR code: 1602, Criminology, 10%

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Preface

I spent the first six months of my PhD candidature designing experiments in which the only participants were undergraduates. My advisor, Jason Tangen, had spent considerable time over past four years working to develop relationships with law enforcement agencies in Canada and Australia, to no avail. Examiners did not seem particularly interested in opening up to academics who wanted understand expert fingerprint matching. We had a breakthrough in 2010 when the Queensland Police Service called us into headquarters to discuss the problem of dealing with uncertainty in fingerprint identifications from casework. After that, we worked closely with police officers who helped us understand the work they do and who helped us create experiment-quality fingerprint stimuli. Given increasing demand for evidence of the abilities of fingerprint examiners I began to focus performance differences between qualified examiners and novices, and what the source of identification errors might be. I wanted to understand the nature of fingerprint expertise from a cognitive psychology perspective—what are the factors that affect matching accuracy, and how do experts do what they do? I was mindful that the results of my research program would be relevant to non-scientists, and so I ensured my work was packaged in a way made it accessible to judges, lawyers, police officers, forensic examiners, and the public.

This thesis is not presented in a traditional format. Figure 1.1 shows the conceptual diagram of this thesis, and moving across the top, from left to right, are the four broad themes: **PART 1** - ESTABLISHING EXPERTISE, **PART 2** - DEPICTING EXPERTISE, **PART 3** - NATURE OF EXPERTISE, and **PART 4** - EXPRESSION OF EXPERTISE. Beneath each of the themes is at least one thesis chapter, and the thesis moves sequentially down through each of the chapters until returning to the top to start the next part. The thesis comprises both published and unpublished works, which is why some redundancy between chapters could not be avoided. Chapters 2, 3, 4, and 8 are quoted directly from published, peer-reviewed journal articles, and I have extracted the text and figures from the final published versions for consistency. Chapters 4, 6, and 7 are currently unpublished, and will be submitted for publication as three separate manuscripts.

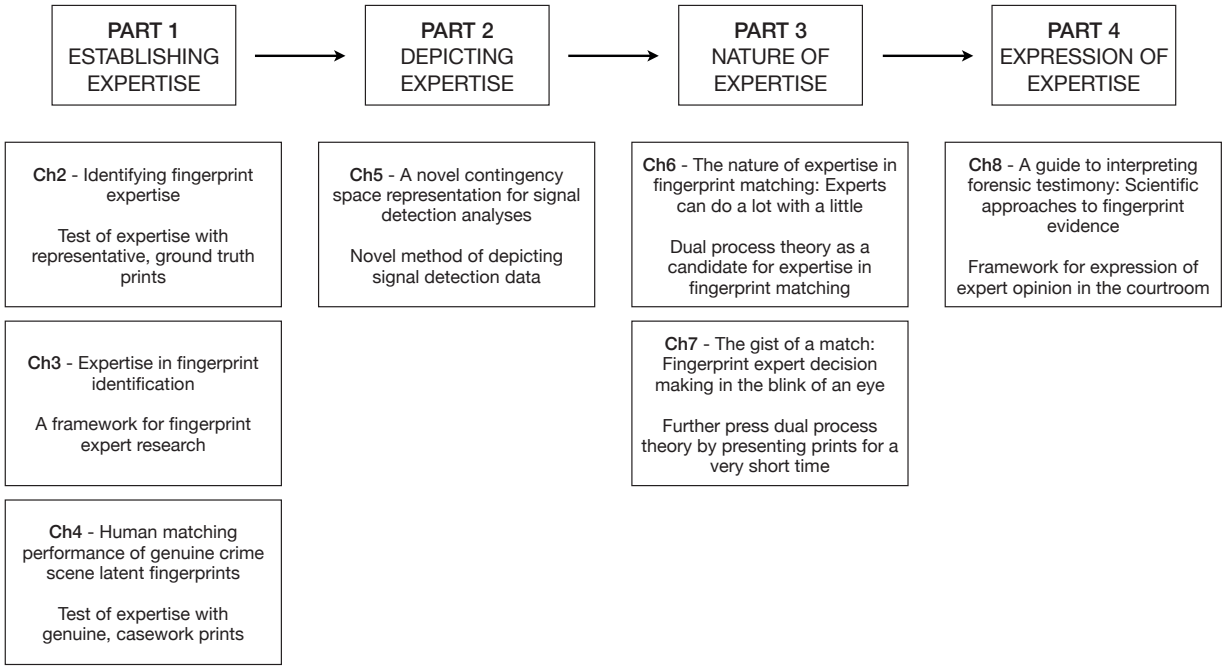

Figure 1.1: Conceptual diagram of the thesis in four parts: (1) Establishing expertise, (2) Depicting expertise, (3) Nature of expertise, and (4) Expression of expertise. Each of the four parts includes at least one thesis chapter.

1.2 PART 1 - ESTABLISHING EXPERTISE

Fingerprint examiners have been active in investigations and have presented identification evidence in criminal courts for more than a century (Cole, 2002). Remarkably, given that testimony about fingerprint matches is a product of human judgment and subjective decision making, there have been few scientific investigations of the human capacity to correctly match fingerprints. Examiners have claimed that fingerprint identification is infallible (Federal Bureau of Investigation, 1984) and that there is a zero error rate for fingerprint comparisons (Cole, 2005; Edwards, 2009b). These claims of individualization and a zero error rate, however, are not supported by evidence and are scientifically implausible (Cole, 2010; National Research Council, 2009; Saks & Faigman, 2008). The fact that humans cannot be detached from forensic decision making has been highlighted in a variety of recent inquiries by the U.S. National Research Council of the National Academy of Sciences (2009), the Scottish Fingerprint Inquiry (Campbell, 2011), and the U.S. National Institute of Justice (2012).

The National Academy of Sciences (NAS, 2009) has highlighted the absence of solid scientific methods and practices in U.S. forensic science laboratories. Harry T. Edwards (a senior U.S. judge and co-chair of the NAS Committee) noted that forensic science disciplines, including fingerprint comparison, are typically not grounded in scientific methodology, and forensic experts do not follow scientifically rigorous procedures for interpretation that ensure that the forensic evidence that is offered in court is valid and reliable (Edwards, 2009b; Risinger, Saks, Thompson, & Rosenthal, 2002; Saks & Koehler, 2005). Most recently, a large multidisciplinary collective—the Expert Working Group on Human Factors in Latent Print Analysis (2012)—was sponsored by the U.S. National Institute of Standards and Technology and the National Institute of Justice to investigate human factors in latent fingerprint identification (2012). The authors recommended that examiners should be familiar with human factors issues such as fatigue, bias, cognitive and perceptual influences, and not claim that errors are inherently impossible or that a method inherently has a zero error rate. They recommended that management foster a culture in which it is understood that some human error is inevitable and that a comprehensive testing program of competency and proficiency should be developed and implemented. Speaking generally, and taking the lead from medical

and aviation research, the authors advocated that fingerprint identification would benefit from the human factors research and systems approaches to improve quality and productivity, and reduce the likelihood and consequences of human error.

Researchers have investigated the effect of contextual bias on fingerprint examiners (Dror & Cole, 2010; Dror & Rosenthal, 2008; Langenburg, Champod, & Wertheim, 2009), the special abilities and vulnerabilities of fingerprint examiners (Busey & Dror, 2010; Busey & Parada, 2010; Busey et al., 2011), the psychophysics of fingerprint identification (Vokey, Tangen, & Cole, 2009), the effect of technology (Dror & Mnookin, 2010; Dror, Wertheim, Fraser-Mackenzie, & Walajtys, 2012), and statistical models of fingerprint identification(Champod & Evett, 2001; Neumann et al., 2007; Neumann, 2012). Despite these contributions to forensic decision making, still very little is known about human fingerprint matching performance, the nature of expertise in fingerprint identification, the factors that affect matching accuracy, and the basis on which examiners can reasonably testify in court. Considering the shift toward viewing the human as an integral part of the forensic identification process (Tangen, 2013), systematic programs of research are needed to understand the skills, abilities, and limits of fingerprint examiners, and to understand the nature of their expertise. Research programs that are underway, or are to be developed, include understanding the nature of forensic expertise, the influence of cognitive and perceptual biases, the impact of technology, how best to present pattern evidence to judges and juries, the best ways to turn novices into experts, and the most effective and efficient work practices, environments, and tools. Before these research programs can advance, however, a foundation for understanding expertise and accuracy in human fingerprint identification is needed.

In PART 1 of this thesis—ESTABLISHING EXPERTISE—I explore expertise in fingerprint identification. I attempt to find evidence for expert-novice differences in fingerprint matching and explore where performance differences might lie. I also describe how, in designing these experiments, I have attempted to balance fidelity, generalizability, and control to answer the most pressing research questions appropriately and efficiently.

In Chapter 2—Identifying Fingerprint Expertise—I set out to determine whether fingerprint experts are any more accurate at matching prints than the average person, and to get an idea of how often experts make errors of the sort that could allow a guilty person to escape

detection compared with how often they make errors of the sort that could falsely incriminate an innocent person. Expert and novice examiners made judgments about representative, ground truth prints that either match, don't match, or don't match but are highly similar. Novices were undergraduates with no previous experience with prints, while experts were qualified, court-practicing fingerprint examiners from Australian national and state law enforcement agencies.

In Chapter 3—Expertise in Fingerprint Identification—I present a framework for fingerprint expertise research and elaborate on the details and implications of the experiment in Chapter 2. I argue that fidelity, generalizability, and control must be balanced to answer important research questions; that validity, proficiency, and the competence of fingerprint examiners are best determined when experiments include highly similar print pairs where the ground truth is known; that a signal detection paradigm can be employed to separate the two ways of being right and the two ways of being wrong, and to distinguish accuracy from response bias; that the most appropriate comparison group to demonstrate expertise should be novices who have no training with fingerprints whatsoever; and that determining error rates with black box studies may be unnecessary at best and ineffective and inefficient at worst. Finally, unless one can demonstrate that a particular qualifier will systematically affect accuracy, the default should be to report accuracy at the broader level.

In Chapter 4—Human Matching Performance of Genuine Crime Scene Latent Fingerprints—I test the matching accuracy of expert, trainee, and novice examiners using pairs of genuine, casework prints that either match, don't match, or don't match but are highly similar. I increased the fidelity of the discrimination task (i.e., the resemblance of the discrimination task to actual casework) by using genuine crime-scene latents (and their matched exemplars) from police training materials, compiled from casework. In doing so, we can understand how performance changes as novices turn into experts.

Taken together, PART 1 of this thesis will help us understand whether, or how, fingerprint matching expertise is similar to other areas of expertise, such as diagnostic medicine, where non-analytic models of cognition account for much of the superior performance of experts (Brooks, 1978; Norman, Young, & Brooks, 2007; Schmidt, Norman, & Boshuizen, 1990).

Providing empirical evidence of the skills and abilities of qualified, court-practicing fingerprint examiners will help provide a basis for legitimate expert testimony in the courtroom.

1.3 PART 2 - DEPICTING EXPERTISE

Signal detection is a method of quantitating a person's (or system's) ability to distinguish signal from noise, and has proved valuable in measuring and understanding human performance. Signal Detection Theory was initially applied to radar operators who were trying to discriminate friendly and enemy aircraft and has since been used to measure all areas of human performance (Green & Swets, 1966). When an examiner compares two fingerprints, there are two ways for her to be right and two ways for her to be wrong. To get a comparison right, she can correctly say the prints match when in fact they do (a hit) or she can correctly say they do not match when in fact they do not (a correct rejection). These decisions could result in correctly incriminating a guilty person, or helping to eliminate potential suspects. To get a comparison wrong, she can incorrectly say that the prints match when in fact they do not (a false alarm), or she can incorrectly say that the prints do not match when in fact they do (a miss). These decisions could result in falsely incriminating an innocent person, or allowing a guilty person to escape detection.

There are several, well-established, ways to describe signal detection data numerically and pictorially. For example, sensitivity indices (e.g., d' and A') and response bias indices (e.g., c and β), and pictorial representations (e.g., detection error tradeoff graphs and confidence-based Receiver Operator Characteristic curves). I suggest that these representations can be difficult to interpret for those not well versed in Signal Detection Theory, and that there may be a complementary representation that better depicts the relationship between sensitivity and response bias.

In PART 2 of this thesis—DEPICTING EXPERTISE—I explore alternative methods for communicating and illustrating sets of signal detection data, and the results of experiments from Part 1 in particular. In Chapter 5—A Novel Contingency Space Representation for Signal Detection Analyses—I describe a method of depicting signal detection data. I describe contingency tables and how to create a contingency space to depict sensitivity, response

bias, chance, and relative performance. I argue that the method makes the theoretical independence between sensitivity and response bias clearer, makes experimental results easier to interpret through the use of natural frequencies, and is especially useful for comparing results between experimental conditions.

1.4 PART 3 - NATURE OF EXPERTISE

Experts are those who consistently performance better than lay people, while *expertise* refers to the mechanisms underlying this superior performance (Ericsson & Charness, 1994). One property of expertise is that experts can perform accurately when given only a small amount of information:

- Chess experts can reconstruct board configurations from memory after only a five second glance whereas novices cannot (Chase & Simon, 1973), and the quality of chess moves selected by experts remains high when the time available is drastically reduced (Ericsson, 1996);

- Fireground Commanders, under extreme time pressure, can rapidly decide an effective course of action for the situation (Klein, 1998);

- Expert radiologists can accurately discriminate normal and abnormal chest X-ray films after a 500 millisecond viewing (Myles-Worsley, Johnston, & Simons, 1988), and chest radiographs after a 2 second viewing (Lesgold et al., 1988) or 200 millisecond viewing (Kundel & Nodine, 1975);

- Expert dermatologists make correct diagnoses of skin disorders after only a few seconds, and are more likely to make an incorrect diagnosis the longer they take to respond (Norman, Rosenthal, Brooks, Allen, & Muzzin, 1989); and

- Expert radiologists and cytologists detect abnormalities above chance in mammogram X-rays and Pap test images after a 200 millisecond viewing (K. K. Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013; Drew, Evans, Vo, Jacobson, & Wolfe,

2013), and they can recall these images from memory better than novices after viewing them for only three seconds (K. K. Evans et al., 2010).

Further examples abound (Klein, 1998), and so it is clear that experts can perform more accurately and consistently than novices when the amount of information is limited. How do experts perform quickly and accurately when there is little time for careful, deliberate analysis?

Expert judgment and decision making is influenced by a combination of intuition and deliberation (Kahneman, 2011). The two-system model (dual-process theory[s] of reasoning) posits two kinds of thinking: System 1 and System 2 (Stanovich & West, 2000). The exact nature of these systems is debated (J. S. B. T. Evans, 2003; Osman, 2004) and having two styles of thinking does not necessarily suggest two distinct cognitive systems (J. S. B. T. Evans, 2012), but dual-process theories have proved to be a useful metaphor for understanding and explaining the abilities of genuine experts. System 1 operates automatically and quickly, and is characterized as intuitive, unconscious, associative, and effortless. System 1 gives no sense of voluntary control and its processing is non-analytic (Brooks, 1978). System 2 operates deliberately and slowly, and is characterized as reasoned, conscious, orderly, and effortful. System 2 gives a sense of voluntary control, choice, and concentration and its processing is analytic.

In PART 3 of this thesis—NATURE OF EXPERTISE—I explore the cognitive processes that might account for the superior performance of expert fingerprint examiners. It is clear that experience with fingerprints provides a real performance benefit, but the basis, or nature, of fingerprint expertise is less clear.

In Chapter 6—The Nature of Expertise in Fingerprint Matching: Experts Can Do a Lot with a Little—I evaluate dual-process theory as a candidate to explain the nature of expertise in fingerprint matching. I expect at least part of the superior performance of fingerprint examiners relative to laypeople to be the result of non-analytic cognition. I present four experiments where I attempt to understand the limits of human fingerprint discrimination and to characterise the influence of non-analytic cognition. I do this by limiting the "information" available to participants, first by adding visual noise to print images and presenting them

either inverted or upright, second by spacing prints in time by a few seconds, third by spacing prints in time by a few minutes, and fourth by limiting the presentation time of prints to a few seconds.

In Chapter 7—The Gist of a Match: Fingerprint Expert Decision Making in the Blink of an Eye—I explore the limits of rapid expert decision making. I further test non-analytic cognition (and dual-process models) as a plausible theory to account for superior expert performance by presenting fingerprints on screen very briefly. Presenting prints for a few hundred milliseconds will give little opportunity for participants to make analytic, rule-based judgments. Experts matching prints more accurately than novices after would suggest that experts are making use of a repertoire of previous instances of matching and nonmatching prints stored in memory in order to judge new instances (Rouder & Ratcliff, 2006, 2004; Brooks, 1978; Norman et al., 2007).

Taken together, PART 3 of this thesis will help us begin to understand the nature of fingerprint matching expertise. I will determine the extent to which experts make use of non-analytic processing (Brooks, 1978) when identifying prints, and whether experts can perform accurately when information is sparse. Understanding the nature and development of fingerprint expertise will create a foundation for evidence-based training and testimony.

1.5 PART 4 - EXPRESSION OF EXPERTISE

For more than a hundred years, and in the absence of experimental support, fingerprint examiners have claimed that fingerprint evidence is basically infallible (Cole, 2010). These assertions are typically justified by reference to training and experience—and the use of a method such as ACE-V: Analysis, Comparison, Evaluation and Verification—and assumptions about the uniqueness of fingerprints, along with legal acceptance and the effectiveness of fingerprint evidence in securing confessions and convictions (B. L. Garrett, 2011; Haber & Haber, 2007; Koehler, 2012; Vokey et al., 2009). In recent decades, however, commentators have questioned these claims of uniqueness (and its significance) and dismissed claims about error-free, positive identification as scientifically implausible.

Notwithstanding long reliance on fingerprint evidence, relatively little is known about the performance of fingerprint examiners or the value of their opinions. Currently, there is a dearth of research. The necessary studies are often beyond the capabilities and competence of fingerprint examiners (and yet to be undertaken, or completed). Understandably, the training of fingerprint examiners is primarily oriented toward comparing fingerprints. Most do not have the methodological skills, funding, time, infrastructure, or experience with research techniques to mount scientific studies of human performance. Moreover, few fingerprint examiners, lawyers, or judges have the time, resources or expertise to track and evaluate extant studies, inquiries and reports, or respond to research as it emerges (Mnookin, Cole, Dror, & Fisher, 2010). Consequently, changes to practices and reporting will require the ongoing assistance of research scientists.

There have been destabilizing epistemic and organizational problems raised in scholarly critiques and the recent authoritative and independent inquiries and reports. It is clear that fingerprint identification cannot be regarded as an infallible 'methodology' that is detached from human judgment (Cole, 2005; Tangen, 2013). Given the long history of claims about uniqueness, individualization, and a disembodied identification process, I argue that examiners and their institutions should now begin to replace traditional practices and reporting with evidence-based claims that reflect actual capabilities. Regardless of what forensic scientists do, criminal courts have a principled obligation to truth and justice (Ho, 2008). Courts, particularly those jurisdictions with a reliability-based admissibility standard, have an obligation to require forensic scientists to present their evidence in ways that embody actual capabilities. This requires evaluating reliability and conveying limitations clearly to the tribunal of fact.

In PART 4 of this thesis—EXPRESSION OF EXPERTISE—I explore ways that fingerprint examiners can communicate (express) their opinions in ways that are responsive to recent epistemological critiques and recent empirical findings. In Chapter 8—A Guide to Interpreting Forensic Testimony: Scientific Approaches to Fingerprint Evidence—I propose a framework for the expression of expert opinion in the courtroom in order to integrate extra-legal recommendations and emerging research and produce a serviceable tool to assist the legal regulation and use of fingerprint evidence. The framework is based on the medical diagnostic

model where the validity, reliability, and accuracy of the test come from the aggregation of many controlled experiments, and which offers information about similar situations in order to help decision-makers reason about the present case. I suggest that an indication of performance (and error) in previous situations, (reasonably) similar to the particular analysis, provides potentially valuable information to those obliged to evaluate fingerprint testimony.

1.6 Purpose

Little is known about the nature and development of fingerprint expertise and, therefore, the best way to turn novices into experts. Little is known about the factors that affect matching accuracy and, therefore, what experts can legitimately testify to in court. With this program of research, I seek to determine the factors that affect matching accuracy, to better understand the development of expert forensic identification, to inform training, and to provide an empirical basis for expert testimony in the courtroom. I will ground the investigation in exemplar, signal detection, and dual-process theories, and draw from several relatively well-understood domains such as expertise and skill acquisition, and familiar and unfamiliar face recognition. I hope to provide general, empirical evidence—evidence that can reasonably be generalized—to form a basis for effective training and legitimate expert testimony.

Chapter 2

Identifying Fingerprint Expertise

2.1 Preface

This chapter is extracted from a published article in the journal *Psychological Science*. As can be seen in Figure 2.1, this chapter is the first in PART 1 - ESTABLISHING EXPERTISE. Contributions to design, running, analysis, writing, etc., were shared equally between Jason Tangen and myself. Reference:

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

After spending time with police examiners in order to understand the task of fingerprint identification, we designed the first test of fingerprint matching expertise. We travelled to forensic conferences and around Australia to several state and federal police agencies. Understandably, given the climate of criticism of forensic science at the time, examiners we initially reluctant to participate. *Nature* had published a news feature, editorial, and opinion piece detailing the lack of science in forensic science, and former *Science* Editor-in-Chief, Donald Kennedy, weighed in on the topic in his editorial, "Forensic science: Oxymoron?" and highlighted the absence of data on error rates in fingerprint identification. Examiners became more receptive after we explained that our approach was to understand expertise,

not to test them directly. This was the most difficult matching task we could conceive at the time; we did not expect examiners to do as well as they did. We published the results in *Psychological Science* because we thought it represented the beginning of an interesting problem for cognitive science to sink its teeth into.

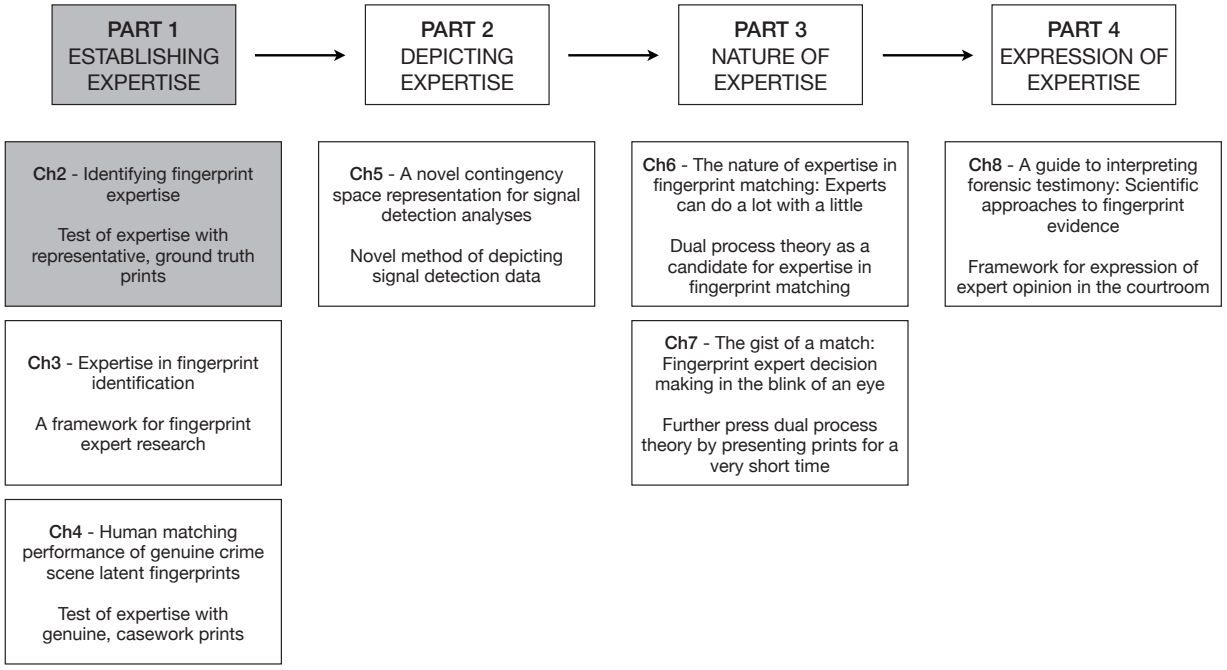

Figure 2.1: Conceptual diagram highlighting Chapter 2, Part 1 of the thesis: "Identifying fingerprint expertise."

2.2 Introduction

"CSI-style" TV shows give the impression that fingerprint identification is fully automated. In reality, when a fingerprint is found at a crime scene, it is a human examiner who is faced with the task of identifying the person who left the print—a task that falls squarely in the domain of psychology. The difficulty is that no properly controlled experiments have been conducted on fingerprint examiners" accuracy in identifying perpetrators (Loftus & Cole, 2004), even though fingerprints have been used in criminal courts for more than 100 years. Examiners have even claimed to be infallible (Federal Bureau of Investigation, 1984). However, the

U.S. National Academy of Sciences has recently condemned these claims as scientifically implausible, reporting that faulty analyses may be contributing to wrongful convictions of innocent people (National Research Council, 2009). Proficiency tests of fingerprint examiners and previous studies of examiners' performance have not adequately addressed the issue of accuracy, and they been heavily criticized for (among other things) failing to include large, counterbalanced samples of targets and distractors for which the ground truth is known (Cole, Welling, Dioso-Villa, & Carpenter, 2008; Vokey et al., 2009). Thus, it is not clear what these tests say about the proficiency of fingerprint examiners, if they say anything at all. Researchers at the National Academy of Sciences and elsewhere (e.g., see Saks & Koehler, 2005; Spinney, 2010b) have argued that there is an urgent need to develop objective measures of accuracy in fingerprint identification. Here we present such data.

2.3 Method

2.3.1 Participants

Thirty-seven qualified practicing fingerprint experts from five police organizations (the Australian Federal, New South Wales, Queensland, South Australia, and Victoria Police) participated in the study. In addition, 37 undergraduates from The University of Queensland participated for course credit, providing comparison data on the performance of novices.

2.3.2 Procedure

We presented the 37 qualified fingerprint experts and the 37 novices with pairs of prints displayed side by side on a computer screen, as illustrated in Figure 2.2. Participants were asked to judge whether the prints in each pair matched, using a confidence rating scale ranging from 1 (sure different) to 12 (sure same); judgments were reported by moving a scroll bar to the left ("different") or right ("same"). Note that the scale forced a "match" or "no match" decision because ratings of 1 through 6 indicated a match, whereas ratings of 7 through 12 indicated no match. Judgments that the information was "inconclusive," which are often made in practice, were not permitted in this two-alternative forced-choice design,

so it was possible to distinguish between accuracy and response bias (Swets, 1992). This task emulates the most forensically relevant aspect of the identification process, namely, the extent to which a print can be accurately matched to its source.

2.3.3 Stimuli

The stimuli consisted of 36 simulated crime-scene prints that were paired with fully rolled prints. Across participants, each simulated print was paired with a fully rolled print from the same individual (match), with a nonmatching but similar exemplar (similar distractor), and with a random nonmatching exemplar (nonsimilar distractor). For each participant, each simulated print was randomly allocated to one of the three trial types, with the constraint that there were 12 prints in each condition.

The simulated prints and their matches were from the Forensic Informatics Biometric Repository, so, unlike genuine crime-scene prints, they had a known true origin (Cole, 2005). Simulated prints were dusted by a research assistant (who was trained by a qualified fingerprint expert), photographed, cropped to 600 $\times$ 600 pixels, and isolated in the frame. A qualified expert (the third author) reported that each simulated print contained sufficient information to make an identification if there was a clear comparison exemplar. The matching exemplars were fully rolled fingerprint impressions made using a standard elimination pad and a 10-print card. Each card was scanned in color as a 600-dpi lossless Tagged Information File Format (TIFF) file, and each print was cropped to 600 $\times$ 600 pixels and isolated in the frame.

Similar distractors were obtained by searching the Australian National Automated Fingerprint Identification System. For each simulated print, the most highly ranked nonmatching exemplar from the search was used if it was available in the Queensland Police 10-print hard-copy archives, which contains approximately 3.3 million prints. The corresponding 10-print card was retrieved from the archives, scanned, and extracted by the same method as before. In practice, highly similar nonmatches retrieved from large national databases are likely to increase the chance of incorrect identifications (Dror & Mnookin, 2010). Distinguishing such highly similar, but nonmatching, prints from genuine matches is potentially the most difficult task that fingerprint examiners face. The nonsimilar distractor for a given simulated print

was randomly selected from the entire set of matching and similar distractors after removing the match and similar distractor for that simulated print.

2.4 Results

For each participant, we calculated the percentage of trials responded to correctly in each condition. The three graphs on the right side of Figure 2.2 depict the average percentage of correct responses for the 37 experts and 37 novices. As the figure shows, experts performed exceedingly well. On the 12 trials in which the prints matched, experts correctly identified 92.12% of the pairs, on average, as matches (hits), misidentifying 7.88% as nonmatches (misses). Misses are the kind of error that can lead to a failure to identify a criminal. On the 12 similar-distractor trials, experts correctly declared nearly all of the pairs (99.32%) to be nonmatches (correct rejections); only 3 pairs (0.68%) out of the 444 in this condition were incorrectly declared to be matches (false alarms). Experts did not misidentify any of the 12 nonsimilar distractor prints as matches. Such errors can lead to false convictions. Even though the novices could reliably distinguish matching and nonmatching prints, they made a large number of errors. In particular, novice participants mistakenly identified 55.18% of the similar, nonmatching distractor prints as matches (the corresponding rate for experts was 0.68%). We subjected the percentages of correct responses to a 2 (expertise: experts, novices) $\times$ 3 (trial type: match, similar distractor, nonsimilar distractor) mixed analysis of variance. The analysis revealed significant main effects of expertise, $F(1, 72) = 416.46$, $MSE = 0.013$, $p < .001$, and trial type, $F(2, 144) = 45.68$, $MSE = 0.011$, $p < .001$, and a significant interaction between the two, $F(2, 144) = 64.32$, $MSE = 0.011$, $p < .001$. Simple-effects analyses revealed a significant benefit of expertise on all trial types—match: $F(1, 72) = 38.49$, $MSE = 0.01$; similar distractor: $F(1, 72) = 476.99$, $MSE = 0.01$; and nonsimilar distractor, $F(1, 72) = 98.46$, $MSE = 0.01$.

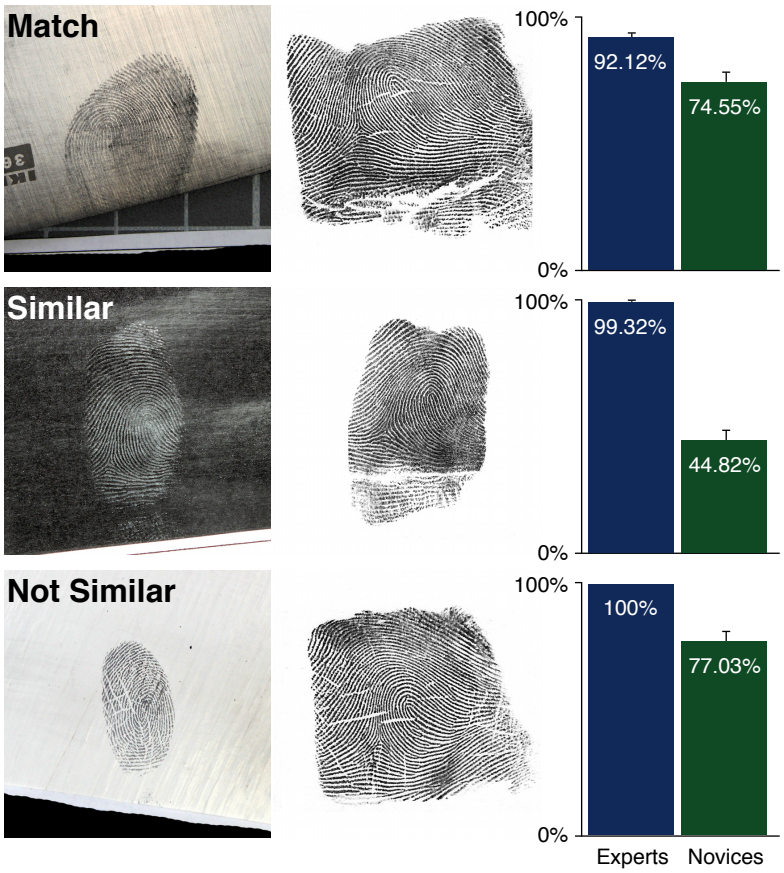

Figure 2.2: Stimuli and results. On each trial, participants were presented with a simulated crime-scene print on the left and a fully rolled candidate print on the right, and they were asked to indicate their level of confidence in whether the prints matched. On some trials, the two prints came from the same individual (top row); on others, the prints were similar but came from two different individuals (middle row); and on others, the prints came from two different individuals and were paired randomly (bottom row).The three graphs on the right depict experts' and novices' mean percentage of correct responses in these three conditions. Error bars represent 95% within-cell confidence intervals.

2.5 Conclusions

We have shown that qualified, court-practicing fingerprint experts are exceedingly accurate compared with novices, but are not infallible. Our experts tended to err on the side of caution by making errors that would free the guilty rather than convict the innocent. Even so, they occasionally made the kind of error that can lead to false convictions. Expertise with fingerprints appears to provide a real performance benefit, but fingerprint experts—like doctors and pilots—make mistakes that can put lives and livelihoods at risk.

Qualified fingerprint examiners now have evidence to legitimately claim specialized knowledge, which may satisfy legal admissibility criteria. It remains unclear, however, how our experiment should affect the testimony of forensic examiners and the assertions that they can reasonably make. The issue is no longer whether fingerprint examiners make errors, but rather how to acknowledge those errors.

We have taken a first step in addressing the call by the National Academy of Sciences for cognitive psychology to establish the limits and levels of performance in forensic science. Considering the central role of humans in forensic identification, the field would benefit from further psychological research. Research on clinical reasoning in medicine, for example, developed over the past 40 years, after it became evident that physicians' decisions too often resulted in adverse consequences for patients. Much has been learned about differences between novice and expert medical practitioners, the influence of cognitive biases in medical decision making, and the most effective ways to incorporate such knowledge into practice. Further research into forensic decision making will help to ensure the integrity of forensics as an investigative tool so that the rule of law is justly applied.

Chapter 3

Expertise in Fingerprint Identification

3.1 Preface

This chapter is extracted from a published article in the journal *Journal of Forensic Sciences*. As can be seen in Figure 3.1, this is the second chapter of PART 1 - ESTABLISHING EXPERTISE. This chapter is very much my own work, with contributions from Jason Tangen on conception and writing making up around 20% of the final publication. Reference:

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Expertise in Fingerprint Identification. *Journal of Forensic Sciences, 58*(6), 1519–30. doi: 10.1111/1556-4029.12203

When I presented the results from Chapter 2 at police departments and forensic conferences, it became clear that I had taken the reasoning behind our approach to measuring fingerprint expert performance for granted. It was difficult to communicate to non-scientists that no experiment can mirror situations in the wild without losing the very control that makes it an experiment. Also, many people misunderstood the goal of the experiment and so were misinterpreting the meaning of the rates of error. I decided to write a "commentary" to explain our approach to the research problem, to put the results of the experiment in context, and to consider the implications for current practice. We published in the top forensic journal,

rather than a psychology journal, to make it more likely that fingerprint examiners and forensic managers would have access to it and read it.

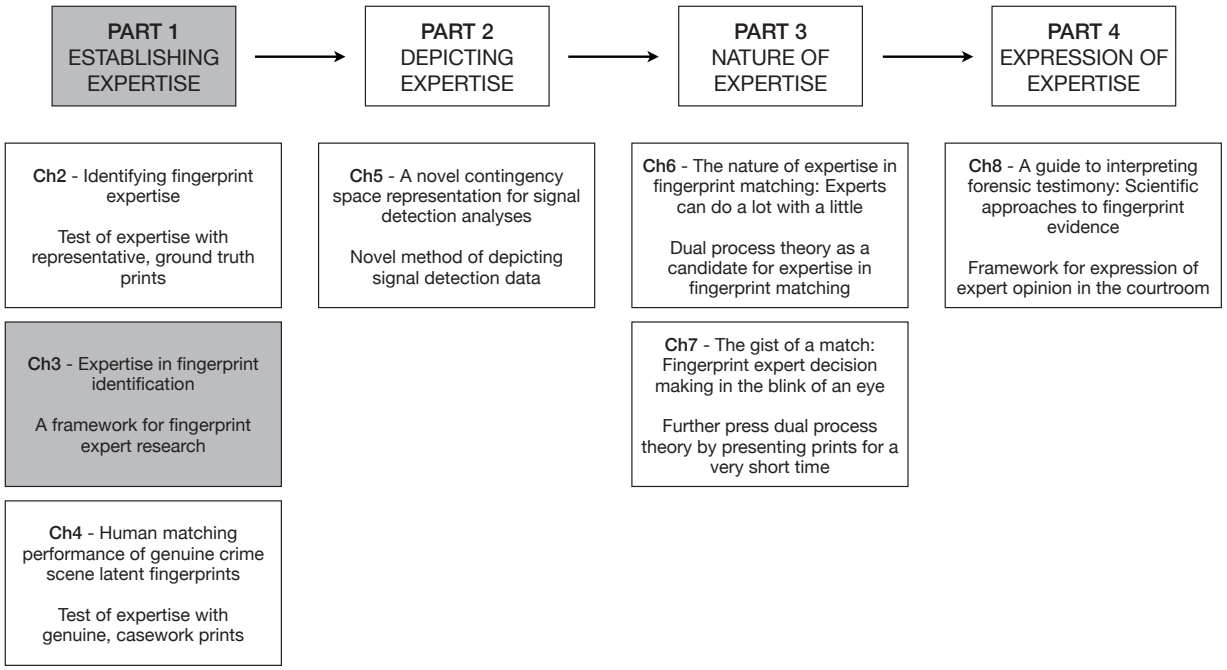

Figure 3.1: Conceptual diagram highlighting Chapter 3, Part 1 of the thesis: "Expertise in fingerprint identification."

3.2 Abstract

Although fingerprint experts have presented evidence in criminal courts for more than a century, there have been few scientific investigations of the human capacity to discriminate these patterns. A recent latent print matching experiment shows that qualified, court-practicing fingerprint experts are exceedingly accurate (and more conservative) compared with novices, but they do make errors. Here, a rationale for the design of this experiment is provided. We argue that fidelity, generalizability and control must be balanced in order to answer important research questions; that the proficiency and competence of fingerprint examiners is best determined when experiments include highly similar print pairs, in a signal detection paradigm, where the ground truth is known; and that inferring from this experiment

the statement "The error rate of fingerprint identification is 0.68%" would be disingenuous. In closing, the ramifications of these findings for the future psychological study of forensic expertise, and the implications for expert testimony and public policy are considered.

3.3 Introduction

Maintaining high standards of forensic evidence is vital for an effective justice system and for ensuring that innocent people are not wrongfully accused. Although fingerprint experts have presented evidence in criminal courts for more than a century, there have been few scientific investigations of the *human* capacity to discriminate these patterns and impressions. Contrary to popular belief (and television shows like CSI), computers are not relied upon to match crime scene fingerprints. Instead, human fingerprint experts decide whether a print belongs to a suspect or not. These experts make thousands of fingerprint identifications, per day, to be used as evidence in courts of law. Until recently, it was unclear what role expertise plays or whether expertise is even necessary to conduct accurate fingerprint comparisons.

In 2011, we (Tangen, Thompson and McCarthy) published the results of an experiment testing the accuracy and claimed expertise of fingerprint examiners. These results showed that qualified, court-practicing fingerprint experts are exceedingly accurate (and more conservative) compared with novices, but they do make errors. Here, the current state of fingerprint testimony, measures of accuracy, and the research culture in forensic science are discussed. A rationale for the "Identifying Fingerprint Expertise" (2011) experimental design is provided, and the steps taken to balance fidelity, generalizability and control; ensure validity and ground truth; create a signal detection framework with highly similar prints; establish expertise with a novice control group; and establish meaningful error rates are described. Given the brevity of the original research article, this rationale will provide context for interpreting the results for the benefit of researchers, forensic examiners, forensic managers, lawyers, and judges. In closing, the ramifications of the findings for the future study of forensic expertise, and the implications for expert testimony and public policy are considered.

3.4 Fingerprint Expert Testimony

In 2009, the National Academy of Sciences (NAS) delivered a landmark report highlighting the absence of solid scientific methods and practices in the forensic science domain (National Research Council, 2009). Harry T. Edwards (a senior US judge and co-chair of the NAS Committee) noted that forensic science disciplines, including fingerprints, are typically not grounded in scientific methodology, and forensic experts are not bound by solid practices that ensure that the forensic evidence that is offered in court is valid and reliable (Edwards, 2009b; Campbell, 2011; Loftus & Cole, 2004; Saks & Koehler, 2005). The NAS report highlighted the absence of experiments on human expertise in forensic pattern matching: "The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine its validity. This is a serious problem." They recommended that the US Congress fund basic research to help the forensic community strengthen their field, develop valid and reliable measures of performance, and establish evidence-based standards for analyzing and reporting testimony.

Courts rely heavily on forensic evidence to convict the guilty and protect the innocent. The presentation of flawed forensic evidence has obvious implications for individual cases, but it also calls into question the integrity of the entire criminal justice system – innocent people may be wrongfully convicted and people may lose trust in the justice system (Edwards, 2009b). It is important, therefore, that the claims made by fingerprint examiners testifying in court are accurate, substantiated and reasonable. Fingerprint examiners have claimed that fingerprint identification is infallible (Spinney, 2010b) and that there is a zero error rate for fingerprint comparisons (Edwards, 2009b; Federal Bureau of Investigation, 1984). Several commentators, however, have suggested that the claims of individualization and a zero error rate are not supported by evidence and, moreover, are scientifically implausible (e.g., National Research Council, 2009; Cole, 2005). Past President of the International Association for Identification suggested that members not assert 100% infallibility (zero error rate) of fingerprint comparisons (R. Garrett, 2009) and the Scientific Working Group on Friction Ridge Analysis, Study and Technology (2011a) have drafted a standard for defining, calculating, and reporting error rates.

These issues are made more complicated by the nature of the legal system. The adversarial approach to the submission of evidence in court is not well suited to establishing 'scientific truth'. According to Edwards (2009b), judges and lawyers generally lack the expertise necessary to evaluate forensic evidence scientifically; defense attorneys often lack the resources to challenge the evidence; judges make admissibility decisions without the benefit of judicial colleagues and time to research; and cases are seldom appealed on the basis of disputed forensic evidence. It may be unwise, therefore, to rely on the judicial system to address the challenges facing fingerprint expert testimony.

3.5 Proficiency Tests and Accuracy

Fingerprint proficiency tests are available—such as those provided by Collaborative Testing Services, Inc.—where the goal is to measure the accuracy of participating laboratories as a unit. The results are sometimes reported as a function of the kind of source print, sometimes as a function of the kind of correct response, sometimes as a function of the kind of error, and sometimes as the proportion of examiners/labs producing various responses. Proficiency tests of fingerprint examiners and previous studies of examiners' performance have been heavily criticized for (among other things) failing to include large, counterbalanced samples of targets and distractors for which the ground truth is known (see Scientific Working Group on Friction Ridge Analysis Study and Technology, 2011a; Cole et al., 2008; Haber & Haber, 2007). A weakness of proficiency tests and previous experiments is that a particular crime scene (or "latent") print either forms part of a match comparison or part of a distractor comparison for every individual who takes the test—a particular latent never serves as part of a target trial for one examiner/lab and a distractor trial for another examiner/lab. The result is that even a single highly distinctive latent on a distractor trial can artificially improve discrimination by reducing false positives. Or a single highly distinctive latent in a match trial can artificially improve discrimination by increasing hits (Haber & Haber, 2007).

There is nothing inherently wrong with the proficiency tests, like those provided by Collaborative Testing Services, Inc. (CTS), if the goal is to measure examiners' performance on exactly the same set items. It may be possible to narrow in on particular features that

cause difficulty (e.g., a peculiar pattern type) or what prints a particular department has trouble with. Indeed, if CTS made their materials and all their results widely available, they may provide a useful tool to measure performance on specific items and for assessing reliability. But the tests are insufficient for measuring accuracy. In order to make general claims—beyond those of specific prints at a specific level (i.e., accuracy with whorl patterns)—different (and randomized) sets of prints for each examiner are needed. Otherwise, information in the specific prints used for the test will influence performance, making it difficult to generalize the results (Haber & Haber, 2007). Proficiency tests have not adequately addressed the general issue of expert matching accuracy and are not designed to disentangle the factors that affect matching accuracy.

There is, however, a growing body of research on fingerprint matching. Researchers have investigated the effect of contextual bias on fingerprint examiners (e.g., Vokey et al., 2009; Dror & Charlton, 2006; Dror, Charlton, & Péron, 2006; Dror & Cole, 2010; Dror, Peron, & Hind, 2005; Dror & Rosenthal, 2008); some of the special abilities and vulnerabilities of fingerprint examiners (Langenburg et al., 2009; Busey & Dror, 2010; Busey & Parada, 2010; Busey & Vanderkolk, 2005); the effect of technology (e.g., Busey et al., 2011; Dror & Mnookin, 2010); statistical models of fingerprint identification (Dror et al., 2012; Champod & Evett, 2001; Neumann et al., 2006, 2007; Neumann, 2012); and the accuracy of fingerprint examiners' decisions (e.g., Ulery, Hicklin, Buscaglia, & Roberts, 2011, 2012; Dror et al., 2011; Langenberg, 2009; Wertheim, Langenburg, & Moenssens, 2006b; Haber & Haber, 2007, 2006; Wertheim, Langenburg, & Moenssens, 2006a). But, despite its 100 year history, there have still been few peer-reviewed studies directly examining the extent to which experts can correctly match fingerprints to one another, how competent and proficient fingerprint experts are, how and on what basis examiners make their decisions, or the factors that affect matching accuracy and what is the effect of expertise. In this paper, we focus our efforts on the claimed matching expertise of fingerprint examiners.

3.6 The Research Culture

There is little doubt, among critics and proponents alike, that fingerprint identification is a valuable tool for law enforcement. Fingerprint identification errors are unlikely to be made because of malicious actions—fingerprint experts do their best to provide accurate fingerprint evidence to the courts and uphold civil liberties. Indeed, in the wake of the Mayfield case of false identification, the FBI stated their intention to make certain they are employing the most effective means to ensure the integrity of their expert fingerprint examinations (Spinney, 2010b). But—unlike other areas of expertise where decisions are safety critical, such as aviation and medicine—there is currently no culture of research in fingerprint identification (Mnookin et al., 2010). Steady advances in the fingerprint development process have been made, but the critical human decision making element has been neglected. Examiners are eager to demonstrate their abilities and advance their field, but rarely receive the support and resources to do so. The Director of the FBI's Investigation Lab describes the gap between basic research and its application in solving crimes as the "valley of death" because "nobody wants to pay for it, nobody really wants to do it," (Spinney, 2010a).

It appears that fingerprint examiners are expected to strengthen the scientific basis of their field while they relentlessly make identifications, search databases, and testify in court. Examiners, however, do not have the time, infrastructure, training, expertise or research culture necessary to mount studies of human performance in order to ensure their field meets scientific and legal standards of evidence. By analogy, it would be like expecting the local doctor to find a cure for cancer. Clearly, examiners are not well positioned to address the challenges leveled at their field alone and, traditionally, there has not been a good working relationship between examiners and researchers. Research on expertise and complex systems is the domain of Cognitive Science and of Human Factors. These researchers have the reward structures already in place for conducting and publishing high quality research, and are well positioned to work with examiners to strengthen the field.

Much of the existing research on the cognitive factors involved in fingerprint judgment has investigated the influence of contextual information on examiners' performance. Dror and colleagues (Dror & Charlton, 2006) used a highly-publicized case of exposed fingerprint

error to determine whether biasing information could lead an examiner to change their prior judgment. They covertly evaluated five examiners, with an average of 17 years of experience, who consented to being tested at an unknown time over twelve months. The five examiners were each given a print to identify by a colleague, who advised them that the fingerprints were from a famous case of misidentification by the FBI for the 2004 Madrid train bombings. One examiner reported that the prints matched, three reported that the prints did not match, and one reported inconclusive. Unbeknownst to the examiners, however, the prints that they were asked to identify were taken from their own previous case history where they had previously declared them a match. With four of the five examiners subsequently changing their previous judgment of the prints as matching, it is clear that there is enough ambiguity in fingerprint patterns to reverse a decision from "match" to "non-match" and that top-down, contextual influences can affect their judgments (see also Vokey et al., 2009; Dror & Cole, 2010). Dror and Rosenthal (2008) also conducted a meta-analysis to determine the degree to which examiners would make the same or conflicting decisions if extraneous information about the case was added. Although good data were sparse, the authors concluded that examiners are indeed susceptible to bias.

It is clear that fingerprint experts have special abilities, but their decisions can be influenced by extraneous contextual information (Busey & Dror, 2010; Dror, 2011) and researchers have suggested ways contextual bias can be mitigated (Dror, 2012). Even with this contribution, relatively little research on human fingerprint identification has been conducted by academics and professionals alike. The US National Academy of Sciences (National Research Council, 2009) and others have called for the development of a research culture within forensic science. Mnookin et al. (2010) argue that there is a legitimate role for experience-based claims of knowledge, but also that pattern identification disciplines must develop a scientific foundation, through research, that is grounded in the values of empiricism and skepticism. The experiment described below is a step towards addressing the call from the National Academy of Sciences for the urgent development of objective measures of accuracy and expertise in fingerprint identification.

3.7 The "Identifying Fingerprint Expertise" Experiment

The *Identifying Fingerprint Expertise* experiment (Tangen et al., 2011) was designed to find out whether fingerprint experts were any more accurate at matching prints than the average person, and get an idea of how often they make errors of the sort that could lead to a failure to identify a criminal compared to how often they make errors of the sort that could lead inaccurate evidence being presented in court. Thirty-seven qualified fingerprint experts and 37 undergraduate students were given pairs of fingerprints to examine and decide whether a simulated crime scene print matched a potential "suspect" or not. Some of the print pairs matched, while others were highly similar but did not match.

Thirty-six simulated crime scene prints were paired with fully rolled exemplar prints. Across participants, each simulated print was paired with a fully rolled print from the same individual (match), with a nonmatching but similar exemplar (similar distractor), and with a random nonmatching exemplar (nonsimilar distractor). The simulated prints and their matches were from our Forensic Informatics Biometric Repository, so, unlike genuine crime scene prints, they had a known true origin (Koehler, 2008, 2012). Similar distractors were obtained by searching the Australian National Automated Fingerprint Identification System. For each simulated print, the most highly ranked nonmatching exemplar from the search was used if it was available in the Queensland Police 10-print archives, which contains approximately 3.3 million prints.

The results were striking. Of the prints that actually matched, the experts correctly declared 92.12% of them as matching (hits). Of the prints that did not actually match, the experts incorrectly declared 0.68% of them as matching (false alarms)—impressive expert performance, considering the corresponding false alarm rate for novices was 55.18%. We concluded that qualified court-practicing fingerprint experts are exceedingly accurate compared to novices, but are not infallible. Experts tended to err on the side of caution by making errors that would fail to identify a criminal rather than provide incorrect evidence to the court. Even so, they made the kind of error that could result in incorrect evidence being presented to the court in a criminal trial.

3.8 Balancing Fidelity, Generalizability and Control

Readers will react differently to the methodology employed in the *Identifying Fingerprint Expertise* experiment. The intuition that the best experiment should resemble 'real-life' as much as possible is understandable. However, this intuition is incorrect (see Mook, 1983, for a defense of external *in*validity), as has been demonstrated in other complex, safety-critical domains, such as aviation and medicine. When designing studies of human performance, the challenge is to find the appropriate balance between fidelity, generalizability and control, to produce data that are best suited to answering the research question (Brinberg & McGrath, 1985; Woods, 1985).

Fidelity is the similarity of an experimental task to its reference domain (i.e., fingerprint identification (Brunswik, 1956; Rasmussen, Pejtersen, & Goodstein, 1994). How well does the task represent the particular work domain? Is the expertise of the participants high or low? Are the experimental situations full-featured or simplified? Are the available tools restricted or complete?

Generalizability is the theoretical depth or breadth of applicability of the results to situations beyond those examined in the study. Can the results and conclusions of the experiment be extended to situations that are different from the experiment, and is this the goal?

Control is the latitude available to the experimenters to isolate and manipulate variables. Is control high or low, and will the data collected be sensitive enough to detect differences between the experimental manipulations?

The perfect, but unattainable, experiment, will have high fidelity, high control and high generalizability. But these variables must be balanced in order to answer the research question appropriately (Sanderson, 2008; Sanderson, Liu, Jenkins, Watson, & Russell, 2010). To understand accuracy rates in fingerprint identification it is tempting to think that the best option is to insert test prints—unbeknownst to the examiners—into their regular workflow and measure the number of errors that come out the other end. This arrangement of high fidelity comes at the cost of reduced generalizability (we cannot apply the results from one experiment in a particular lab to fingerprint identification more broadly) and reduced control

(when errors occur, we have no way of knowing how they arose and, therefore, what we might do to prevent them).

Our goal in measuring fingerprint expertise was to compare expert and novice performance at identifying fingerprints, and get an idea of how often they fail to declare matching prints as such (misses) in a matching task compared to how often they declare that prints match when they actually do not (false alarms). In designing the experiment, fidelity, generalizability, and control were balanced in order to answer these questions. The goal was not to generalize the results from these lab-based experiments to the 'real world' (Mook, 1983). The fingerprint examiners who participated in this experiment did not have their usual tools available to allow them to zoom, rotate, or apply filters to the images; the latent prints that were used were collected as part of the Forensic Informatics Biometric Repository; and examiners conducted the experiment during their break on laptops that were provided in bureau conference rooms. This situation is not—purposefully—analogous to casework. The majority of expert participants, nonetheless, reported that the task represented their day-to-day work.

Unlike opinion polls and surveys, scientific experiments are not about seeing how well a sample approximates the general population from which it was selected. Lab-based experiments are intentionally artificial, because they allow for the control of all factors that are not of interest (e.g., the benefit of software tools, the role of verification, the type of crime, inconclusive responses, lifting agents, etc.), and for the systematic manipulation of only the factors of interest (e.g., the difference between novices and qualified experts, comparing performance on match trials and nonmatch trials, ensuring the ground truth of the prints, using highly similar distractors from a national database search, etc.). Generalizability in this context refers to the extent to which the difference between expert and novice performance is 'real', not the extent to which the laboratory setting resembles the everyday operations of a fingerprint bureau.

The unit of analysis in this particular experiment is the comparison between experts and novices or between matching and nonmatching prints, not their absolute performance. So even though a false alarm rate for experts of 0.68% is impressive in its own right, we cannot determine from this experiment whether this rate reflects the false alarm rate of the field more generally. We can conclude, however, that a false alarm rate of 55.18% for

novices pales in comparison to experts under the same conditions. These results provide sufficient evidence for examiners to legitimately claim specialized knowledge, which may satisfy legal admissibility criteria. These results do not allow one to conclude that 0.68% is the misidentification rate for the field. A full scale "black box" experiment would allow us to pinpoint the precise rate of accuracy in the current system, but it is inappropriate and inefficient to conduct a black box experiment to make simple claims like, "Are experts better than novices?" and "Do experts make errors?" Much more will be said about black box experiments in the *Determining Error Rates* section below. Put simply, the design of an experiment needs to be targeted specifically at the question that one sets out to address.

3.9 Validity and Ground Truth

Validity is a cornerstone of the scientific method. It is a measure of whether a method, instrument, questionnaire, construct, etc., measures what it is supposed to measure. Validity can be demonstrated by comparing the outcomes of a method with the ground truth. So, in order to demonstrate the validity of human fingerprint identification, the conclusion of the identification process (i.e., match or nonmatch) should be compared to that which is known (i.e., the ground truth). For example, if the ground truth of a pair of prints is that they were left by two different individuals, but the examiner incorrectly declares that the prints match, then the examiner has made a "false alarm" type of error; if the ground truth of a pair of prints is that they were left by the same finger from the same individual, but the examiner concludes that the prints don't match, then the examiner had made a "miss" type of error. The same goes for the two ways the examiner can reach the correct conclusion: that the prints actually match and the examiner correctly declares them as such (a "hit") or that the prints don't actually match and the examiner correctly declares them a nonmatch (a "correct rejection").

Most tests of proficiency and studies of accuracy (with the exception of Ulery et al., 2011) used print pairs from casework where the ground truth was uncertain (see Cole et al., 2008; Haber & Haber, 2007; Vokey et al., 2009; Haber & Haber, 2004). Only when experiments

make use of print pairs where the ground truth is known, can the validity of fingerprint identification can be demonstrated.

To make use of ground truth prints in our expertise study, fingerprint pairs were sourced from the Forensic Informatics Biometric Repository—an open biometric repository that we created to increase the availability of high quality forensic materials where the ground truth is known. Details on the Forensic Informatics Biometric Repository are available at `www.FIB-R.com`. FIB-R contains a range of crime related materials: fingerprints, palm-prints, shoe-prints, face photographs, handwriting samples, voice samples, and iris photographs. The materials are collected from participants in using a standardized methodology and vary systematically in quality. The repository contains multiple types of materials converging on a single source, and the ground truth of the source is built into the system. Materials are also collected from participants over two sessions to approximate the natural variation that is commonly found in forensic evidence (e.g., changes in facial hair, clothes, and shoe decay). Participants are first-year undergraduates who participate in one hour of data collection for course credit and who provide informed consent.

The fingerprint materials contained in FIB-R are 10-prints, palm-prints and latent prints. Ink is used to capture each fingerprint onto standard 10-print cards, rolled fully from nail-edge to nail-edge, as well as 'slap impressions' (pressing, not rolling, the fingers on the card) and fully rolled palms. Latent prints are taken from common crime scene surfaces (determined in consultation with fingerprint examiners) including: gloss-painted timber, smooth metal, glass and smooth plastic. Participants are instructed to interact with the surfaces by "pushing on the gloss-painted timber to open the door" or "safely grabbing the knife by the blade." By interacting with objects in this way, the aim is to approximate the variation in materials that are commonly found at actual crime scenes.

In our experiment, the latent prints used were mated with their matching fully rolled exemplar so that the ground truth of match trials was known. The use of ground truth print pairs means that we can compare participants' responses to reality.

3.10 Signal Detection

A signal detection framework was used to measure the matching performance of fingerprint examiners (see also Phillips, Saks, & Peterson, 2001; Vokey et al., 2009). Signal Detection is a method of quantifying a person's ability to distinguish signal from noise. In fingerprint identification, for example, *signal* refers to print pairs that truly match and *noise* refers to print pairs that do not truly match. Signal detection was initially applied to radar operators who were trying to discriminate friendly and enemy aircraft, and has since been used to measure all areas of human performance. Several factors may affect a person's ability to distinguish signal from noise, such as experience, expectations, context, physiological and psychological states. In order to conduct a signal detection analysis of novice and expert fingerprint matching performance the two ways of being right and the two ways of being wrong were separated; performance on matching and nonmatching prints was compared; and accuracy and response bias were separated.

3.10.1 Two Ways of Being Right and Two Ways of Being Wrong

When an examiner compares two fingerprints, there are two ways for her to be right and two ways to be wrong, as shown in Fig. 1. To get a comparison right, she can correctly say the prints match when they actually do (a hit) or she can correctly say they don't match when they don't (a correct rejection). These decisions could result in providing correct evidence to the court or help eliminate potential suspects. To get a comparison wrong she can incorrectly say that the prints match when they don't (a false alarm), or she can incorrectly say that the prints don't match when in fact they do (a miss). These decisions could lead to providing incorrect evidence to the court or a failure to identify a criminal.

3.10.2 Compare Performance on Matching and Nonmatching Prints

In order to properly measure performance, examiners must compare both matching and nonmatching prints. As discussed in the section on *Proficiency Tests and Accuracy* above, most previous studies have included no or few distractors, making it impossible to measure

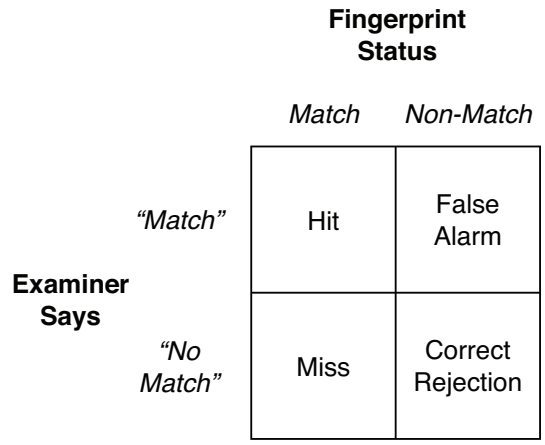

Figure 3.2: A 2×2 contingency table depicting the four possible outcomes of a forced choice fingerprint matching task where two prints match or not and an examiner declares them as a "match" or "no match."

the two ways of being right and the two ways of being wrong, leading to artificially inflated accuracy rates.

In this experiment, to avoid the problem of distractors, each latent print in the set formed part of a match, similar distractor, and non-similar distractor trial. (The reasoning for providing similar distractors is described in the *Similarity* section below.) This way, match trials can be directly compared to the same number of nonmatch trials in the other two conditions. For each participant, each latent print was randomly allocated to one of the three trial types, with the constraint that there were 12 prints in each condition. This way, each latent print has an equal chance to act in either a match, similar distractor or non-similar distractor trial, and so eliminating the possibility that a particularly easy/difficult/distinctive/high quality latent print could artificially influence examiners' performance.

3.10.3 Accuracy and Response Bias

There are two distinct measures of performance in a fingerprint comparison task. The obvious one is accuracy—an examiner's ability to discriminate matches from non-matches. The less obvious measure is response bias—the decision rule employed by an examiner when they are uncertain about a comparison. That is, their tendency to say "match" or "no match"

regardless of whether the prints match or not. If an examiner is unsure whether two prints match, and declares that they do, then they have made a 'liberal' decision. If an examiner is unsure whether two prints match, and declares that they don't, then they have made a 'conservative' decision. Averaged across several comparisons, the criteria used to make these decisions add up to so-called liberal and conservative response biases. Put simply, a response bias is a measure of a person's willingness to say, 'yes': if they say 'yes' a lot, then they have a liberal response bias; if they say 'no' a lot, then they have a conservative response bias.

Two examiners can be equally accurate in their ability to discriminate or 'see' matching prints, but—if they have a different response bias—they may come to opposite conclusions. It follows that there is no universal best response bias. There is an optimal criterion that minimizes false alarms and misses, but the appropriate decision criterion will depend on the costs and benefits of committing both types of error and both types of success. Only when the number of response alternatives is limited can an examiner's response bias be separated from their ability to discriminate prints.

Forcing a choice is a widely used paradigm for measuring human performance. In our experiment, participants were asked to judge whether print pairs matched, using a confidence rating scale ranging from 1 ("sure different") to 12 ("sure same") anchored at the centre (i.e., 6.5). The response scale forced a "no match" or "match" decision because ratings of 1 through 6 indicated no match, whereas ratings of 7 through 12 indicated a match. (Note that these ratings were described incorrectly in the *Procedure* section of our original paper.) That is, subjects were required to move the scrollbar either left (to 6 or less, "different") or right (to 7 or more, "same"); they could not make a rating of 6.5. This 12-point confidence scale was not designed to reflect the decisions made, and terms used, by examiners during casework. Judgments that the information was 'inconclusive,' which are often made in practice, were not permitted in this match/no match forced-choice design, making it possible to distinguish between accuracy and response bias (54). Interestingly, experts responded much more towards the extreme ends of the scale compared to novices: 92% of expert responses were either a 1 or a 12 compared to 32% for novices.

Aside from the capacity of the forced-choice procedure to differentiate the roles of accuracy and response bias, there are difficulties with measuring 'inconclusive' judgments. There is

no ground truth for sufficiency, that is, there is no way of knowing whether a print contains sufficient information for a human to discriminate it. The best that can be done is to ask several experts about the sufficiency of the information in several prints to see whether they agree with each other and themselves on repeated examinations (i.e., between and within participant reliability). What one means by 'insufficient' is also tricky. Does it mean, "There is not sufficient information in the latent print to make an identification," or does it mean, "I am unwilling to make a judgment (match/identification, no match/exclusion) either way." In fact, if a sufficient amount of information or signal was present (whatever that means), and an examiner declared it "inconclusive," then this ought to be regarded as a 'miss' type of error. Sufficiency of information in this experiment was partially controlled by only using prints that an expert declared as having sufficient information to make an identification. Participants' uncertainty in their judgments was also controlled by using a 12 point confidence scale where a rating of 6, for example, would be counted as a "nonmatch" decision.

3.11 Similarity

A pair of fingerprints will appear similar or dissimilar to each other (or somewhere in-between), depending on the amount of information in each and depending on the experience of the examiner. There is no agreed upon definition or measure of similarity for the comparison of prints, but there have been attempts to create an objective measure of similarity. For example, Vokey et al. (2009) converted a set of fingerprint images into their raw pixel values (i.e., the brightness values in each fingerprint image) and projected each print into the multidimensional space of all the prints in a set to return a vector, where the similarity of one print to another is given by the cosine of the angle between their vectors. A cosine value close to 1 indicates that the prints are virtually identical; whereas cosines close to zero indicate that the prints are highly dissimilar. This technique, therefore, provides an objective measure of similarity because it uses only the raw pixel values in the images and so requires no human input. We did not make use of this objective measure of accuracy for this experiment but, instead, used a national fingerprint database search.

For over twenty years, examiners have had the ability to search large databases, with the aid of computer algorithms, for potential matches to latent crime scene fingerprints. Although no formal data exist, it is likely that the majority of fingerprint comparisons made today use database queries (suspect-absent cases) rather than with a closed set of known prints from a suspect (suspect-present cases). A database query on a latent print will return a list of candidates that are most similar (according to the algorithm) to that latent. As Vokey et al. (2009) and Dror and Mnookin (2010) discuss, a database query makes an examiner's task much more difficult by returning a set of highly similar distractor prints—prints that look very much alike (according to the algorithm) but come from different people. These searches, by their very nature, are maximizing the conditions conducive to false positives. What's more, Vokey et al. found that novices made more false alarms on comparisons that were similar—as measured by the distance between vectors of pixel maps—than those that were not similar. Given that distinguishing highly similar, but nonmatching, prints from genuine prints is likely to be the most difficult and common task that examiners face, similarity was included as a factor in our experiment.

Similar distractor prints were obtained by searching our simulated crime scene latents on the Australian National Automated Fingerprint Identification System (NAFIS). The latents were first auto-coded and then hand-coded by a qualified expert. For each simulated crime scene print, the most highly ranked nonmatching exemplar from the search was used if it was available in the Queensland Police 10-print hard-copy archives, which contains approximately 3.3 million prints. The corresponding 10-print card was retrieved from the archives, scanned, and the individual print of interest was extracted. Due to the proprietary nature of NAFIS, the information it uses in its search algorithms is unknown; but, it almost certainly relies on the minutiae, features, direction, relative and spatial relationships of the fingerprints as identified by human examiners, rather than low-level pixel values. These print comparisons were labeled 'similar distractors', but it is important to note that the NAFIS algorithm is a search aid and was not designed to model human performance; so what the NAFIS algorithm regards as similar may or may not correspond to what a human examiner considers similar.

3.12 Establishing Expertise: Novices as a Control Group

Research on the nature of human expertise has developed over decades and ranges from the seemingly disparate domains of chess to medical diagnosis. One goal of expertise research in cognitive science is to understand the mechanisms that account for the superior performance of experts, across domains (Ericsson & Smith, 1991). Another is understanding the domain specific nature of expertise: Why does superior performance in one domain not transfer to others? Can we hasten the transition from novice to expert? Broadly, what makes an expert, an expert?

In order to study the nature of expertise in fingerprint identification, it first needs to be demonstrated that expertise actually exists; that is, are there people who possess exceptional abilities for matching latent fingerprints to their source? Despite over 100 years of fingerprint testimony—and although implicit in the terminology referring to fingerprint examiners who are qualified to testify in court (i.e., a 'fingerprint expert')—there had been no study demonstrating that qualified examiners have specialized discrimination skills or abilities superior to those of the person on the street (but see Langenburg et al., 2009, for a bias experiment with novices) for a bias experiment with novices. Superior expert discrimination performance in fingerprint identification had been assumed and the nature of that expertise had not been proposed or demonstrated. But should people who have no training or experience be expected to accurately match prints to their source?

It is clear from Vokey et al. (2009) that novices generally have substantial abilities to match fingerprints. In a fingerprint matching task, naive undergraduates were able to discriminate fingerprint matches from non-matches quite well, or well above chance at least. With these findings in mind—and without any requisite experiments of expert performance in the forensic use of fingerprint identification—it was not obvious (to us at least) that experts would outperform novices in our experiment as much as they did. Furthermore, and as Vokey et al. note, one pioneer of fingerprinting, Sir Francis Galton (1893), believed that experts would quickly become unnecessary and that lay juries would eventually evaluate fingerprint evidence.

In 2002, Louis H. Pollak (a senior federal judge in Philadelphia) ruled, in United States v. Llera Plaza, that fingerprint evidence does not meet the standards set for scientific testimony and that experts in the field cannot testify that a suspect's prints definitely match those found at a crime scene (Cho, 2002a). Pollak ruled that fingerprint experts could still point out the similarities between prints from a crime scene and those of a defendant, but the ultimate decision should be left to the jury. This decision was eventually overturned (Cho, 2002b), but it is clear that one option for expert testimony under consideration is for experts to present the physical evidence, with commentary attached, and allow lay juries to decide whether a latent crime scene print matches the suspect.

Considering both the evidence for the reasonable performance of novices and the notion that juries should make the ultimate decision, it seems that the most appropriate comparison group to demonstrate expertise should be novices who have no training with fingerprints whatsoever. In the *Identifying Fingerprint Expertise* experiment, the matching performance of qualified fingerprint examiners was compared to the performance of novice undergraduates who had no experience or training with prints in order to establish the supposition of expertise in fingerprint identification. Novices were thirty-seven psychology undergraduates from The University of Queensland who participated for course credit. Experts were thirty-seven qualified practicing fingerprint experts from five police organizations (the Australian Federal, New South Wales, Queensland, South Australia, and Victoria Police) who volunteered during our visit to their department. Their experience with prints ranged from 5 to 32 years and was 17.45 ($SD = 7.53$) years on average. We found that qualified, court-practicing fingerprint experts are exceedingly accurate compared with novices. Even though the novices could reliably distinguish matching and nonmatching prints, they made a large number of errors.

The performance difference between experts and novices on trials in which the prints matched was relatively small (92.12% correct for experts vs. 74.55% for novices). Comparably, the performance difference between experts and novices on trials in which the prints did not match, and were not similar, was also relatively small (100% correct for experts vs. 77.03% for novices). The performance difference between experts and novices on trials in which the prints were highly similar but did not match, however, was substantial; novice participants mistakenly identified 55.18% of the similar, nonmatching distractor prints as matches, whereas

the corresponding rate for experts was 0.68%. The largest performance difference between novices and experts seems to lie in identifying highly similar, but nonmatching prints, as such. A comparison to novices was important for demonstrating expertise and shows that the matching task was difficult enough for experts to perform accurately, but for novices to perform relatively poorly.

3.13 Error Rates

Much has been made about 'error rates' in fingerprint identification, and more so in light of the National Academy of Sciences report (National Research Council, 2009). In this section we attempt to characterize error rates in our experiment and for fingerprint identification more generally.

3.13.1 Expert Matching Accuracy

In our experiment, 37 experts each compared 36 print pairs for a total of 1332 comparisons. Of the 444 comparisons in which the prints matched (targets), 22 of the 37 examiners incorrectly declared at least one of these matching prints as nonmatches, for an absolute total of 35 misses (hits = 92.12%; misses = 7.88%). Misses are the kind of error that can lead to a failure to identify a criminal. Of the 444 comparisons in which the prints did not match, and were not similar (nonsimilar distractors), all of the examiners correctly declared the prints as nonmatches (correct rejections = 100%; false alarms = 0%). Of the 444 comparisons in which the prints did not match, but were highly similar (similar distractors), three examiners incorrectly declared three of these print pairs as matches (correct rejections = 99.32%; false alarms = 0.68%). These three print pairs were of three different latents. False alarms are the kind of error that can lead providing incorrect evidence to the court. (As an aside, it is not possible to link a participant's performance to the identity of a particular individual because experts and novices participated anonymously.) What, then, can be concluded about error rates from this experiment?

3.13.2 Determining Error Rates

Our study was not designed to determine the likelihood of errors in practice, nor the performance of individual practitioners or departments. It was designed to demonstrate expertise in fingerprint identification. Inferring, from our results, that "Fingerprint examiners are 99.32% accurate," or "The error rate of fingerprint identification is 0.68%," would be disingenuous. Any claim of accuracy would have to be followed by a list of qualifiers. For example, "Some qualified fingerprint examiners are 99.32% accurate at correctly declaring nonmatching prints as such when the prints were obtained from the most similar non-match according to the NAFIS system, and when examiners are not provided with their usual tools, or independent verification," and so on. The qualifiers are limitations only when trying to make an overgeneralization like, "fingerprint examiners are 99.32% accurate," which is close to impossible to make in any area of expertise (let alone on the basis of our results). It is, however, legitimate to conclude that experts are more accurate (and conservative) than novices, for example. If this is what we are concluding (and we are), then all of the qualifying remarks above are completely irrelevant.

Readers might be now considering some particular qualifiers to explain these results. For example, one might imagine that the ability to zoom and rotate prints will improve expert performance or that verification will reduce experts' error rate to zero. But it is unlikely that one particular qualifier will be enough to entirely explain the results. Regardless, these qualifiers (and many others) are all testable hypotheses, should the answers be seen as necessary and important. If our goal—in addressing the critics and advancing the field—is to determine the precise error rate for each fingerprint examiner in each department, or even the field as a whole, then the necessary experiments become unwieldy.

We would need to unpack the different types of error that are possible (e.g., clerical, identification, sufficiency, misses, false alarms, disagreement, inadmissible rulings, etc.). We would need to establish who is going to be tested (e.g., trainees, intermediates, qualified experts, supervisors), under what conditions (e.g., distractions, interruptions, sleep deprivation, time and resource constraints, etc.), and on which part of the process (e.g., latent development, analysis, comparison, testifying, etc.). Will we include verification (and will it be blind)?

How will we ensure ground truth? What sort of materials will be examined (e.g., lifts, surface types, lifting agents, powders, whorls, arches, low quality, distorted, highly similar, etc.)? What tools will be available (e.g., image enhancements, digital markers, zoom, rotation, computer search algorithms, statistical models, etc.)? How much time will they have? Can they collaborate? How many items will they be tested on, and so on. In the end, we would be left with some numbers (e.g., 85% hits and 6% false alarms for Jones under x, y, and z conditions). What, then, do we do with this information? It is difficult to see how the incredible amount of time, money and resources required to get such an answer, would pay off. As discussed earlier, this approach may be ineffective and inefficient. And we will be unable to locate the source of errors, to say nothing of taking steps to avoid them.

3.13.3 Levels of Analysis

Is a measure of individual error necessary for the science or for the court? The fingerprint profession, of course, will be concerned with their performance to make sure that they are on track and to ensure continuous improvement. And, of course, the courts will be concerned with data that will help fact finders make optimal decisions. But, for example, demands are not made for individual error rates for a medical doctor or a field-wide error rate in medical diagnosis; only performance measures of the instrument or test on average are sought. To ask for error rates associated with a particular individual on a particular test seems, rightly, inappropriate in medicine. Similarly, focusing on the individual is the wrong level of analysis when attempting to characterize the accuracy of the forensic fingerprint identification system.

A broader question concerns the level of analysis that is appropriate for presenting evidence and associated rates of error in court. At the extreme, an examiner could report how accurate they are at matching a whorl type print, lifted from a crime scene, on a wooden surface, using magnetic black powder, in a particular department, in a particular country, on a Tuesday, and so on. It may be necessary, or the courts may demand, that particular rates of error are established for particular situations. But unless it has been demonstrated that the level of accuracy (or proficiency, or reliability, or competence) varies systematically

at any one of these levels, then the default should be to opt for reporting accuracy at the broader level.

3.13.4 Error Rates in Other Domains

Comparing the matching performance of fingerprint experts to experts in similar domains may give us an appreciation for their relative performance. Unlike with fingerprints, all people have expertise with faces. Psychologically, faces are similar to prints in that they are both complex visual patterns (Busey & Parada, 2010). People can easily recognize familiar faces despite changes in facial expression, context and viewpoint (Johnston & Edmonds, 2009). Unfamiliar faces, however, are extremely difficult to identify across these changes. Even in ideal conditions, where an unfamiliar target face is presented next to a set of candidate faces—with similar lighting, poses and no time constraints—people only match 68% of the faces correctly (Megreya & Burton, 2008). Even people whose job requires them to identify unfamiliar faces from identity cards perform poorly at this simultaneous matching task (Kemp, Towell, & Pike, 1997). Based on the results of Tangen et al. (2011), it is clear that fingerprint experts have impressive pattern matching abilities.

The accuracy of fingerprint experts becomes more impressive when compared to medical experts. Just as in fingerprint identification, however, it is difficult or impossible to determine general rates of field-wide error. But it is known that roughly 5% of autopsies reveal lethal diagnostic errors for which a correct diagnosis coupled with treatment could have averted death, and an estimated 40,000 to 80,000 US hospital deaths result from misdiagnosis annually (Kohn, Corrigan, & Donaldson, 2000; Newman-Toker & Pronovost, 2009). These figures suggest that more Americans are killed in US hospitals every 6 months than died in the entire Vietnam War, and is equivalent to three fully loaded jumbo jets crashing every other day (but see Hayward & Hofer, 2001). The prevalence of false positive diagnostic errors in perceptual specialties, such as radiology and pathology, is typically less than 5%, and increases to the range of 10-15% in emergency room type settings (Berner & Graber, 2008; Norman & Eva, 2010). Researchers are now focused on ways to reduce (not eliminate) diagnostic errors and on creating policy that defines acceptable rates of error (Newman-Toker & Pronovost, 2009).

Of course it is difficult to define 'error rates', let alone compare them across domains. But it is clear, from the available data, that fingerprint experts demonstrate impressive pattern matching abilities that may rival those of medical diagnosticians; even despite the distinction that, arguably, identification (as in fingerprints) is a more difficult task than categorization (as in medical diagnosis).

3.14 Summary

Thus far, we have expanded on the results of the *Identifying Fingerprint Expertise* experiment (2011), and explained that previous experiments and tests of proficiency were problematic and that the expertise of human fingerprint examiners had been assumed but not demonstrated. We have described the decisions and compromises that we made in order to design an experiment that tests the claimed expertise of human fingerprint examiners. In summary:

- Fidelity, generalizability and control must be balanced in order to answer research questions. Our experiment was 'artificial' for good reason. The goal was to understand the extent to which the difference between expert and novice performance is real, not the extent to which the experimental setting resembles the everyday operations of a fingerprint bureau.

- The validity, proficiency, and competence of fingerprint examiners is best determined when experiments include highly similar print pairs where the ground truth is known. Prints from the Forensic Informatics Biometric Repository were used to ensure ground truth.

- In order to quantify matching performance, a signal detection paradigm must be employed to separate the two ways of being right and the two ways of being wrong, to compare performance on matching and nonmatching prints, and to separate accuracy and response bias.

- Distinguishing highly similar, but nonmatching, prints from genuine prints is likely to be the most difficult and common task that examiners face. Similar distractor prints were obtained by searching simulated crime scene latents on the Australian National Automated Fingerprint Identification System (NAFIS) to emulate this task.

- Considering both the evidence for the reasonable performance of novices and the notion that juries should make the ultimate decision, the most appropriate comparison group to demonstrate expertise should be novices who have no training with fingerprints whatsoever.

- Our study was not designed to determine the likelihood of errors in practice, nor the performance of individual practitioners or departments. As such, inferring from our results that, "Fingerprint examiners are 99.32% accurate," or "The error rate of fingerprint identification is 0.68%," would be disingenuous.

- Determining error rates with black box studies may be unnecessary at best and ineffective and inefficient at worst, and unless one can demonstrate that a particular qualifier will systematically affect accuracy, the default should be to report accuracy at the broader level.

- Fingerprint experts posses impressive pattern matching abilities that may rival those of medical diagnosticians.

It appears that expertise in fingerprint identification does exist. That is, there are people who have demonstrable and specialized abilities for matching latent fingerprints to their source, and those abilities are superior to the person on the street. An examiner's expertise seems to be situated, not in their ability to match prints per se, but in their superior ability to identify highly similar, but nonmatching fingerprints as such. These results, and their comparison to novices, shows that the accuracy of qualified examiners is substantially higher than inexperienced novices. Moreover, the experiment was designed to be difficult. The fact that experts made so few errors is evidence for impressive human pattern matching performance possibly exceeding that of experts in other comparable domains. It seems that

some combination of training and the daily comparison of untold numbers of fingerprints leads to an uncanny ability to match fingerprints to their source. Experts are drawing on an entire career of experiences in making their decisions, as well as their training in fundamentals of fingerprint impressions to understand the 'behavior' of minutiae. Experts, likely implicitly, understand the structure, regularities and acceptable variation of fingerprint impressions. Future experiments could pinpoint the nature of this expertise. That is, whether expertise arises mainly from formal rules or the accumulation of instances (Norman et al., 2007).

3.15 Implications for Expert Testimony

The results from Tangen et al. (2011) demonstrate that qualified fingerprint experts perform much better than novices at matching fingerprints and their rates of error may be lower than those in diagnostic medicine, for example. Below the implications of these results for current models of expert admissibility, testimony, and policy are discussed.

3.15.1 The Current Model

Current models of expert testimony vary from country to country and from state to state. The two largest bodies that provide consensus guidelines and standards for fingerprint identification—the Scientific Working Group on Friction Ridge Analysis, Science and Technology (2011b) and the International Association for Identification (2007)—both stipulate that examiners are only permitted to testify to three conclusions: exclusion, inconclusive and individualization. An individualization is defined as, "The determination by an examiner that there is sufficient quality and quantity of detail in agreement to conclude that two friction ridge impressions originated from the same source." When testifying, examiners often do not provide evidence of their claimed expertise or attempt to characterize their level of proficiency. Examiners may, when pushed by the courts, report that all fingerprints are unique or point to prenatal development and persistence. Despite considerable acrimony, examiners continue to make claims of individualization or similar (Cole, 2009, 2010).

3.16 The Implications of the Identifying Fingerprint Expertise Experiment

3.16.1 Admissibility

Information about accuracy and performance—along with the relative performance of laypersons—is required for courts to make informed decisions about the admissibility of expert testimony. Experts outperform novices, but they do make errors (Tangen et al., 2011). These results make it less likely that examiners themselves will suffer unfounded attacks on their expertise. If an examiner's expertise is challenged, then the methodology and design of the experiment ought to be the target of criticism rather than the examiners themselves; assuming, of course, that their testimony does not extend beyond what experiments can support.

The distinction, between the performance of experts and novices, is fundamental to the question of expert testimony, because it demonstrates specialized knowledge. This experiment could be used as evidence for this distinction in order to satisfy legal admissibility criteria. And the results suggest that relying on juries to evaluate fingerprint evidence—when presented with the physical evidence alone, without expert commentary (Galton, 1893)—could result in a substantial number of false identification errors. The National Academy of Sciences Report (National Research Council, 2009; Edwards, 2009b) noted the frequent absence of solid scientific research demonstrating the validity of forensic methods in general; of quantifiable measures of the reliability and accuracy of forensic analyses; and of quantifiable measures of uncertainty in the conclusions of forensic analyses. Fingerprint examiners have taken a first step in demonstrating their claimed expertise in controlled, representative situations in which ground truth is known. Examiners are now working with researchers towards understanding the source of identification errors, the factors that influence performance, and the nature of expertise in identification. In light of the National Academy of Sciences report, and the model for demonstrating expertise provided here by fingerprint examiners and researchers, it behooves other forensic pattern identification disciplines—such as shoeprints, bloodstains,

DNA, ballistics, toolmarks, bitemarks, CCTV face identification, etc.—to conduct similar experiments to demonstrate expertise and performance in their own disciplines.

3.16.2 Testimony

Documented cases of false identification, issues of plausibility reported by the National Academy of Sciences (National Research Council, 2009), and recent experiments (Tangen et al., 2011; Ulery et al., 2011, 2012; Dror et al., 2011) highlight the need for a contemporary model of forensic testimony. Following developments in the US and Canada, Edmond (Edmond, 2008) has suggested that Australia adopt a reliability standard, and the UK Law Commission (Campbell, 2011) has announced similar recommendations for admissibility practice in England and Wales. Indeed, science and legal commentators are beginning to call for empirical demonstrations of accuracy and performance, along with details about the relative performance of laypersons, across forensic science. A failure to respond to criticism means that judges are in danger of acting irrationally and being left behind by practical and ongoing reforms in the forensic sciences. While it is likely that courts will start to develop an admissibility and testimony jurisprudence more directly concerned with reliability in the near future, there is an independent need for forensic scientists and technicians to pay much closer attention to the evidence for ability and reliability (Edmond, 2011). Edmond (2008) suggests that reliability standards will help to make criminal trials fairer and ensure outcomes reflect the known value of expert evidence.

It is clear that an alternative to the current model of fingerprint testimony is required. But what should an acceptable alternative and contemporary model look like? Several factors must be considered; these include the role of scientific experiments on the accuracy and reliability of forensic identification; whether it is necessary to report department or individual scores on properly controlled proficiency tests; the state of the science in other areas of pattern and impression identification; the impact that the testimony has on jury decision making; finding the right balance between accurate scientific reporting and the ability of judges and juries to understand expert testimony; and decisions about whether to report on the degree to which the specimen matches the source (e.g., “lends limited support”), the

degree of confidence in a match (e.g., "highly confident that x matches y"), opinions about the evidence (e.g., "it is my opinion that..."), or statements about the particular hypotheses in question (e.g., the evidence is more consistent with x than y).

There is much research and consideration needed to develop an acceptable alternative model of fingerprint testimony. Several debates on this topic are raging internationally between academics, statisticians, lawyers, forensic examiners and managers. We are working towards proposing recommendations that do not extend beyond the capabilities of examiners or experimental findings while substantially engaging with critics in order to develop robust empirical guides to practice.

3.17 To Develop a Research Culture in Forensic Science

Researchers and professionals (e.g., Mnookin et al., 2010) have highlighted the need for a research culture in forensic science. Currently, however, it appears that professionals are expected to strengthen the scientific basis of their field but are not provided with the financial or intellectual support to do so. It is clear that examiners are doing their best to capture criminals and uphold civil liberties. But the lack of funding and resources in already overworked forensic departments makes basic research exceedingly unlikely. In addition, few have the methodological skills and expertise in the psychology of perception, cognition, bias, memory, accuracy, and decision making to ensure that their practice meets legal admissibility standards emerging internationally.

It is essential that we move beyond the adversarial system currently impeding advancement of the field, and develop a culture of cooperation between researchers and examiners. The emergence of such a culture would fundamentally change examiners' relationship with empirical data and affect how evidence is understood and reported. Indeed, forensic examiners have expressed a desire to address the shortfalls of their discipline and engage in research.

Considering that forensic identification is based on human judgment, the field would benefit from further research on expert decision making. Clinical reasoning in medicine, for

example, has developed over the last 40 years after it became increasingly apparent that physicians' decisions resulted in adverse consequences for patients (Kohn et al., 2000). Much has been learned about the nature of medical expertise, the influence of perceptual and cognitive biases, and how to best incorporate such knowledge into practice. Researchers need to provide a scientific basis for demonstrating the validity of forensic methods and measures of uncertainty in the judgments of forensic analyses.

From here, more sophisticated questions can be asked than those about error rates and expertise. For example, what is the most effective way to train novices? What information is the most important for matching or excluding prints? What elements of the matching task best distinguish experts and novices? How do experts and novices differ in their use of this information? How does expertise with fingerprints develop over time? What is the relationship between the Analysis and Comparison phase of the identification process? How does time pressure influence performance? What is best practice in providing feedback and self-assessment? What is the most effective way to present fingerprint evidence to juries? The practical outcomes from answering questions such as these include a better understanding of the source of potential identification errors and factors that influence performance, a reduction in training time from novice to expert, more effective recruitment and training methods, and greater validity in presenting forensic evidence in court.

Maintaining high standards of evidence is vital for an effective justice system and ensuring that innocent people are not wrongfully accused. The reliability of forensic evidence and the value of expert testimony in the criminal justice system can be maximized by examining forensic reasoning and decision making. Given the inevitability of human error, the move should be towards fostering resilient systems capable of minimizing and acknowledging errors (Hollnagel, Woods, & Leveson, 2012). We—collectively, forensic professionals, researchers, legal scholars and the courts—need to define acceptable rates of error, foster a work environment conducive to learning from error, and promote a blame-free safety culture, as medicine is working towards (Newman-Toker & Pronovost, 2009; Woods, Johannesen, Dekker, Cook, & Sarter, 2010).

This approach will allow police, intelligence systems and investigators to interpret evidence more effectively and efficiently, assist forensic examiners in the development of evidence-based

training programs, discourage exaggerated interpretations of forensic evidence, and help in the development of a model of expert testimony that does not extend beyond the capabilities of examiners or beyond the scope of experimental findings. Further psychological research into forensic decision making will help to ensure the integrity of forensics as an investigative tool available to police, so the rule of law is justly applied.

Chapter 4

Human Matching Performance of Genuine Crime Scene Latent Fingerprints

4.1 Preface

This chapter is extracted from a published article in the journal *Law and Human Behavior*. As can be seen in Figure 4.1, this is the third and final chapter of PART 1 - ESTABLISHING EXPERTISE. The majority of the work is my own, with Jason Tangen contributing 50% to the experimental design, 10% to other areas, and 25% to writing of the final publication. Reference:

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Human matching performance of genuine crime scene latent fingerprints. *Law and Human Behavior.* doi: 10.1037/lhb0000051

The experiment from Chapter 2 was tightly controlled, making use of stimulated crime scene prints where the ground truth is known, which was especially important because the ground truth of previous experiments was uncertain. This high control, however, came at the cost of fidelity—we did not know the extent to which our simulated crime scene prints were

representative of case work. We worked with police officers to get access to the materials they use during formal training (collected over several years of casework) and used them in this experiment. The advantage here is that the prints are highly representative of the prints examiners deal with in everyday casework. We also added two trainee groups.

A minimum of five years training and experience is required to become a qualified, court-practicing fingerprint examiner in Australia, and they are often under pressure from management to reduce training time. Training times in the US vary widely, but, at the extreme, police officers have been known to complete a weekend training course in fingerprint identification and then testify as an expert witness in court. In this experiment we tested new and intermediate trainees to measure their performance relative to novices and experts, and to see how expertise might develop over time. Again, we did not expect examiners to perform as well as they did.

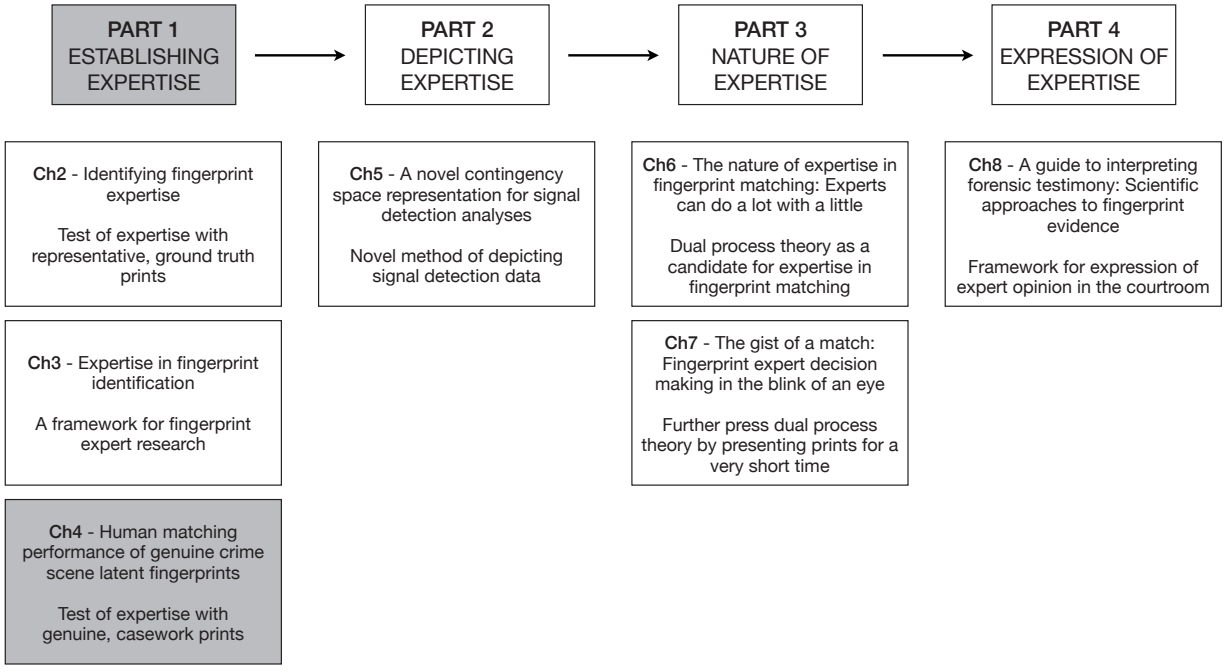

Figure 4.1: Conceptual diagram highlighting Chapter 4, Part 1 of the thesis: "Human matching performance of genuine crime scene latent fingerprints."

4.2 Abstract

There has been very little research into the nature and development of fingerprint matching expertise. Here we present the results of an experiment testing the claimed matching expertise of fingerprint examiners. Expert ($n = 37$), intermediate trainee ($n = 8$), new trainee ($n = 9$), and novice ($n = 37$) participants performed a fingerprint discrimination task involving genuine crime scene latent fingerprints, their matches, and highly similar distractors, in a signal detection paradigm. Results show that qualified, court-practicing fingerprint experts were exceedingly accurate compared with novices. Experts showed a conservative response bias, tending to err on the side of caution by making more errors of the sort that could allow a guilty person to escape detection than errors of the sort that could falsely incriminate an innocent person. The superior performance of experts was not simply a function of their ability to match prints, *per se*, but a result of their ability to identify the highly similar, but nonmatching fingerprints as such. Comparing these results with previous experiments, experts were even more conservative in their decision making when dealing with these genuine crime scene prints than when dealing with simulated crime scene prints, and this conservatism made them relatively less accurate overall. Intermediate trainees—despite their lack of qualification and average 3.5 years experience—performed about as accurately as qualified experts who had an average 17.5 years experience. New trainees—despite their 5-week, full-time training course or their 6 months experience—were not any better than novices at discriminating matching and similar nonmatching prints, they were just more conservative. Further research is required to determine the precise nature of fingerprint matching expertise and the factors that influence performance. The findings of this representative, lab-based experiment may have implications for the way fingerprint examiners testify in court, but what the findings mean for reasoning about expert performance in the wild is an open, empirical, and epistemological question.

4.3 Introduction

Fingerprint examiners have been active in investigations and have presented identification evidence in criminal courts for more than a century (Cole, 2002). Remarkably, given that testimony about fingerprint matches is a product of human judgment and subjective decision making, there have been few scientific investigations of the human capacity to correctly match fingerprints. Examiners have claimed that fingerprint identification is infallible (Federal Bureau of Investigation, 1984) and that there is a zero error rate for fingerprint comparisons (Cole, 2005; Edwards, 2009b). These claims of individualization and a zero error rate, however, are not supported by evidence and are scientifically implausible (Dror & Cole, 2010; National Research Council, 2009; Saks & Faigman, 2008; Saks & Koehler, 2005). As a result, former President of the International Association for Identification suggested that members not assert 100% infallibility (zero error rate) of fingerprint comparisons (R. Garrett, 2009) and the Scientific Working Group on Friction Ridge Analysis, Study and Technology (2011a) has drafted a standard for defining, calculating, and reporting error rates. Recently, there has been a shift in the way fingerprint identification is regarded (Tangen, 2013). The acknowledgment that humans cannot be detached from forensic decision making has been highlighted in a variety of recent inquiries by the U.S. National Research Council of the National Academy of Sciences (2009), the Scottish Fingerprint Inquiry (Campbell, 2011), and the National Institute for Standards and Training and the U.S. National Institute of Justice (Expert Working Group on Human Factors in Latent Print Analysis, 2012). The National Academy of Sciences (NAS, 2009) has highlighted the absence of solid scientific methods and practices in U.S. forensic science laboratories. Harry T. Edwards (a senior U.S. judge and cochair of the NAS Committee) noted that forensic science disciplines, including fingerprint comparison, are typically not grounded in scientific methodology, and forensic experts do not follow scientifically rigorous procedures for interpretation that ensure that the forensic evidence that is offered in court is valid and reliable (Edwards, 2009b; Risinger et al., 2002; Saks & Koehler, 2005). The NAS report (2009) highlighted the absence of experiments on human expertise in forensic pattern matching: “The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine

its validity. This is a serious problem." They recommended that the U.S. Congress fund basic research to help the forensic community strengthen their field, rectify the lack of basic research, develop valid and reliable measures of performance, understand the effects of bias and human error, and establish evidence-based standards for analyzing and reporting forensic testimony. Subsequent reports in the United Kingdom and United States have focused directly on fingerprint evidence. An inquiry into fingerprint evidence was conducted by Lord Campbell (2011) following the controversial McKie case in Scotland. The former police detective, Shirley McKie, was accused by fingerprint examiners of leaving her fingerprint on the bathroom door frame of a murder crime scene, a charge she denied. The report recommends that fingerprint evidence should be recognized as opinion evidence, not fact; examiners should discontinue reporting conclusions on identification or exclusion with a claim to 100% certainty or infallibility; and that examiners should receive training that emphasizes that their findings are based on their personal opinion and subjective interpretation. Most recently, a large multidisciplinary collective—the Expert Working Group on Human Factors in Latent Print Analysis (2012)—was sponsored by the U.S. National Institute of Standards and Technology and the National Institute of Justice to investigate human factors in latent fingerprint identification. The authors recommended that examiners should be familiar with human factors issues such as fatigue, bias, cognitive and perceptual influences, and not state that errors are inherently impossible or that a method inherently has a zero error rate. They recommend that management foster a culture in which it is understood that some human error is inevitable and that a comprehensive testing program of competency and proficiency should be developed and implemented. Speaking generally, and taking the lead from medical and aviation research, the authors advocate that fingerprint identification would benefit from the human factors research and systems approaches to improve quality and productivity, and reduce the likelihood and consequences of human error. As a result of these reports and of scholarly criticism, changes in policy and research programs have begun. There are two proposed bills currently before the United States Congress calling for more research into forensic identification and changes to the funding, organization, standards, and regulation of forensic science (Carle, 2011; Maxmen, 2012), and research into fingerprint identification is well underway. Researchers have investigated the effect of contextual bias

on fingerprint examiners (Dror & Cole, 2010; Dror & Rosenthal, 2008; Langenburg et al., 2009), the special abilities and vulnerabilities of fingerprint examiners (Busey & Dror, 2010; Busey & Parada, 2010; Busey et al., 2011), the psychophysics of fingerprint identification (Vokey et al., 2009), the effect of technology (Dror & Mnookin, 2010; Dror et al., 2012), and statistical models of fingerprint identification(Champod & Evett, 2001; Neumann et al., 2007; Neumann, 2012). Two recent experiments have been conducted to directly address the matching accuracy and expertise of examiners. Ulery, Hicklin, Buscaglia, and Roberts (2011) set out to measure the matching performance of latent print examiners. They had 169 latent print examiners each compare around 100 pairs of latent and exemplar fingerprints from a pool of 744 pairs. They focused on examiners' accuracy in the comparison process (i.e., the extent to which examiners can accurately match a latent print to its source). The researchers manufactured their own latent fingerprints so the ground truth is known, and they included similar, but nonmatching, distractors from a search of a national computer database containing approximately 580 million individual fingerprints. They reported an overall false alarm rate of 0.1% (i.e., incorrectly judging nonmatching prints to be a "match"). And 85% of examiners made at least one miss (i.e., incorrectly judging matching prints to be a "nonmatch") for an overall miss rate of 7.5%. Refer to Figure 4.2 for a description of the two ways of being right and two ways of being wrong in a basic fingerprint comparison task. Note, however, that Ulery et al. (2011) allowed examiners to give "inconclusive" and "no value" responses, and when the no value responses are discounted and the inconclusive responses are translated into misses, the overall miss rate is closer to 60% (an extremely conservative response bias). Although the experiment did not include a comparison group of participants (e.g., laypersons), it is clear that fingerprint examiners demonstrate impressive pattern matching abilities that may rival those of medical diagnosticians (M. B. Thompson, Tangen, & McCarthy, 2013a). The rigorous experimental design, coupled with the large number of participants and stimuli, makes it one of the most important contributions to our understanding of expert matching performance. Tangen, Thompson, and McCarthy (2011) set out to determine whether fingerprint experts are any more accurate at matching prints than the average person, and to get an idea of how often experts make errors of the sort that could allow a guilty person to escape detection compared with how often they make

errors of the sort that could falsely incriminate an innocent person. In a two-alternative forced choice design, 37 qualified fingerprint experts and 37 undergraduate students were presented with pairs of fingerprints and asked to indicate whether a simulated crime scene print matched a potential "suspect" or not. Some of the print pairs matched, and others were highly similar but did not match. Thirty-six simulated crime scene prints were paired with fully rolled exemplar prints. Across participants, each simulated print was paired with a fully rolled print from the same individual (match), with a nonmatching but similar exemplar (similar distractor), and with a random nonmatching exemplar (nonsimilar distractor). The simulated prints and their corresponding fully rolled print were from the Forensic Informatics Biometric Repository (see `FIB-R.com` for details), so, unlike genuine crime scene prints, they had a known true origin (Cole et al., 2008; Koehler, 2008). Similar distractors were obtained by searching the Australian National Automated Fingerprint Identification System. For each simulated print, the most highly ranked nonmatching exemplar from the search was used if it was available in the Queensland Police 10-print hardcopy archives, which contains approximately 1 million 10-print cards (10 million individual prints) from approximately 300,000 to 400,000 people (one person may have more than one 10-print card on record). Of the prints that actually matched, the experts correctly declared 92.12% of them as matching (hits). Of the prints that did not actually match, the experts incorrectly declared 0.68% of them as matching (false alarms). The rate of expert false alarms is impressive considering the corresponding false alarm rate for novices was 55.18%. Tangen et al. (2011) concluded that the qualified court-practicing fingerprint experts were exceedingly accurate compared with novices, and that the experts tended to err on the side of caution by making errors of the kind that would fail to identify a criminal rather than provide incorrect evidence to the court. The experiment by Tangen et al. (2011) did not focus on the absolute performance of experts but on the comparison between experts and novices and between matching and nonmatching prints. So even though a false alarm rate for experts of 0.68% is impressive in its own right, this experiment cannot determine whether this rate reflects the false alarm rate of the field more generally. But it can be concluded that a false alarm rate of 55.18% for novices pales in comparison with experts (Thompson et al., 2013). Despite the above contributions to forensic decision making, still very little is known about human fingerprint matching performance,

the nature of expertise in fingerprint identification, the factors that affect matching accuracy, and the basis on which examiners can reasonably testify in court. Considering the shift toward viewing the human as an integral part of the forensic identification process, systematic programs of research are needed to understand the skills, abilities, and limits of fingerprint examiners, and to understand the nature of their expertise. Research programs that are under way or are to be developed include understanding the nature of forensic expertise, the influence of cognitive and perceptual biases, the impact of technology, how best to present pattern evidence to judges and juries, the best ways to turn novices into experts, and the most effective and efficient work practices, environments, and tools. Before these research programs can advance, however, a foundation for understanding expertise and accuracy in human fingerprint identification is needed. Here, we present a first step in our research program into the nature of forensic expertise in fingerprint identification.

4.4 Overview of the Present Research

In the experiment reported here we investigated the matching performance and expertise of human fingerprint examiners by replicating and extending on the work of Tangen et al. (2011). We increased the fidelity of the discrimination task (i.e., the resemblance of the discrimination task to actual casework) by using genuine crime-scene latents (and their matched exemplars) from police training materials, compiled from casework. The increased fidelity, however, reduces experimental control because the ground truth of the matched fingerprint pairs cannot be certain (Thompson et al., 2013). We also made the addition of two trainee groups and asked four groups of people to perform a fingerprint discrimination task—novices, new trainees, intermediate trainees, and qualified experts—in order to compare their relative performance.

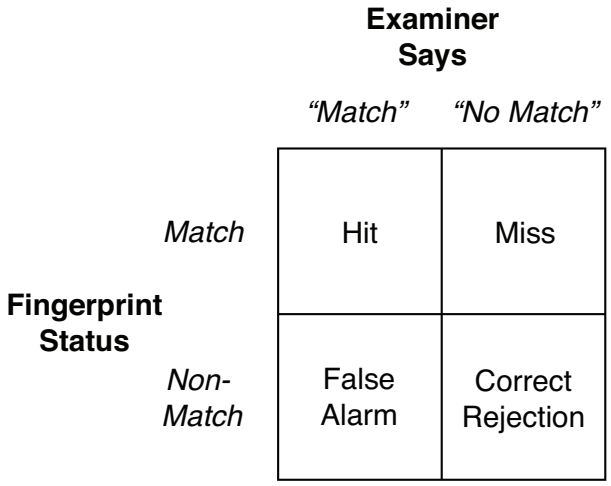

Figure 4.2: A 2×2 contingency table depicting the four possible outcomes of a forced choice fingerprint discrimination task where two prints match or not and an examiner declares them as a "match" or "no match".

4.5 Method

4.5.1 Participants

Four distinct groups participated in the experiment: novices, new trainees, intermediate trainees, and qualified, court-practicing experts. Novices were 37 undergraduates from The University of Queensland who participated for course credit and who had no experience with fingerprints. New trainees, intermediate trainees, and qualified experts were from five police organizations: The Australian Federal, New South Wales, Victoria, South Australia, and Queensland Police. New trainees included nine people who were training to be fingerprint experts. Five of these trainees had completed a 5-week training program on the day of testing, and four had been working in a fingerprint department for 5 or 6 months. Intermediate trainees included eight people who were training to be fingerprint experts. Of these, one had 1 year of experience, one had 2 years of experience, two had 3 years of experience, one had 4 years of experience, and three had 5 years of experience ($M = 3.5$, $SD = 1.51$). The distinction between the two types of trainees is arbitrary and was decided before the data were analyzed. Experts were 37 qualified court-practicing fingerprint experts with experience ranging from 5 to 32 years ($M = 17.45$, $SD = 7.53$).

4.5.2 Procedure

Participants were presented with pairs of prints displayed side-by-side on a computer screen, as illustrated in Figure 4.3. They were asked to judge whether the prints in each pair matched, using a confidence rating scale ranging from 1 (sure different) to 12 (sure same). Judgments were reported by moving a scroll bar to the left ("different") or right ("same"). The scale forced a "match" or "no match" decision, where ratings of 1 through 6 indicated no match, and ratings of 7 through 12 indicated a match. Judgments of "inconclusive" which are often made in practice, were not permitted in this two-alternative forced choice design, so it was possible to distinguish between accuracy and response bias (Green & Swets, 1966). A thorough explanation of the advantages of this approach can be found in Thompson et al. (2013). The methodology of this experiment emulates one aspect of the identification process, namely, the extent to which a print can be accurately matched to its source.

4.5.3 Stimuli

The stimuli consisted of 45 latent prints from a larger police training examination set and were paired with fully rolled prints. The latent prints were taken from actual crime scene casework and were used for training purposes. An examiner (the third author) developed the training set to provide comparison materials that would expose trainee experts with a larger volume of latent comparisons, and chose the experimental stimuli from the larger set such that they provided clear ridges for the NAFIS system to search on. The corresponding fully rolled matches were declared previously as identifications, and were verified by at least three expert examiners. Information about whether these identifications had any associated, and potentially corroborating, information such as a guilty plea, conviction, or independent DNA match was not available. Given that the prints were matched during casework, a qualified expert must have decided that each matching fingerprint pair contained sufficient information to make an identification. Across participants, each latent print was paired with a fully rolled print from the "same" individual (match), with a nonmatching but similar exemplar (similar distractor), and with a random nonmatching exemplar (nonsimilar distractor). For each participant, each latent print was randomly allocated to one of the three trial types, with the

constraint that there were 15 prints in each condition. Unlike the simulated latent prints taken from the Forensic Informatics Biometric Repository (as used by Tangen et al., 2011), the ground-truth of the matches cannot be certain. Similar distractors were obtained by searching each crime scene latent print on the Australian National Automated Fingerprint Identification System. For each latent print, the most highly ranked, nonmatching exemplar from the search was used if it was available in the Queensland Police 10-print hardcopy archives, which contains approximately 1 million 10-print cards (10 million individual prints) from approximately 300,000 to 400,000 people (one person may have more than one 10-print card on record). The corresponding 10-print card was retrieved from the archives, scanned, and extracted. In practice, highly similar nonmatches retrieved from large national databases are likely to increase the chance of incorrect identifications (Dror & Mnookin, 2010). Distinguishing such highly similar, but nonmatching, print pairs from actual matching print pairs is potentially the most difficult task that fingerprint examiners face (Dror & Mnookin, 2010; M. B. Thompson et al., 2013a). Latent prints were from printed photographs and were scanned in color as a 600-dpi lossless Tagged Information File Format (TIFF) file, converted to grayscale, cropped to 600 × 600 pixels, and isolated in the frame. The matching and nonmatching exemplars were fully rolled fingerprint impressions made using a standard elimination pad and a 10-print card or were digitally scanned via LiveScanTM. Each card was photocopied at 600-dpi and scanned in color as a 600-dpi lossless Tagged Information File Format (TIFF) file. Each print was then converted to grayscale, cropped to 600 × 600 pixels, and isolated in the frame.

4.6 Results

While analyzing the data, the pattern of results at the level of the trial type suggested that one of the latent prints in the set might not truly match the target exemplar. We sent the target pair to a qualified fingerprint examiner who declared that the prints, in fact, did not match. The source of the error arose from the police training materials spreadsheet that incorrectly labeled the finger type of a 10-print card, and so the incorrect print was extracted from the 10-print card. This transcription error has no relation to casework. As a result,

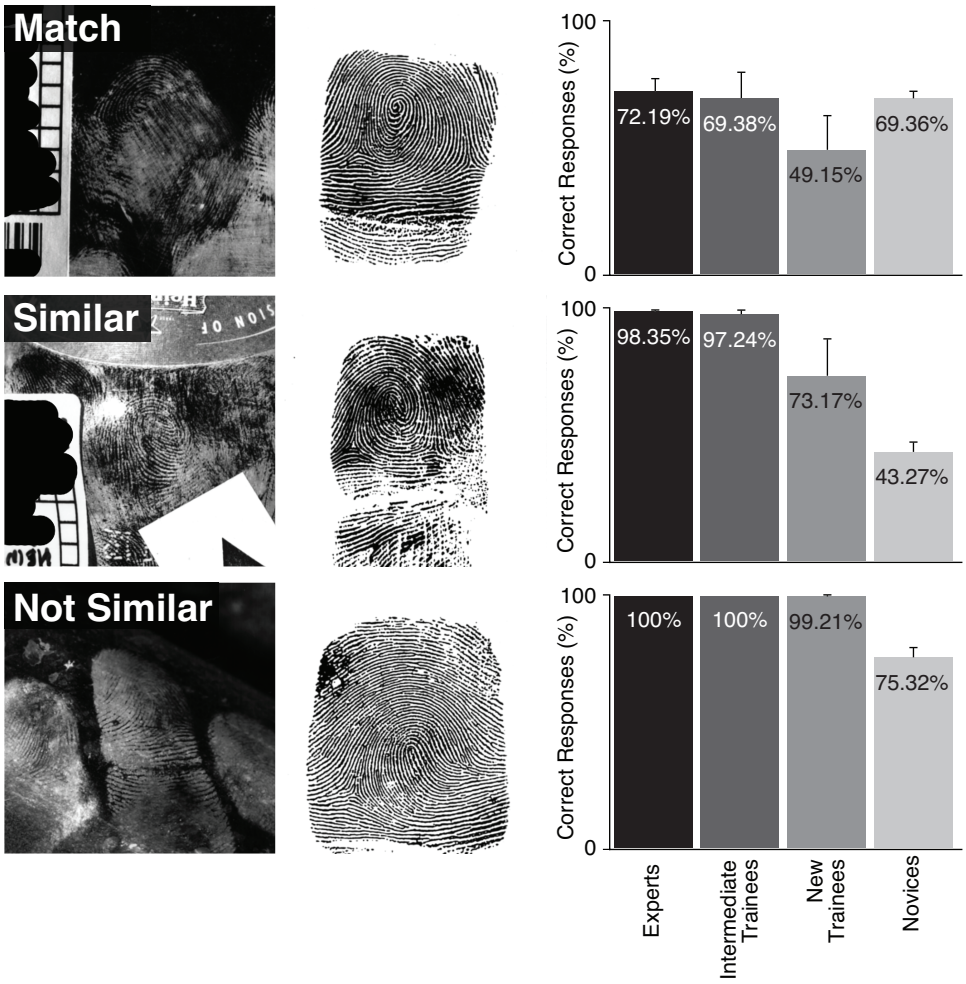

Figure 4.3: Stimuli and results. On each trial, participants were presented with a genuine crime scene latent print on the left and a fully rolled candidate print on the right, and they were asked to judge whether the prints in each pair matched using a confidence rating scale. On some trials, the two prints came from the same individual (top row); on others, the prints were similar but came from two different individuals (middle row); and on others, the prints came from two different individuals and were paired randomly (bottom row). The three graphs on the right depict the mean percentage of correct responses in these three conditions for experts, intermediate trainees, new trainees, and novices. Error bars represent 95% within-cell confidence intervals.

all trials containing the latent (even the unaffected similar and nonsimilar distractor trials) were removed from the analysis. Because the latents were randomly allocated to either a target, similar or nonsimilar distractor pair, the proportion of trials removed was randomly distributed across trial types. With the offending latent removed there were 44 fingerprint comparison data points from each participant, rather than 45. For the 37 experts, for example, there were 537 matching trials, 547 similar nonmatching trials, and 544 nonsimilar nonmatching trials, rather than 555 per condition. For each participant, we calculated the percentage of trials that were responded to correctly in each condition. The three graphs on the right side of Figure 4.3 depict the average percentage of correct responses for the 37 experts, 10 intermediate trainees, nine new trainees, and 37 novices. Participants were anonymous, so it is not possible to link a participant's performance to a particular person or police agency. Matching trials included pairs of prints that originated from the same source. We use the term match here as shorthand, but as indicated above, the ground truth of the print pairs is uncertain. As depicted in Figure 4.3, experts correctly labeled 72.19% (SD = 18.10%) of the matching pairs on average "match" (hits), but incorrectly labeled 27.81% of the matching pairs "no match" (misses). Intermediate trainees correctly labeled 69.38% (SD = 17.34%) of the matching pairs "match" (hits), but incorrectly labeled 30.62% of them "no match" (misses). New trainees correctly labeled 49.15% (SD = 21.66%) of these matching pairs "match" (hits), but incorrectly labeled 50.85% "no match" (misses). Novices correctly labeled 69.36% (SD = 13.02%) of these matching pairs "match" (hits), but incorrectly labeled 30.64% "no match" (misses).

Highly similar nonmatching trials included pairs of prints that did not originate from the same source but are, according to the national database search algorithm, highly similar. We use the term similar nonmatch here as shorthand. Experts correctly labeled 98.35% (SD = 4.01%) of the highly similar nonmatching pairs "no match" (correct rejections), but incorrectly reported 1.65% of them "match" (false alarms) on average. Specifically, seven experts incorrectly labeled nine pairs out of the 547 highly similar nonmatching pairs a "match"—six experts made one false alarm each and one expert made three false alarms. Confidence ratings for the nine false alarms were 8, 9, 9, 10, 12, and 12 for the six experts with one false alarm each and 8, 9, 9, for the one expert with three false alarms. The nine

false alarm errors occurred on eight different print pairs (i.e., each false alarm was made on a different latent and similar exemplar pair except for one). The fact that the nine false alarms were spread across prints and across people suggests that human factors are likely to be good predictors of these errors, rather than factors in the prints themslves. We caution readers, however, to avoid over interpreting individual confidence reports because the appropriate level of analysis in this experiment is expertise, not the individual confidence ratings by individual examiners on individual trials. Intermediate trainees correctly labeled 97.24% ($SD = 4.91\%$) of these pairs "no match" (correct rejections), but incorrectly labeled 2.76% "match" (false alarms). Specifically, four intermediate trainees made four false alarms on four different print pairs. New trainees correctly labeled 73.17% ($SD = 23.22\%$) of these pairs "no match" (correct rejections), but incorrectly labeled 26.83% "match" (false alarms). Novices correctly labeled 43.27% ($SD = 14.79\%$) of these pairs "no match" (correct rejections), but incorrectly labeled 56.73% "match" (false alarms). Most striking is the difference in the rate of false alarms between experts and novices: 1.65% for experts compared with 56.73% for novices. Nonsimilar nonmatching trials included pairs of prints that did not originate from the same source and were sampled randomly from the set. We use the term nonmatch here as shorthand. Of the trials in which the prints did not match, and were not similar, both experts and intermediate trainees correctly labeled 100% of these pairs "no match" (correct rejections), and so they did not incorrectly label any pairs (false alarms). New trainees correctly labeled 99.21% ($SD = 2.39\%$) of these pairs on average "nonmatch," but incorrectly labeled 0.79% "match." Novices correctly labeled 75.32% ($SD = 15.48\%$) of these pairs on average "nonmatch," but incorrectly labeled 24.68% "match." Experts and intermediate trainees responded much more toward the extreme ends of the confidence scale compared with new trainees and novices: 83% of expert and 75% of intermediate responses were either one or 12 compared with 53% for new trainees and 20% for novices. We subjected the percentages of correct responses to a 4 (expertise: experts, intermediate trainees, new trainees, novices) $\times$ 3 (trial type: match, similar nonmatch, nonsimilar nonmatch) mixed analysis of variance (ANOVA). The analysis revealed significant main effects of expertise, $F(3, 89) = 109.450$, $MSE = 0.014$, $p = .001$, $\eta^2 = .79$, 95% CI [.71, .82], and trial type, $F(2, 178) = 66.038$, MSE

= .019, p = .001, η^2 = .43, 95% CI [.33, .50], and a significant interaction between the two, $F(6, 178) = 29.385$, MSE = .019, p = .001, η^2 = .50, 95% CI [.40, .55].

Simple effects analyses revealed a significant benefit of expertise on all trial types: match, $F(3, 89) = 4.759$, MSE = .027, p = .004, η^2 = .14, 95% CI [.03, .23], similar nonmatch, $F(3, 89) = 142.391$, MSE = .015, p = .001, η^2 = .83, 95% CI [.77, .86], and nonsimilar nonmatch, $F(3, 89) = 45.999$, MSE = .010, p = .001, η^2 = .61, 95% CI [.49, .67]. Follow-up pairwise comparisons revealed that, for matches, only new trainees were different from all other levels of expertise: new trainees versus novices, p = .001, d = 1.13, 95% CI [32.0, 8.0]; new trainees versus intermediate trainees, p = .009, d = 1.03, 95% CI [35.0, 5.1]; new trainees versus experts, p = .001, d = 1.15, 95% CI [35.0, 11.0]. For similar nonmatches, both novices and new trainees were different from all other levels of expertise: novices versus new trainees, p = .001, d = 1.54, 95% CI [38.8, 21.0]; novices versus intermediate trainees, p = .001, d = 4.89, 95% CI [62.5, 45.4]; novices versus experts, p = .001, d = 5.08, 95% CI [60.7, 49.5]; new trainees versus intermediate trainees, p = .001, d = 1.43, 95% CI [35.1, 13.0]; new trainees versus experts, p = .001, d = 1.51, 95% CI [34.1, 16.3]. For nonsimilar nonmatches, only novices were different from all other levels of expertise: novices versus new trainees, p = .001, d = 2.16, 95% CI [32.2, 16.6]; novices versus intermediate trainees, p = .001, d = 2.25, 95% CI [31.7, 17.7]; novices versus experts, p = .001, d = 2.25, 95% CI [29.2, 20.1].

4.7 Discussion

We set out to determine whether fingerprint experts are any more accurate at matching prints than trainees and lay people. We also wanted to get an idea of how often these groups make "misses" (i.e., errors comparable with allowing a guilty person to escape detection) compared with how often they make "false alarms" (i.e., errors comparable with falsely incriminating an innocent person). In this experiment, we made use of genuine crime scene prints that are highly representative of casework, but where the ground truth is uncertain. We found that experts and novices were equally accurate at identifying print pairs that actually matched; both groups were around 70% accurate. Experts, however, were much more accurate than novices at identifying prints that did not actually match, but were highly similar; experts were

98.35% accurate compared with 43.27% for novices. It seems that superior expert performance lies, not in the ability to match prints per se, but in the ability to identify highly similar, but nonmatching, prints as such. The comparison with novices is important for demonstrating expertise, and shows that the discrimination task was difficult enough for experts to perform accurately, but for novices to perform relatively poorly. The results of this experiment are similar to those reported by Tangen et al. (2011). It is possible to compare performance across these two experiments because it is only the stimulus sets that differ—simulated crime scene latents were used by Tangen et al. and genuine crime scene latents were used in this experiment. The performance difference between experts and novices for similar nonmatches was about the same in both experiments; experts were around 55% more accurate than novices in both experiments. The performance difference between experts and novices for matches, however, was different in the two experiments. In Tangen et al., experts were around 18% more accurate than novices for prints that matched. In the present experiment, experts and novices were equally accurate for prints that matched. It appears that experts are even more conservative in their decision making when dealing with genuine crime scene prints than when dealing with the simulated crime scene prints: 72% hits on matching genuine latents compared with 92% hits on matching simulated latents. The performance of trainees, in the present experiment, was more nuanced. For print pairs that matched, intermediates were just as accurate as experts, although new trainees were less accurate than any other group. For the similar print pairs, experts and intermediates were equally accurate, although new trainees were less accurate than both experts and intermediates, but they were more accurate than novices. There was very little difference between the overall performance of experts with an average of 17.5 years of experience and intermediate trainees with an average of 3.5 years of experience. Although these results provide some insight into the development of expertise (i.e., how long it takes to turn a novice into an expert), much more research needs to be conducted. For example, one could track the development of novice examiners over time to determine precisely what aspects of their performance change as novices become experts and how quickly these capabilities develop. Future experiments could also pinpoint the nature of this expertise. That is, the relative contribution of formal rules compared with the accumulation of experience (Norman et al., 2007; Norman & Brooks, 1997), as well

as the role of corrective feedback (Eva & Regehr, 2005). In addition to the development of expertise, we need to better understand how age affects identification separately from years of experience. In medicine, for example, older/more experienced doctors generally have greater diagnostic accuracy (Eva, 2002), but are less likely to be influenced by the presentation of clinical features that are inconsistent with their initial hypothesis (Eva, Link, Lutfey, & McKinlay, 2010). Similar analyses need to be conducted in forensic reasoning to establish the relationship between age, experience, and fingerprint matching performance. This experiment was not designed to determine the likelihood of errors in practice, nor the performance of individuals or departments, and examiners were not provided with their usual tools or independent verification. It was designed to determine performance differences based on expertise using genuine crime scene latents. Inferring from these results that experts are 98.35% accurate in practice or that the overall error rate of fingerprint identification is 1.65%, would be unjustified. It may be necessary, or the courts may demand, that particular rates of error are established for particular situations. At the extreme, an examiner could report how accurate they are at matching an arch type print, lifted from glass, using white powder, in a particular department, with particular training, on a Tuesday, and so on. But unless it has been demonstrated that accuracy (or proficiency, or reliability, or competence) varies systematically in any one of those situations, then it may be best to report measures of accuracy at a broader level (Koehler, 2008; M. B. Thompson et al., 2013a).

4.7.1 Discrimination and Response Bias

In describing how well someone performs a given task, people usually count the number of correct items relative to the total number of items in the task. But in our experiment, when an examiner compares two fingerprints, there are two ways to be right and two ways to be wrong. To get a comparison right, as shown in Figure 4.2, one can correctly say the prints are from the same source when they actually are (a hit), or correctly say the prints are not from the same source when they actually are not (a correct rejection). To get a comparison wrong, one can incorrectly say that the prints are from the same source when they actually are not (a false alarm), or incorrectly say the prints are not from the same source when they actually

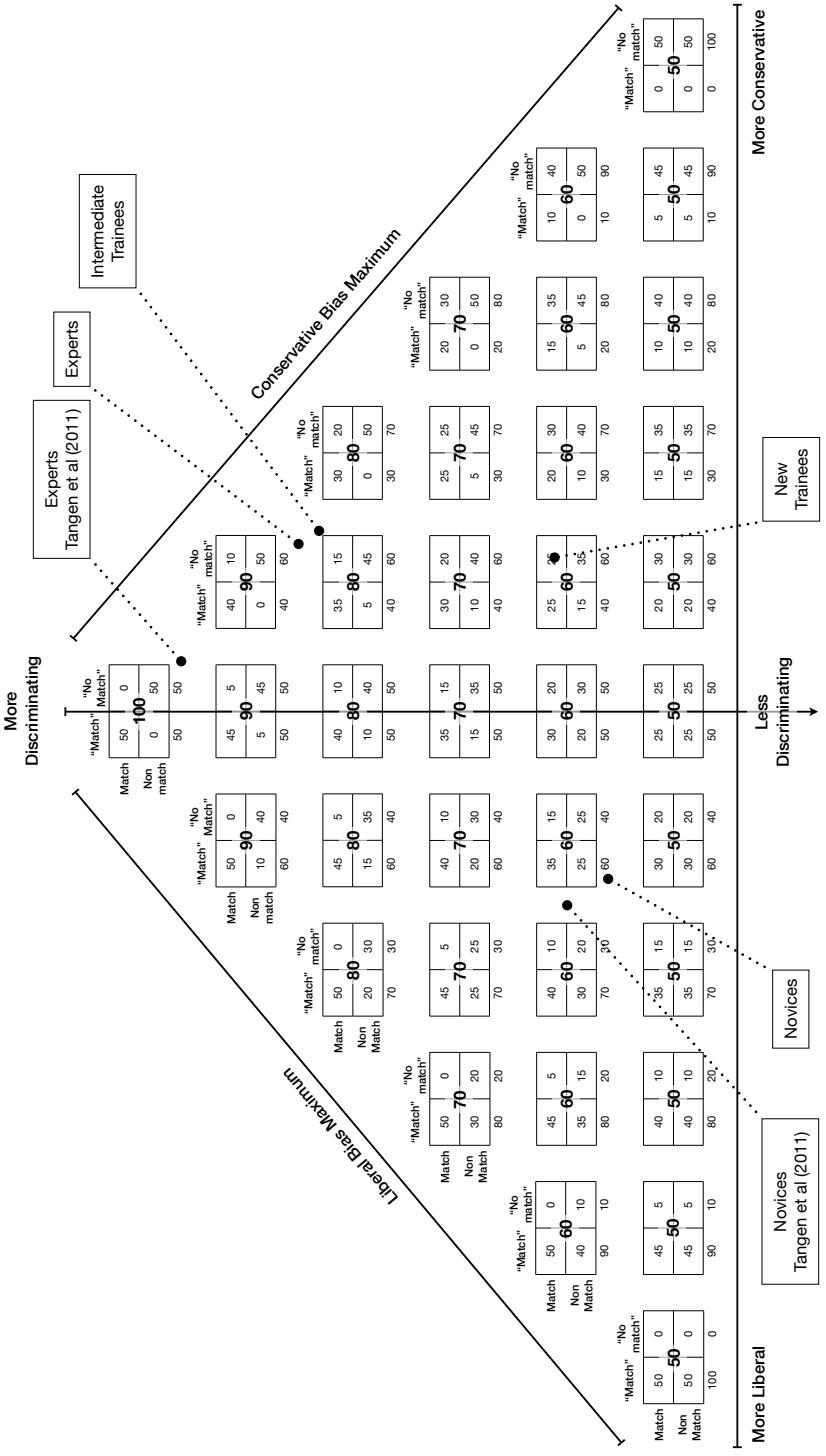

Figure 4.4: The space represents all possible performance results from a fingerprint discrimination task and the relationship between discrimination and response bias. Pinpointed in the space are the locations of the actual results for each of the six groups from the experiments in Chapter 2 and 4, with nonsimilar nonmatches omitted and the number of trials scaled to give a total of 100. Each filled circle represents the center of the 2×2 contingency table based on the data from each of the conditions.

are (a miss). Therefore, simply counting up, say, just the number of similar, but nonmatching prints that experts correctly judged to be "no match" (i.e., 98.35%) is only half of the story. These examiners could have scored 100% correct on these nonmatching prints by simply saying "no match" to every pair of prints. By adopting such a conservative response bias, however, they would have incorrectly deemed every pair of matching prints a "no match" as well. On the other hand, they could adopt an extremely liberal response bias and say "match" to every pair of prints. These examiners would score 100% correct for all the prints that actually match, but they would incorrectly declare every nonmatching pair a "match" as well. The only way to perform perfectly in this experiment is to adopt a neutral response bias and correctly label all of the matching pairs "match" and label all of the nonmatching pairs "no match." By adopting a signal detection methodology, we can distinguish people's tendency to say "match" or "no match" from their ability to distinguish prints that actually match from those that actually do not match—we can separate an examiner's response bias from their ability to discriminate matching and nonmatching fingerprints (Greene & Oliva, 2009; Phillips et al., 2001). The results from our experiment indicate that experts and intermediate trainees both have a tendency to say "no match" regardless of whether the prints actually match or not. Adopting such a strong conservative response bias certainly reduces the rate of false alarm errors (i.e., errors that could lead to falsely incriminating an innocent person), but it will also necessarily increase the rate of miss errors (i.e., errors that could lead to a guilty person escaping detection). A false alarm rate of 1%–3% is indeed impressive, particularly compared with a 57% false alarm rate for novices. But there is a direct tradeoff between preventing a false alarm and allowing a miss. The cost of such a low false alarm rate for experts and intermediate trainees in the current experiment equates to a substantial miss rate of roughly 30%. The relationship between discrimination and response bias for each of the four groups in the above experiment, and the novices and experts in Tangen et al. (2011), is depicted in Figure 4.4. The figure represents the space of all possible results from experiments like ours. Each of the tables that comprise Figure 4.4 is a version of the contingency table in Figure 4.2 with different combinations of average "match" and "no match" responses from participants when we ask them to compare 50 print pairs that actually match and 50 prints pairs that don't actually match. Moving along the y-axis from the bottom to

the top of the figure, participants become more capable of discriminating matching and nonmatching prints. That is, they correctly say "match" to matching prints and "no match" to nonmatching prints, thereby increasing the values in the top left cell (hits) and bottom right cell (correct rejections) in each of the tables that comprise Figure 4.4. The table at the apex depicts perfect discrimination—50 hits and 50 correct rejections. Participants here can distinguish between matching and nonmatching prints perfectly. The tables along the bottom depict chance discrimination. Participants here cannot distinguish between matching and nonmatching prints (their performance is like a coin flip), but there are several ways to reach the same level of overall performance. Overall performance—the number of comparisons they got correct—is depicted by the large number in bold at the center of each table and is the sum of the two diagonal cells (hits and correct rejections). Overall performance ranges from 50 (chance discrimination) at the bottom of the figure to 100 (perfect discrimination) at the top. Results from novices and new trainees lie toward the bottom of this figure—they are reasonably poor discriminators—compared with intermediate trainees and experts, who are closer to the top. Moving along the x-axis from the left to the right, participants become more conservative in their responses; on the left side of Figure 4.4, they say "match" much more often than they say "no match," regardless of whether the prints actually match or not. The opposite is true on the right side of the figure; they say "no match" much more often than they say "match." A liberal response bias (on the left of the figure) is represented by a higher column total for the two cells on the left side of each table (i.e., a tendency to say "match"), compared with the two cells on the right. A conservative response bias (on the right of the figure) is represented by a higher column total for the two cells on the right side of each table (i.e., a tendency to say "no match"), compared with the two cells on the left. An extremely liberal response bias, coupled with low accuracy, means that participants say "match" to every comparison. They will get half of the comparisons correct in this case, but they will also get half incorrect; they make many hits and many false alarms, while not making any misses or correct rejections. An extremely conservative response bias, on the other hand, coupled with low accuracy, means that participants say "no match" to every comparison. Again, they will get half of the comparisons correct in this case, but they will also get half incorrect; they make many misses and many correct rejections, while not making any hits or

false alarms. Results from novices lie closer to the left of the figure—they have a reasonably liberal response bias—compared with trainees and experts, who lie closer to the right and have a very conservative response bias. Theoretically, there is an optimal decision criterion, that minimizes errors, where the participant shows no response bias and is equally likely to say "match" or "no match" across all comparisons (i.e., straight up and down the middle of Figure 4.4 where the row totals are equal). This is true only when the base rates—the signal and noise distributions—are equal (i.e., the column totals are equal), as in Figure 4.4. The picture changes dramatically, however, when the base rates are unequal, which would add a third dimension to Figure 4.4. For example, if there are many more matches than there are nonmatches—as may be the case in practice when examiners compare crime scene latent prints to suspects already in custody—a liberal response bias would result in a high number of hits and a low number of false alarms. If, on the other hand, there are many more nonmatches than there are matches—as may be the case in practice when examiners search large databases in the absence of a suspect—a liberal response bias would result in a high number of false alarms and a low number of hits. The decision criterion (i.e., the propensity to say "match" or "no match") that a search algorithm, examiner, or department adopts will depend on the real world costs and benefits. Policy decisions about the ideal decision criterion and subsequent response bias will be ideological, not empirical, in nature. Interventions in training, technology, management, safety culture, and public policy will influence the signal and noise distributions, and the ratio of errors (false alarms vs. misses) that examiners will make in practice (Clark, 2012; Wixted & Mickes, 2012).

4.8 Conclusions

We found that qualified, court-practicing fingerprint experts were exceedingly accurate at discriminating prints compared with novices. Our experts showed a conservative response bias, tending to err on the side of caution by making more errors of the sort that could allow a guilty person to escape detection than errors of the sort that could falsely incriminate an innocent person. The performance difference between experts and novices provides further evidence for expertise in fingerprint identification. How novices would perform under different

levels of motivation and incentive, after brief training, after the costs of a false alarm versus a miss are conveyed, and so forth, is still unknown. The superior performance of experts in this experiment was not simply a function of their ability to match prints, per se, but a result of their ability to identify highly similar, but nonmatching fingerprints as such. Novices, nonetheless, correctly identified almost the same number of matching prints as experts. This experiment was designed to be difficult. The fact that experts made so few errors is evidence for impressive human pattern matching performance possibly exceeding that of experts in other comparable domains of expertise (Thompson et al., 2013). When these results are compared with those of Tangen et al., (2011) we see that experts were even more conservative in their decision making when dealing with genuine crime scene prints than when dealing with simulated crime scene prints, and this conservatism made them relatively less accurate overall. How experts would perform with their usual tools, peer verification, statistical models, different lifting agents and surface types, different response types, time and resource constraints, different types of training, experience, and qualifications, and so forth, is unknown. The performance of the trainee groups was surprising. First, intermediate trainees—despite their lack of qualification and average 3.5 years experience—performed about as accurately as qualified experts who had an average 17.5 years experience. This finding provides some insight into the development of expertise, that is, how long it takes to turn a novice into an expert. It appears that people can learn to distinguish matching from similar nonmatching prints to roughly the same level of accuracy as experts after a few years of experience and training. Much more research needs to be conducted, however, to make precise and definitive conclusions about the factors that lead to fingerprint matching expertise. Second, new trainees—despite their 5-week, full-time training course or their 6 months of experience—were not any better than novices at discriminating matching and similar nonmatching prints, they were just more conservative (see Figure 4.4). It appears that early training and/or experience may not necessarily result in more accurate judgments, but may simply result in a more conservative response bias (i.e., a tendency to say "no match" more often). This experiment was limited by the small number of trainee participants, so one needs to be cautious when interpreting the relative performance of these groups. Small sample sizes in trainee comparison groups will be difficult to overcome considering

that we recruited the majority that existed across Australia. Also, more trials (i.e., the number of fingerprint comparisons) per participant across conditions would help bear out false alarm errors in order to understand their nature. Given that appropriate casework stimuli are so rare and manufacturing stimuli is difficult and expensive, future experiments could attempt to pull expert performance off ceiling by adding artificial noise or constraining the task environment. Measuring the relative performance of trainees is a useful first step, but programmatic or longitudinal experiments are needed to answer questions such as: What sets an expert apart from a novice? How does fingerprint expertise develop over time? Does training help and can training time be reduced without compromising performance? What is the best way to provide feedback to examiners about their performance? More research is needed to determine the nature of forensic reasoning, the influence of deadline pressure (Brewer, Weber, Wootton, & Lindsay, 2012), the role of feedback and self-assessment (Eva & Regehr, 2005), and, more generally, the respective contribution of training (the formal rules) versus daily exposure to a multitude of prints (the accumulation of instances; Norman et al., 2007). The findings of this representative, lab-based experiment may have implications for the way fingerprint examiners testify in court (Edmond, Thompson, & Tangen, 2013). What the findings mean for understanding expert performance in the wild is an open, empirical, and epistemological question that is part of an ongoing conversation (e.g., Koehler, 2012, 2008; M. B. Thompson et al., 2013a). Whether performance data come from lab-based experiments, statistical models, proficiency tests, or full-scale black box interrogations of a system, one still needs to make an inference to performance in a particular manifestation of practice or to reason about the value of the evidence in a particular case. Edmond, Thompson, & Tangen (2013) have proposed a guide for the reporting of emerging empirical data about the performance of fingerprint examiners in order to help nonexperts understand the value of fingerprint evidence. The question, "What is the error rate in practice?" may not be the right one. Better questions might be: What is known about expert performance in situations similar to practice or to the particular case? What can reasonably be inferred from the general (data from experiments like this one) to the particular (the evidence in the case)? What information, and in what form, will help a trier of fact make optimal decisions? Taking the lead from research in health care and aviation, having empirical evidence from several

experiments that address various research questions, at multiple levels of analysis, is sure to be the best way to help researchers reason about performance in the wild and to help triers of fact reason about forensic evidence.

Chapter 5

A Novel Representation of Sensitivity and Response Bias for Signal Detection Analysis

5.1 Preface

This chapter is unpublished, but will eventually be submitted for publication a research methods journal. As can be seen in Figure 5.1, this is the first and only chapter of PART 2 - DEPICTING EXPERTISE. It is very much my own work, with Jason Tangen contributing 20% to the design of the contingency space, and 10% to contrasting with traditional depictions.

When I was presenting the results of Chapters 2 and 4 to colleagues at conferences and examiners at police stations, I found that simple accuracy bar graphs were not properly depicting the highly conservative response bias of fingerprint experts, or the conservative but inaccurate performance of new trainee examiners. Police examiners, especially, were focusing on the false alarm rate as the sole indicator of performance, and I found it difficult to communicate the independence of accuracy and bias in a Signal Detection Theory framework. I then tried presenting the results of each condition as a set of contingency tables, but the relationship accuracy and bias relationships *between* the conditions was still unclear. I realised that the results of each condition would need to be represented in a space of all possible ways

in which the experiment could have turned out. I then presented the results from Chapters 2 and 4 at a recent conference using this contingency space representation, and several people encouraged me to publish the contingency space representation in a research methods journal. I think this representation could be useful for communicating signal detection data to people who don't have a background in statistics and research methods, and even for fellow researchers as a compliment to traditional representations.

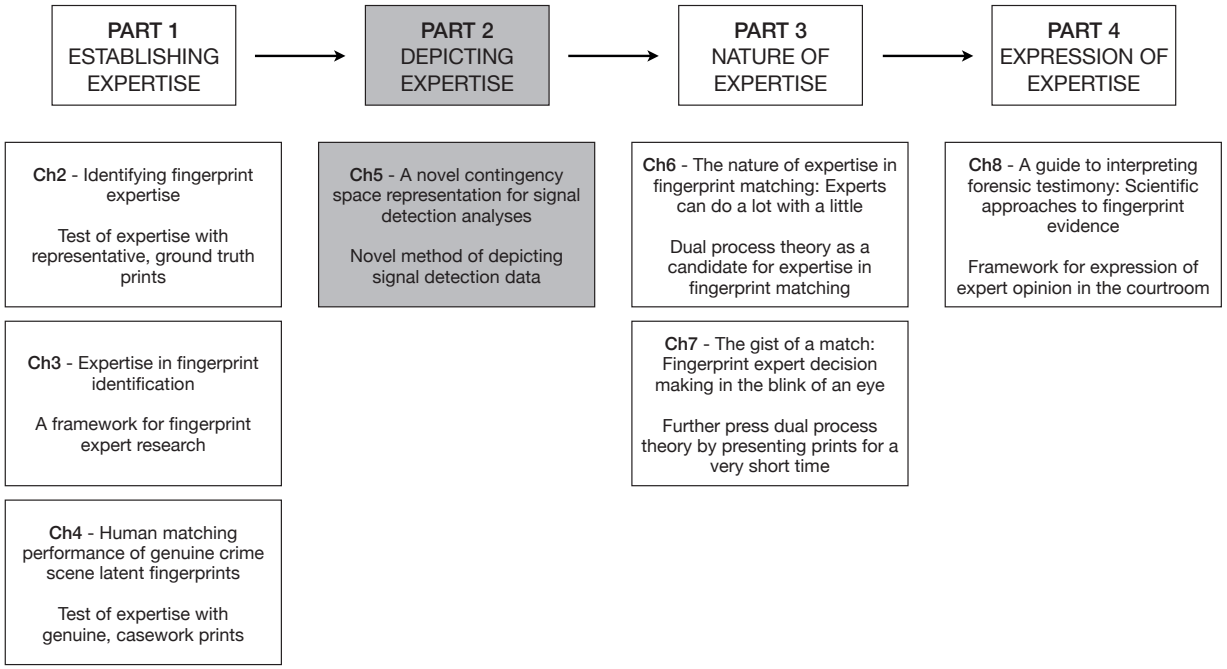

Figure 5.1: Conceptual diagram highlighting Chapter 5, Part 2 of the thesis: "A novel contingency space representation for signal detection analyses."

5.2 Introduction

Data from perception, decision making, and recognition memory experiments are commonly analysed using Signal Detection Theory (SDT). Signal detection is a method of quantitating a person's (or system's) ability to distinguish signal from noise. Experiments consist of target and distractor stimuli and because people are asked to make a decision about the stimuli, they must adopt a criterion upon which to report the stimuli as a target or a distractor

(Higham, Perfect, & Bruno, 2009). Signal detection was initially applied to radar operators who were trying to discriminate friendly and enemy aircraft and has since been used to measure several areas of human performance (Green & Swets, 1966). It allows us to separate the two ways of being right and the two ways of being wrong (Phillips et al., 2001). In SDT, the rate of hits and the rate of false alarms can be used to calculate an examiner's accuracy (or discrimination)—that is, their ability to tell the difference between two prints that came from the same source and two prints that came from a different source. The hit and false alarm rates can also be used to calculate the examiner's response bias—that is, their tendency to respond "match" or "no match" regardless of whether the prints actually match or not.

There are several, well-established, ways to describe signal detection data numerically and pictorially. For example, sensitivity indices (e.g., d' and a') and response bias indices (e.g., c and β), and pictorial representations (e.g., detection error tradeoff graphs and confidence-based receiver operator characteristics). I suggest that these representations can be difficult to interpret for those not well versed in SDT, and that there may be a complementary representation that better depicts the relationship between sensitivity and response bias.

5.3 Signal Detection in Fingerprint Matching

A decision about whether two fingerprints match or not is based on the judgment of a human examiner, not a computer. In the experiments in Chapter 2 and 4, I set out to determine whether fingerprint experts are any more accurate at matching prints than the average person, to explore the development of matching expertise, and to get an idea of how often experts make errors of the sort that could allow a guilty person to escape detection compared with how often they make errors of the sort that could falsely incriminate an innocent person. In Chapter 2, I asked experts and novices to compare fingerprints in a two-alternative forced choice design, and in Chapter 4, I added two trainee groups—new trainees and intermediate trainees. Here I will present the results from these experiments in Receiver Operator Characteristics and a bias plots.

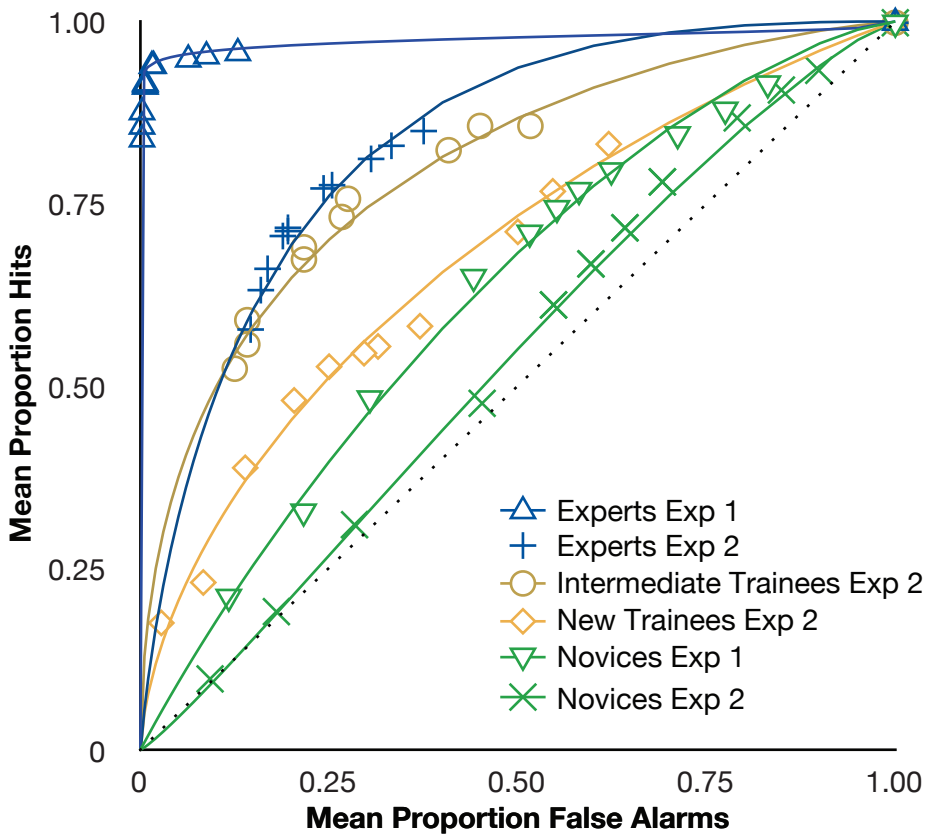

Figure 5.2: Receiver Operator Characteristics of the the six groups from the experiments in Chapters 2 (Exp 1) and 4 (Exp 2).

5.3.1 Receiver Operator Characteristics

A Receiver Operator Characteristic (ROC) curve plots the hit rate against the false alarm rate depending on different levels of confidence. Figure 5.2 shows a set of 12-point ROC plotted from the 12-point confidence scale of my experiments. In general, the extent to which the curve bows from the major diagonal depicts accuracy, with more bowing indicating higher accuracy (greater discrimination). The shape of the ROC depicts the response bias. Figure 5.2 shows the ROC curves based on data of the six groups from the experiments in Chapters 3 and 5.

For readers who are experienced in interpreting ROCs, it may be relatively easy to get a feel for the relative accuracy of experts, trainees, and novices across the two experiments. I contend, however, that interpreting ROCs is difficult for those without such experience.

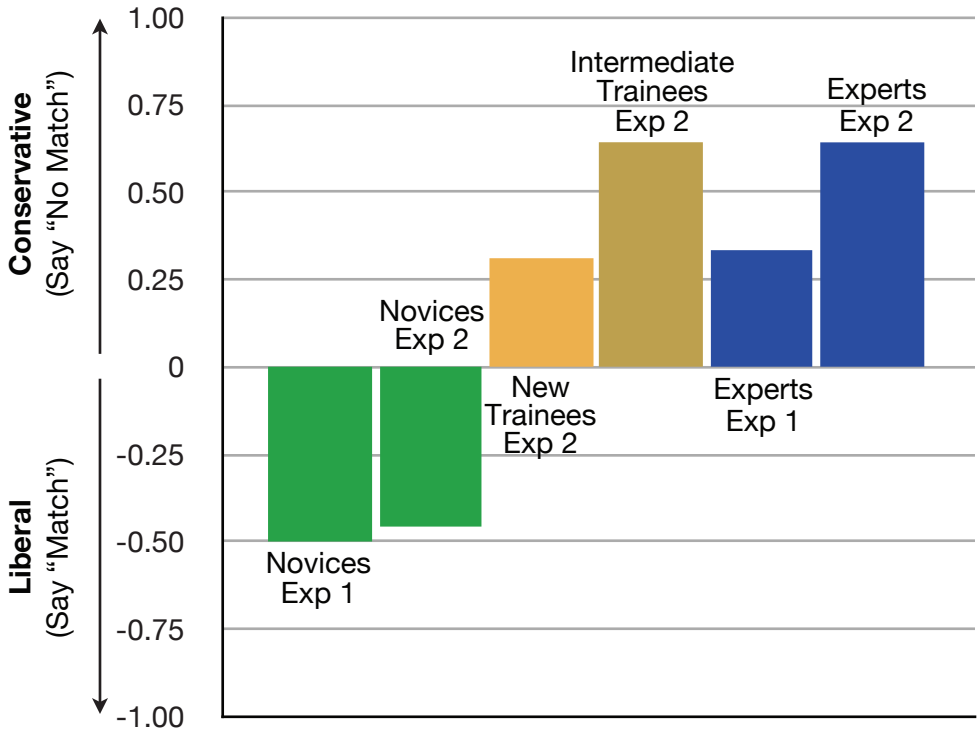

Figure 5.3: Average response bias (B''_D) values for each of the six groups. B''_D varies from -1.0 to +1.0, with positive numbers indicating a bias to respond "No Match," negative numbers indicating a bias to respond "Match," and 0.0 indicating no bias.

Those without experience are often the very people for whom the results of experiments like these are of particular relevance—fingerprint examiners, forensic managers, judges, lawyers, and jurors. (This difficulty is partly why I opted to depict the results from the experiments in Chapters 3 and 5 as percent correct for each of the trial types, rather than ROCs.) I also contend that getting a feel for the response bias of the groups—and the relationship between the biases of the groups—is difficult, even for those with training in signal detection theory. A possible solution is to create a separate response bias plot for each group.

5.3.2 Bias

Figure 5.3 shows the response bias—B''_D in this case—for each of the six groups. B''_D is a convenient statistic because it varies from -1.0 to $+1.0$, with positive numbers indicating a bias to respond "No Match," negative numbers indicating a bias to respond "Match," and 0.0

indicating no bias (see Donaldson, 1992, for discussion). In Figure 5.3 we can see that novices have a liberal response bias and trainees and experts have a conservative response bias. We can also see the bias relationship between the groups. The downside is that we cannot see the accuracy of the groups without switching between the bias figure and the ROC figure. A potential solution is to create a single figure that depicts accuracy and response bias, which allows us to depict the relationship between the groups on these two performance measures.

5.4 Contingency Space Representation

Here I describe a novel method of depicting signal detection data. I suggest that the method makes the theoretical independence between discriminability and response bias clearer, it is especially useful for comparing results between experimental conditions, and the use of natural frequencies makes the results easier to interpret. This contingency space representation of signal detection data could complement traditional representations. I first describe contingency tables, then the contingency space and then how to interpret sensitivity and response bias.

5.4.1 Contingency Table

Panel A of Figure 5.4 shows a 2 × 2 contingency table depicting the four possible outcomes of a forced choice discrimination task where the actual value (the ground truth) is true or false and where a person responds "True" or "False." (Ground truth is labeled on the left, rather than the top, which is traditional, for reasons that will become clear.) The four possible outcomes are a hit (true positive), a miss (false negative), a correct rejection (true negative), or a false alarm (false positive). Panel B of Figure 5.4 shows a similar table with with numbers representing the results of an experiment where a fingerprint examiner was asked to make a "match" or "no match" decision about 100 fingerprint pairs (50 from the same source and and 50 from a different source—equal base rates). In this example, the examiner performed perfectly (perfect discrimination) by correctly labeling all of the same source prints as matching and all of the different source prints as not matching. The number

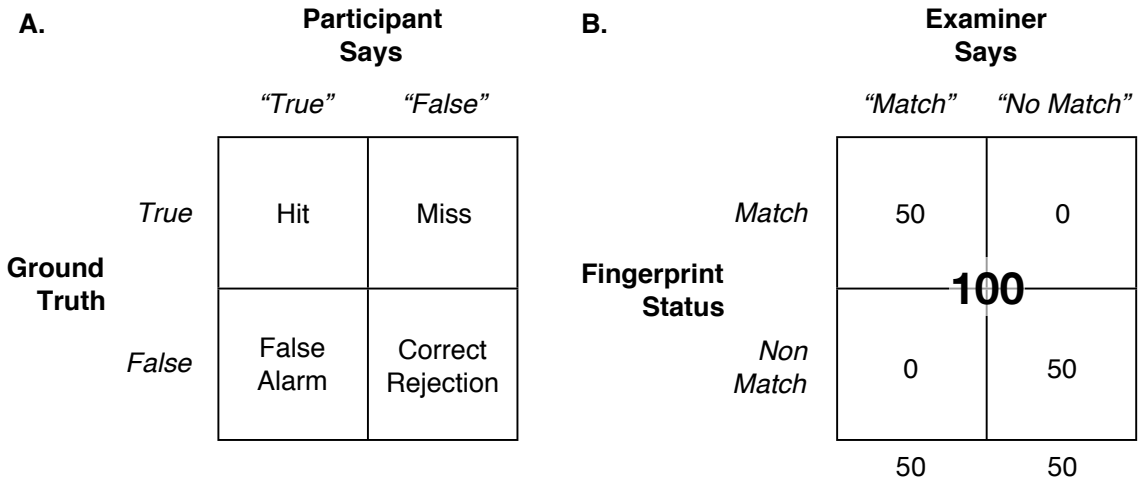

Figure 5.4: Panel A shows a 2×2 contingency table depicting the four possible outcomes of a forced choice discrimination task where the ground truth of the stimuli is either true or false and an agent reports the stimuli as "True" or "False". Panel B shows a 2×2 contingency table depicting the four possible outcomes of a forced choice discrimination task where two prints match or not and an examiner declares them as a "match" or "no match". The numbers align with hits, false alarms, misses, and correct rejections in Panel B. The large number in bold at the center of each table depicts the sum of the two diagonal cells ranging from 50 (chance discrimination) at the bottom of the figure to 100 (perfect discrimination) at the top. The column totals at the bottom of each table depict response bias with liberalism represented by a higher column total for the two cells on the left side of each table and conservatism represented by a higher column total for the two cells on the right side of each table.

in bold at the center of the table depicts the sum of the two diagonal cells (hits and correct rejections), and ranges from 0 (defective discrimination) to 50 (chance discrimination) to 100 (perfect discrimination). The column totals at the bottom of the table depict response bias, where liberalism would be represented by a higher column total for the two cells on the left side of each table and conservatism represented by a higher column total for the two cells on the right side of each table. In this case, the response bias is neutral because the column totals are equal.

5.4.2 Contingency Space

Figure 5.5 is a contingency space representation of all possible results from experiments employing signal detection. Each of the tables that comprise Figure 5.5 is a version of the

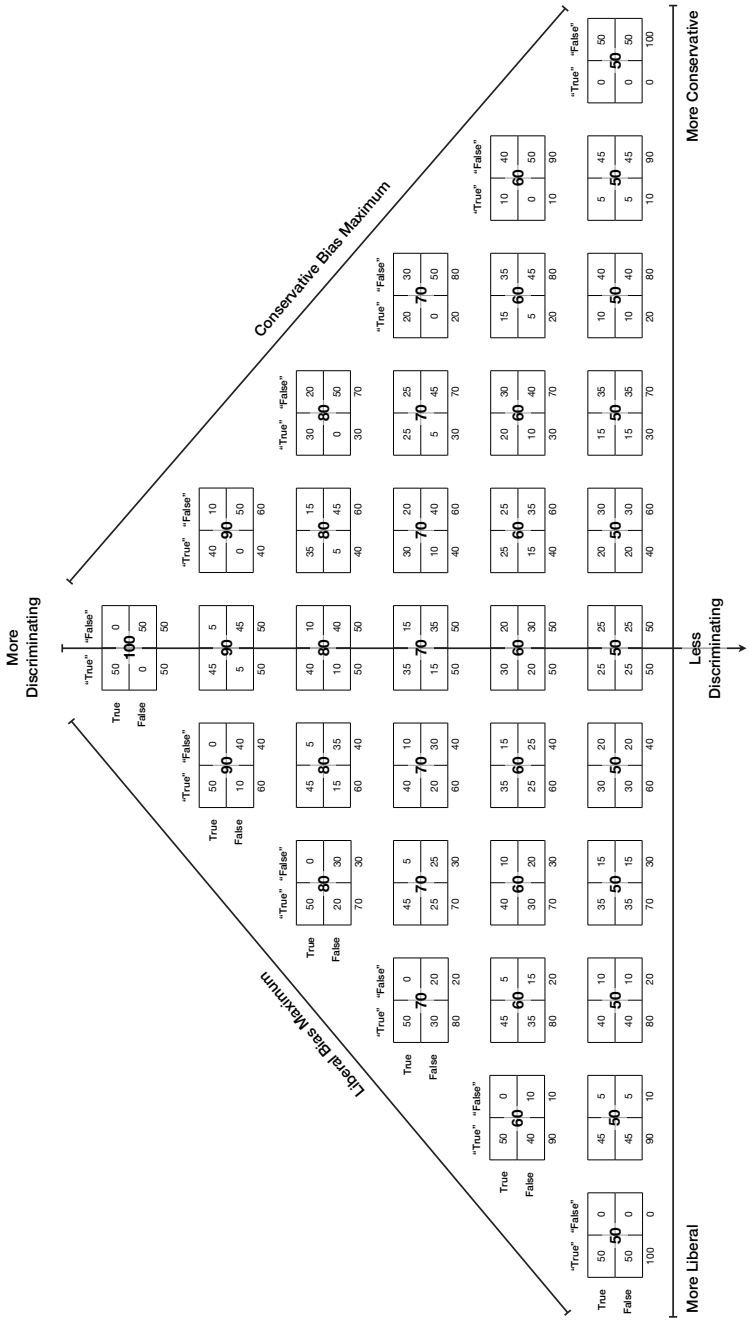

Figure 5.5: The space represents all possible performance results from a discrimination task and the relationship between discrimination and response bias. Each of the tables that comprise the figure is a 2×2 contingency table depicting the four possible outcomes of a forced choice discrimination task where the ground truth of the stimuli is either true or false and an agent reports the stimuli as "True" or "False". The numbers align with hits, false alarms, misses, and correct rejections in Figure 5.4. The large number in bold at the center of each table depicts the sum of the two diagonal cells ranging from 50 (chance discrimination) at the bottom of the figure to 100 (perfect discrimination) at the top. The column totals at the bottom of each table depict response bias with liberalism represented by a higher column total for the two cells on the left side of each table and conservatism represented by a higher column total for the two cells on the right side of each table.

contingency table in Figure 5.4A with different combinations of average "True" and "False" responses from participants when they were asked to make a judgment about a set of stimuli.

5.4.3 Discrimination (Sensitivity)

Moving along the y-axis from the bottom to the top of Figure 5.5, participants become more capable of discriminating true and false stimuli. That is, they correctly say "true" to true trials and "false" to false trials, thereby increasing the values in the top left cell (hits) and bottom right cell (correct rejections) in each of the tables that comprise Figure 5.4B. The table at the apex depicts perfect discrimination—50 hits and 50 correct rejections. Participants here can distinguish between true and false stimuli perfectly. The tables along the bottom depict "chance" discrimination. Participants here cannot distinguish between true and false stimuli (their performance is like a coin flip), but there are several ways to reach this same level of overall accuracy. Overall accuracy—the number of trials participants got correct—is depicted by the large number in bold at the center of each table and is the sum of the two diagonal cells (hits and correct rejections). Overall performance ranges from 50 (chance discrimination) at the bottom of the figure to 100 (perfect discrimination) at the top. The space of below chance performance is not depicted in this figure, but can be added if necessary to form a diamond shape, rather than a triangle.

5.4.4 Response Bias

Moving along the x-axis from the left to the right, participants become more conservative in their responses; on the left side of Figure 5.5, they say "True" much more often than they say "False," regardless of whether the prints are from the same or a different source. The opposite is true on the right side of the figure; they say "False" much more often than they say "true." A liberal response bias (on the left of the figure) is represented by a higher column total for the two cells on the left side of each table (i.e., a tendency to say "True"), compared with the two cells on the right. A conservative response bias (on the right of the figure) is represented by a higher column total for the two cells on the right side of each table (i.e., a tendency to say "False"), compared with the two cells on the left. An extremely liberal response bias,

coupled with chance discrimination, means that participants say "True" to every comparison. They will get half of the trials correct in this case—because the experiment has equal base rates—but they will also get half incorrect; they make many hits and many false alarms, while not making any misses or correct rejections. An extremely conservative response bias, on the other hand, coupled with chance discrimination, means that participants say "False" to every trial. Again, they will get half of the trials correct in this case, but they will also get half incorrect; they make many misses and many correct rejections, while not making any hits or false alarms. A neutral response bias means that participants say "True" and "False" equally often. A neutral response bias is depicted along the altitude of the triangle. The left vertex of the triangle shows the maximum liberal bias and the right vertex shows the maximum conservative bias, which changes as a function of sensitivity.

5.4.5 Locating the Results

Data from a two-alternative forced choice experiment, with equal base rates, can be located in the contingency space of all possible results. To do this, the data—that is, the absolute number of the four possible responses to each trial from all participants—can be placed into the four cells of a contingency table, similar to Figure 5.4B. Those frequencies can then be converted into a percentage of the total number of trials in the entire experiment. Depending on the level of precision required, the resolution of the space can be increased or decreased (Figure5.5 has a resolution of 5).

5.5 Contingency Space in Fingerprint Matching

As an illustration, Figure 5.6 is a version of Figure 5.5 altered for the fingerprint matching context (i.e., "Match" and "No Match" rather than "True" or "False"). The average performance of each of the groups from the experiments in Chapters 2 and 4 have been placed into the space.

Moving along the y-axis from the bottom to the top of the figure, participants become more capable of discriminating matching and nonmatching prints. That is, they correctly

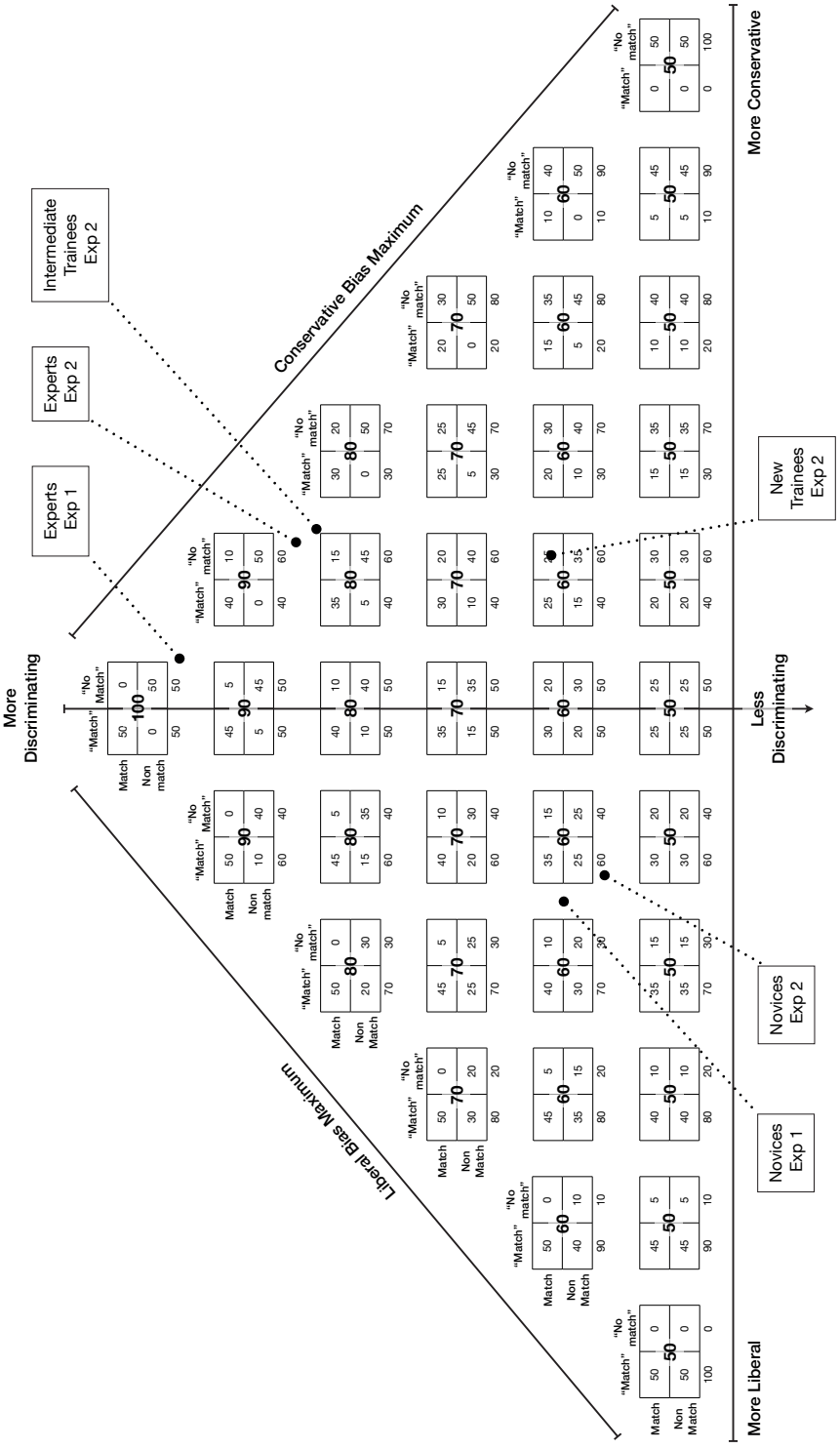

Figure 5.6: A contingency space for the fingerprint matching context. The space represents all possible performance results from a fingerprint discrimination task and the relationship between discrimination and response bias. Pinpointed in the space are the locations of the actual results for each of the six groups from the experiments in Chapters 2 and 4, with nonsimilar nonmatches omitted and the number of trials scaled to give a total of 100. Each filled circle represents the center of the 2×2 contingency table based on the data from each of the conditions.

say "match" to matching prints and "no match" to nonmatching prints, thereby increasing the values in the top left cell (hits) and bottom right cell (correct rejections) in each of the tables that comprise Figure 5.6. Overall performance—the number of comparisons they got correct—is depicted by the large number in bold at the center of each table and is the sum of the two diagonal cells (hits and correct rejections). Results from novices and new trainees lie toward the bottom of this figure—they are reasonably poor print discriminators— compared with intermediate trainees and experts, who are closer to the top.

Moving along the x-axis from the left to the right, participants become more conservative in their responses; on the left side of Figure 5.6, they say "match" much more often than they say "no match," regardless of whether the prints actually match or not. The opposite is true on the right side of the figure; they say "no match" much more often than they say "match." Results from novices lie closer to the left of the figure—they have a reasonably liberal response bias—compared with trainees and experts, who lie closer to the right and have a very conservative response bias. When the data for each of the groups are displayed in this way, we can see that trainees are not any more accurate than novices but they are more conservative. We can also see that experts and intermediate trainees are far more accurate than novices, and that experts and trainees are conservative while novices are liberal. Thinking longitudinally about the progression from novice, to trainee, to expert, it is also clear that the relationship between accuracy and bias changes over time. If desired, this progression could be highlighted in the figure by drawing a continuous pathway from novice (bottom left) to expert (top right). More generally, we can see the effect of moving around the space. Participants can alter their response bias by favouring one type of error over another, but it is clear that this change is independent of accuracy. Similarly, we can see that, as accuracy increases, the degrees of freedom for bias decrease—there are fewer permutations in which to be biased as accuracy increases. Making similar representations in an ROC curve or in a bar chart would, I suggest, be much harder to interpret.

5.6 Conclusion

Here I have described a novel method of depicting the results of a signal detection analysis. First we saw the results of the experiments from Chapter 2 and 4 depicted as a Receiver Operator Characteristic curve (rather than as percent correct), and then in a bias plot. Both representations provided useful ways to visualise measures of accuracy and bias, but neither allowed us to clearly see the relationship of accuracy and bias between groups in a single figure. Locating data in a contingency space of all the possible outcomes of the experiment seems to provide a solution. I suggest that this representation makes the theoretical independence between sensitivity and response bias clearer, and is especially useful for comparing results between experimental conditions. Further, using natural numbers (to make a total of 100) should make results easier to interpret. The figure can be tweaked in several ways to suit the purpose, such as adding sensitivity and bias measures to the labels of each of the groups, adding a longitudinal pathway, or changing the resolution. Anecdotally, my police colleagues seemed to interpret our results more easily when the data were presented in the contingency space, as compared to ROCs or even bar charts. A disadvantage of the space is that it is difficult to pinpoint (locate) a large set of results. For example, Figure 7.4 from Chapter 7, with its 16 groups presented at once, has probably reached the limit of the numbers that can be depicted clearly.

Also important to consider is that the depiction presented here does not account for unequal base rates. In Figure 5.6, 50% of the print pairs where matches, and 50% were nonmatches—equal base rates. Changing the base rates to, for example, 60 matches and 40 nonmatches means that the space would shift to the right, and the two left most bottom contingency tables would contain negative numbers. This would be nonsensical and so the two tables would have to be removed from the left of the space. Going further, base rates of 70 matches and 30 nonmatches would mean that a further 4 of the contingency tables would have to be removed from the left of the space. Eventually, with base rates of 100 matches and 0 nonmatches, the space would be comprised of only 6 contingency tables. The same would happen when changing base rates in the opposite direction. In other words, the total number of possible ways to perform above chance decreases as the ratio of matches to

nonmatches increases. Changing the figure in this way also helps to make clear that the only way to achieve perfect discrimination is to adopt a response bias that reflects the underlying base rates. For example, the only way to achieve perfect discrimination with base rates of 60 matches and 40 nonmatches is to adopt a somewhat liberal response bias, and respond "Match," 60 times and "No match," 40 times.

Further, representing all components of the space—that is, with all the ways to be accurate and biased, and with all permutations of base rates—means that the space would have to be represented in three-dimensions. I am in the initial stages of devising a software tool that creates figures dynamically when given a set of frequencies, a certain resolution, size dimensions, and base rates. The tool could convert raw data into percentage values and create contingency tables that it could then locate in the space. At the extreme, the tool could create a three-dimensional representation with all possible permutations which would be pyramid shaped. One could "fly" through the pyramid to see where the experimental results lie in the space of all possible outcomes and experimental permutations.

A contingency space representation could be applied to many of the research areas that make use of signal detection analyses. The space could also be a useful communication tool for those not well-versed in signal detection theory, or even statistics. Managers and policy makers, for example, could visualise the costs and benefits (payoffs) of different medical tests and interventions (such as breast screening; Gigerenzer, Mata, & Frank, 2009), seeing the relationship between accurate, inaccurate, liberal, and conservative tests. In this way a contingency space depiction of signal detection data could complement traditional signal detection representations.

Chapter 6

The Nature of Expertise in Fingerprint Matching: Experts Can Do a Lot With a Little

6.1 Preface

This chapter is unpublished, but will eventually be submitted for publication in a basic psychology journal. As can be seen in Figure 6.1, this is the first of two chapters in PART 3 - NATURE OF EXPERTISE. This is very much my own work, with Jason Tangen contributing to the overall conceptualisation and 30% to the design of the four experiments.

In PART 1, I established that expertise in fingerprint identification does exist. That is, there are people who have demonstrable and specialized abilities for matching latent fingerprints to their source, and those abilities are superior to lay persons. The fact that experts made so few errors is evidence for impressive human pattern matching performance, possibly exceeding that of experts in other comparable domains. It seems that some combination of training and the daily comparison of untold numbers of fingerprints leads to an uncanny ability to match fingerprints to their source. Experts are drawing on an entire career of experiences in making their decisions. Next I wanted to investigate the basis for their superior performance—how do fingerprint experts do what they do? Experts, likely implicitly,

understand the structure, regularities, and acceptable variation of fingerprint impressions based on their accumulation of instances through experience. In other areas of expertise, such as clinical reasoning in medicine, experts rely heavily on non-analytic processing to make accurate diagnoses. Here, in PART 2, I test whether qualified fingerprint examiners can perform more accurately and consistently than novices when the amount of information is limited—a hallmark of non-analytic cognition. I found that experts can match prints accurately when there is little visual information, little opportunity for direct comparison, and little time to engage in deliberate reasoning. The manuscript to be submitted for publication will include much of the review of literature and motivation from Chapter 1 and implications and conclusions from Chapter 9, but I have not repeated them here so as to reduce redundancy.

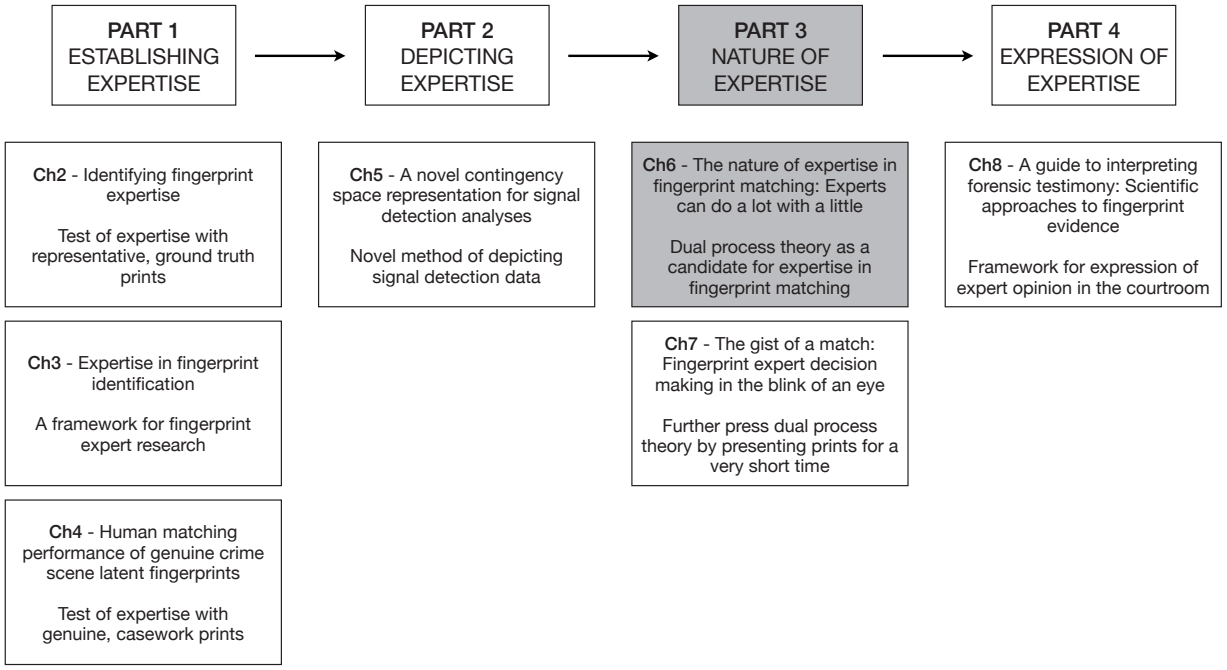

Figure 6.1: Conceptual diagram highlighting Chapter 6, Part 3 of the thesis: "The nature of expertise in fingerprint matching: Experts can do a lot with a little."

6.2 Introduction

Experts are those who consistently performance better than lay people, while *expertise* refers to the mechanisms underlying this superior performance (Ericsson & Charness, 1994). As discussed in Chapter 1, a property of expertise is that experts can perform accurately when given a small amount of information (Klein, 1998; K. K. Evans et al., 2013), and it is common for experts, such as chess players and musicians, to have 10 years experience or to have engaged in 10,000 hours deliberate practice to develop their expertise (Ericsson, 2006; Ericsson, Krampe, & Tesch-Römer, 1993). Compared to novices, experts have a larger number of effective strategies and schemas for performing their tasks accurately and efficiently (Chi, 2006; Schmidt et al., 1990; Shanteau, 1988). Expertise is domain-specific, such that superior performance in a particular domain does not usually guarantee superior performance in another domain, even when the domains have similar surface characteristics (Ericsson, 1996). Experts deploy their skills automatically, largely outside conscious awareness, and experts see objects and situations differently (Chi, 2006; Ericsson, 1996; Kahneman, 2011; Proctor & Dutta, 1995). Bukach, Gauthier, and Tarr (2006) suggest that an expertise framework is powerful way of thinking about common types of perceptual learning, such as face recognition (Gauthier, Williams, Tarr, & Tanaka, 1998).

In order to develop this genuine expertise, the task needs to be regular enough for a person to learn the regularities and receive valid feedback. Experts are exposed to many situations and exemplars—often through deliberate practice—and they receive feedback on their performance. Through experience and feedback, their associative memory recognises situations from environmental cues and generates solutions and decisions accurately and quickly. Experts can rapidly retrieve, from memory, previous instances and decisions relevant to the current situation, while novices rely on formal rules and procedures (Brooks, 2005; Norman et al., 2007). In clinical reasoning, for example, experts cannot predict errors of other experts; experts misinterpret ambiguous clinical signs; experts come to a decision more quickly when they are correct but more slowly when they are incorrect (Norman et al., 2007); doctors' verbal reports of their decision making are insufficient to describe all the aspects of clinical presentations that determine diagnoses (Brooks, 2005); and, although they use

diagnostic criteria in justifying their diagnoses, they do not appear to use them in arriving at their diagnoses (Brooks, Norman, & Allen, 1991).

Norman et al. (2007) suggest that the judgments of diagnosticians are based, in part, on similarity to a previous instance (Schmidt et al., 1990). In dermatology, for example, specific similarity accounts for about 30% of diagnosis, suggesting that much of clinical diagnostic thinking is based on rapid and unconscious recognition of previously encountered situations (Norman & Brooks, 1997). Drew et al. (2013) suggest that rapid memory and decision making by clinicians is explained by a two pathway (selective and nonselective) architecture for visual processing. Moreover, there is very little evidence that diagnostic errors are the result of non-analytic (System 1) reasoning, and experts are as likely to commit errors when they are attempting to be systematic and analytical (System 2; Norman & Eva, 2010).

Experts can intuit the right answer and they are often right. Novices, however, cannot rely on their intuition for accurate performance (Kahneman, 2011). Expertise in a domain does not necessarily include the ability to articulate the basis of that expertise. Asking experts to describe what they are doing and thinking can hurt performance, and experts may not have knowledge or access to the basis of their decisions (Norman & Eva, 2010; Dreyfus & Dreyfus, 2005). Expert intuition is recognition and, if the expert has genuine expertise, their decisions are often accurate.

6.3 Expertise in Fingerprint Identification

When a fingerprint is found at a crime scene it is a human examiner, not a machine, who is faced with the task of identifying the person who left it. Examiners are usually sworn police officers who use image enhancement tools, such as Photoshop or a physical magnifying glass, and database tools to provide a list of possible matching candidates. They place a crime scene print and a suspect print side-by-side—physically or on a computer screen—and visually compare the prints to judge whether the prints match or not, and conclude "identification" or "exclusion," or "inconclusive." Remarkably, given that fingerprint examiners have testified in court for over one hundred years, there have been few experiments directly investigating the extent to which experts can correctly match fingerprints to one another, how competent

and proficient fingerprint experts are, how examiners make their decisions, or the factors that affect performance (M. B. Thompson et al., 2013a; Saks & Koehler, 2005; Spinney, 2010b; Loftus & Cole, 2004; Vokey et al., 2009; Mnookin, 2008a)

Vokey, Tangen, and Cole (2009) found that novices (with no prior experience with prints) can discriminate prints surprisingly well. Tangen, Thompson, and McCarthy (2011) found the qualified court-practicing fingerprint experts were exceedingly accurate in discriminating prints compared to novices. Thompson, Tangen and McCarthy (2013b) replicated the experiment with genuine crime-scene prints and again found that experts and intermediate trainees were more accurate and more conservative than novices. Ulery, Hicklin, Buscaglia, and Roberts (2011) found that fingerprint experts made very few false alarms (approximately 0.01%) but also few hits (approximately 40%), due to an extremely conservative response bias, and also found some variability in consistency of examiners' decisions (Ulery et al., 2012). (I recalculated hits here by discounting the approximately 30% "no value" decisions and translating the "inconclusive" decisions to misses when the prints were from the same source and into correct rejections when the prints were from a different source.) Despite these contributions (and others, e.g., Wertheim et al., 2006b; Dror et al., 2011; Langenberg, 2009), we still know very little about the performance of experts and the nature of their decision making (Mnookin et al., 2010), especially given the history of claimed infallibility (Federal Bureau of Investigation, 1984) and zero error rate for fingerprint comparisons (Cole, 2005).

Several differences between expert and novice fingerprint examiners have been demonstrated previously (see Busey & Dror, 2010; Busey & Parada, 2010, for reviews). Tangen et al. (2011) and Thompson et al. (2013a, 2013b), found that experts showed a conservative response bias, tending to err on the side of caution by making more errors of the sort that could allow a guilty person to escape detection than errors of the sort that could falsely incriminate an innocent person. The superior performance of experts was not simply a function of their ability to match prints, per se, but a result of their ability to identify the highly similar, but nonmatching fingerprints as such. Thompson (2013b) found that intermediate trainees—despite their lack of qualification and average 3.5 years experience—performed about as accurately as qualified experts who had an average 17.5 years experience. And new trainees—despite their 5-week, full-time training course or their 6 months experience—were

not any better than novices at discriminating matching and similar nonmatching prints, they were just more conservative.

Busey et al. (2011) found that experts move their eyes differently from novices and is one of very few published experiments with the aim of understanding the nature of expert performance. Busey and Vanderkolk (2005) found that experts performed better than novices at identifying the matching fragments of fingerprints in noise after a short delay. They also found that inverted fingerprints produced a delayed N170 ERP response in experts but not in novices, suggesting that experts process upright fingerprints configurally. Busey, Wyatte, Vanderkolk, Parada, and Akavipat (2011) found that experts matched fingerprints more accurately than novices when print pairs were shown for 20 seconds. They also found that experts' eye fixations were more variable than novices on latent to ink comparisons, but experts' eye fixations were less variable than novices' on ink to ink comparisons. Fingerprint examiners are also susceptible to a variety of contextual influences (Dror & Charlton, 2006; Busey & Dror, 2010; Dror, 2011).

More generally, the US National Academy of Sciences (National Research Council, 2009) has reported on the absence of solid scientific practices in the forensic sciences, including fingerprints. They highlight the absence of experiments on human expertise in forensic pattern matching and suggest that faulty analyses may be contributing to wrongful convictions of innocent people. The Committee recommend that the US Congress fund basic research to help the forensic community strengthen their field, rectify the lack of basic research, develop valid and reliable measures of performance, understand the effects of bias and human error, and establish evidence-based standards for analyzing and reporting forensic testimony. Further reports from the Scottish Fingerprint Inquiry (Campbell, 2011) and the National Institute for Standards and Training and the US National Institute of Justice, (Expert Working Group on Human Factors in Latent Print Analysis, 2012) as well as US Senate Committee hearings (Carle, 2011; Maxmen, 2012) have echoed these sentiments (Edwards, 2009b). And legal scholars, scientists, and forensic scientists have lamented the lack of a research culture in the forensic sciences (Mnookin et al., 2010).

For expertise to develop, the task needs to stable enough for a person to learn the regularities while receiving valid feedback. Is the task of fingerprint matching one in which

genuine expertise could develop? Examiners are exposed to untold numbers of prints daily, and a routine part of their work is to match an unknown print to a known print. The task of fingerprint identification could conceivably be such that examiners can learn the necessary regularities. When examiners declare a match or "identification," they may ask a colleague to verify their decision. But examiners rarely hear about the outcome of a particular case and, even when they do, it could be months or years after the identification. This situation is in contrast to medical diagnosis, where the effect of a treatment, based on a diagnosis, is much sooner. Further, training materials in fingerprint identification come largely from casework where the ground truth is uncertain. The kind of feedback examiners receive may not be valid or sufficiently immediate to learn the task regularities, and there may not be sufficient opportunity to practice. It is, therefore, not clear whether examiners have the right task and environment to develop genuine expertise, and they could be learning the wrong lessons from experience (Hogarth, 2001).

6.4 Overview of the Present Research

The nature of fingerprint matching expertise is largely unknown, and there are urgent calls for more and better research. In PART 1 of the thesis, I showed that qualified examiners were more accurate than novices and that their superior performance comes largely from their ability to better distinguish highly similar, nonmatching prints from matching prints. Here I begin to explore the nature of fingerprint expertise by probing the cognitive processes that might account for the superior performance of expert fingerprint examiners. Examiners claim that careful, deliberate analysis is the basis of the work that they do (Busey & Parada, 2010; Cole et al., 2008), but a hallmark of genuine expertise is the ability to accurately perform a domain relevant task quickly (Kahneman & Klein, 2009). The exemplar theory of categorisation posits that categorisation is easy for people who have acquired a large number of exemplars from the various categories, because this experience allows them to categorise new items based on the similarity of the new item to the previously encountered exemplars (Brooks, 1978, 2005). Much of diagnostic medicine, for example, is thought to be accounted for by the rapid retrieval of previous instances—non-analytic processing(Norman & Brooks,

1997). If, indeed, fingerprint examiners do draw on a repository of previous instances when making judgments about new prints, then they should be able to perform accurately even when the amount of information in the prints is severely reduced. Fingerprint examiners performing accurately given little information would indicate that, to some extent, they are processing prints non-analytically. Here I ask, to what extent do examiners rely on non-analytic cognition when identifying fingerprints? In doing so, I can evaluate dual-process and exemplar theories as candidates to explain the nature of expertise in fingerprint matching.

In the four experiments that follow, I manipulate the amount of "information" available to decision makers in order to characterise the influence of non-analytic cognition. In Experiment 1, I limit information by adding visual noise to print images and present them either inverted or upright. In Experiment 2, I limit information by spacing prints in time by a few seconds. In Experiment 3, I limit information by spacing prints in time by a few minutes. In Experiment 4, I limit information by presenting fingerprints on screen for just a few seconds.

6.5 Experiment 1: Inversion in Noise

Recognising faces is more difficult when the faces are upside down compared to rightside up, and the impairment is larger than for other kinds of visual stimuli (Yin, 1969; Valentine, 1988). The effect that is best explained by the fact that we have much more experience with upright faces than inverted faces, rather than there being something special about faces per se (Valentine, 1988). As we have seen, there are large expertise effects in fingerprints matching, and experience (exposure) seems to play an important role (M. B. Thompson et al., 2013b). Just as in face identification, I expect that fingerprint identification will be more difficult when prints are inverted, and that inversion will disproportionately affect experts when compared to novices. If so, it would be evidence that experts rely—at least partly—on global rather than local features, and would indicate a reliance on non-analytic processing. In this experiment we show novice and expert fingerprint examiners pairs of artificially noisy prints, that were either upright or inverted, and asked them to indicate whether the prints were a match or not. I expect expert performance to be disproportionately degraded with prints that are inverted compared to upright.

6.5.1 Method

Participants

Novices were 30 undergraduates from The University of Queensland who participated for course credit and who had no experience with prints. Experts were 13 qualified, court-practicing fingerprint experts with an average 13.5 years ($SD = 8.2$) experience from four police organizations: The Australian Federal, New South Wales, Victoria, and Queensland Police.

Procedure

The experiment was a mixed 2 × 3 × 2 design: 2 (Expertise: expert vs novice; between subjects) × 3 (Trial: target vs similar vs random; within subjects) × 2 (Orientation: upright vs inverted; within subjects). We presented participants with a pair of prints displayed side-by-side on a computer screen. The prints appeared onscreen for 60 seconds, after which they disappeared and a slider bar appeared asking participants to judge whether the two prints were the same or different, using a confidence rating scale ranging from 1 (sure different) to 12 (sure same). Judgments were reported by moving a scroll bar to the left ("different") or right ("same"). The scale forced a "match" or "no match" decision, where ratings of 1 through 6 indicated "no match," whereas ratings of 7 through 12 indicated a "match." Half of the prints in the set were presented upright and half were inverted. For each participant, a total of 36 pairs of prints were randomly allocated to the Orientation condition (with the restriction that 18 were upright and 18 were inverted) and to Trial condition (with the restriction that 12 were target pairs, 12 were similar pairs, and 12 were random pairs). The order of presentation for Trial type was random, but Orientation was counterbalanced such that half the participants saw the first half of the set of prints inverted and the second half upright, and the other half of participants saw the first half of the set of prints upright and the second half inverted.

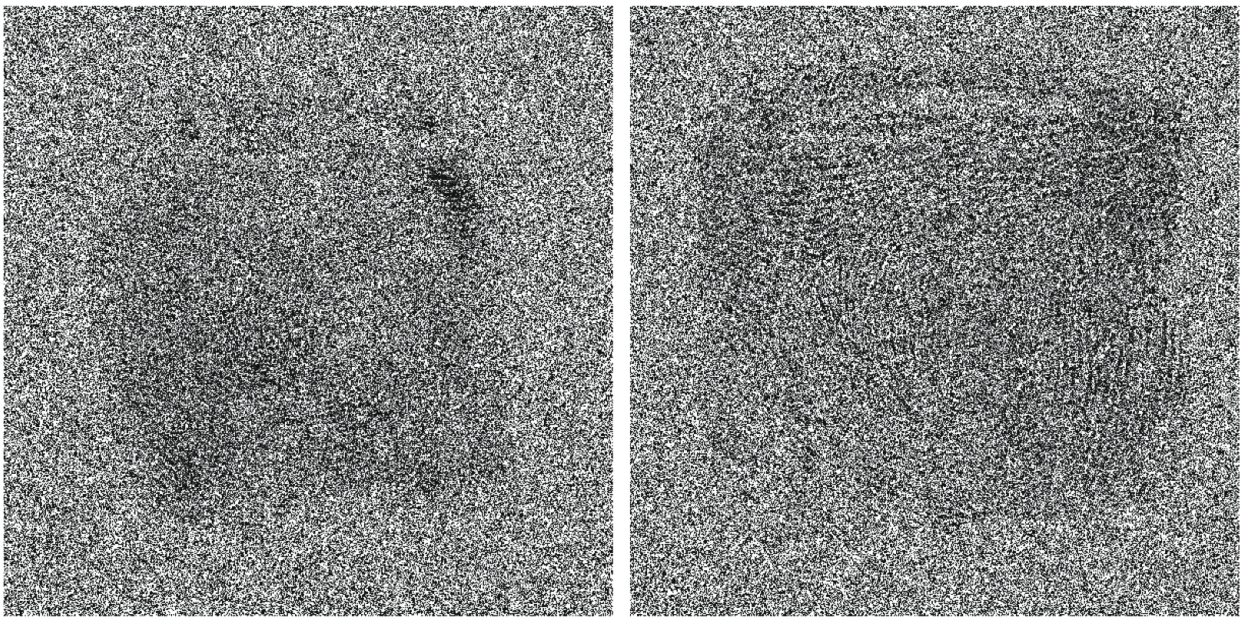

Figure 6.2: Stimuli. An example of a pair of inverted fingerprints with artificial noise. The print on the left is the 'crime scene' print and the print on the right is a similar nonmatching 'suspect' print.

Stimuli

All prints were individual, 'fully-rolled' exemplar prints scanned and extracted from 10-print cards. As shown in Figure 6.2, each trial was made up of two print images, one on the left and one on the right. There were 3 types of trials: (1) matches, where the two images were prints of the same finger of the same person, and were separate instances; (2) similar nonmatches, where the two images were prints from two different people but were deemed similar by a database search algorithm; and (3) nonsimilar nonmatches, where the two images were prints from two different people, and the print on the right was randomly sampled from the set of targets. Each left side print acted as one of the three trial types across the experiment.

There were 36 left side prints in total, and each was paired with a matching print (i.e., a new instance of a print from the same finger of the same person) to create a match trial, or a similar nonmatching print (the result of a national database search as described below) to create a similar nonmatch trial, or a nonsimilar nonmatching print (the right side print was randomly selected from the set of new instance target images) to create a nonsimlar

nonmatching trial. For each participant, each simulated print was randomly allocated to one of the three trial types, with the constraint that there were 12 prints in each condition. So the total number of possible pairs across the experiment was $36 \times 3 = 108$, but each participant responded to only 36 trials in total. In addition to the three trial types, each trial could be present either upright or inverted. The two print images in each trial were always either both upright or both inverted. Each participant saw 12 matching pairs (6 upright and 6 inverted), 12 similar nonmatching pairs (6 upright and 6 inverted), and 12 nonsimilar nonmatching pairs (6 upright and 6 inverted). Allocation to either upright or inverted was random, but was counterbalanced by having either all upright prints presented first or all inverted prints presented first for each participant.

As described in Chapter 4, matching and nonsimilar nonmatching prints were sourced from the Forensic Informatics Biometric Repository (`FIB-R.com`), and similar nonmatching prints were obtained by searching the Australian National Automated Fingerprint Identification System. Artificial noise was added to each of the print images using the "Speckle" function in MATLAB®. "Speckle" is multiplicative noise using the equation J = I + n × I, where I is the image and n is uniformly distributed random noise with mean 0 and variance v. The value for v was set to 10.0 for each image. The amount of visual noise was informally piloted with a three qualified fingerprint experts who were asked to indicate when they thought there was no longer enough detail to make an identification.

6.5.2 Results

Figure 6.3 shows percent correct for experts and novices on each of the three trial types. We subjected the percentages of correct responses to a 2 (Expertise: experts, novices) × 2 (Orientation: upright, inverted) × 3 (Trial: match, similar nonmatch, nonsimilar nonmatch) mixed analysis of variance (ANOVA). Overall, experts were more accurate (87.2% correct) than novices (71.9% correct), $F(1, 41) = 20.869$, $p < .001$. The main effect of Trial was not of particular interest, but it was significant $F(2, 82) = 15.320$, $p < .001$ and follow-up contrasts revealed that non similar nonmatch trials (87.4%) were not significantly different from match trials (84.7%), $F(1, 41) = .406$, $p = .528$, *ns*, but similar nonmatch trials (66.4%)

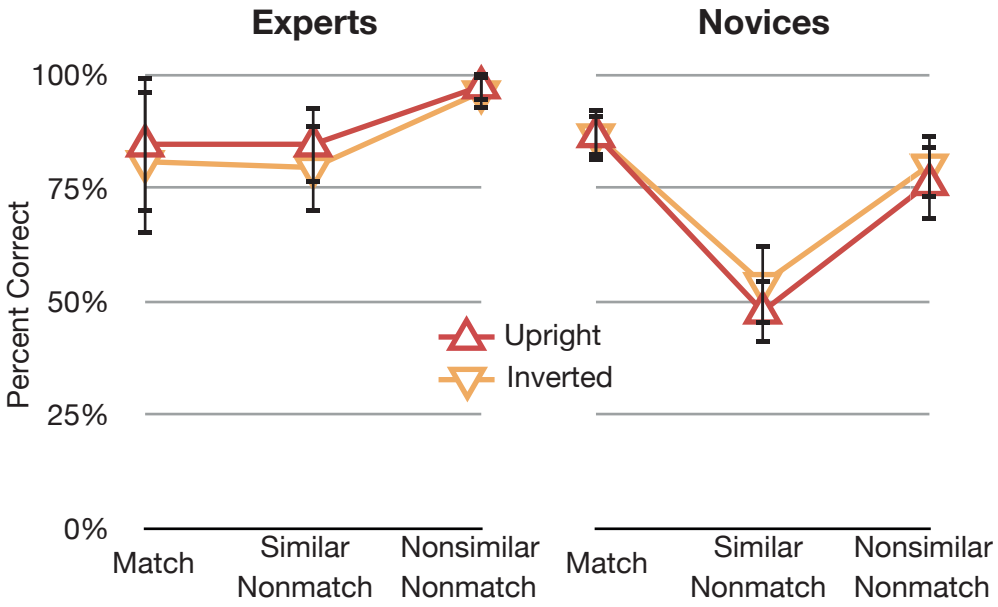

Figure 6.3: Results. Experts' and novices' mean percentage of correct responses for the three trial types (match, similar nonmatch, and nonsimilar nonmatch) and the two orientations (upright and inverted). Error bars represent 95% within-cell confidence intervals.

were significantly different from match trials (87.4%), $F(1, 41) = 14.621$, $p < .001$. Looking at Figure 6.3 it appears that there is no overall difference between upright (79.5%) and inverted prints (79.5%), and so there was no main effect of Orientation, $F(1, 41) < 0.001$, $p = .983$, *ns*.

Of particular interest was the relationship between orientation and expertise, but, as can been seen in Figure 6.3, there was no interaction, $F(1, 41) = 2.799$, $p = .102$, *ns*—experts are no more affected by inversion than are novices. As expected on the basis of previous experience, there was a significant interaction between Trial and Expertise, $F(2, 82) = 9.374$, $p < .001$, and follow-up comparisons reveal that the increased accuracy for nonsimilar nonmatch trials compared to match trials was greater for experts than for novices, $F(1, 41) = 6.946$, $p = .012$, and the decrease in accuracy for similar nonmatch trials compared to match trials was greater for novices than for experts, $F(1, 41) = 13.612$, $p = .001$. Finally, there was no significant interaction between Orientation and Trial, $F(2, 82) = .305$, $p = .738$,

ns, and no significant interaction between Orientation, Expertise, and Trial, $F(2, 82) = .421$, $p = .658$, *ns*.

6.5.3 Discussion

In this experiment, I set out to reduce the amount of information available to experts by inverting fingerprint pairs and adding visual noise. Experts were more accurate than novices overall, and experts were just as accurate with similar nonmatching and matching prints, but novices were far less accurate with similar nonmatching prints than with matching prints. The ability of experts to more accurately discriminate similar nonmatching prints replicates the results from Chapters 2 and 4. I expected that novices would be just as accurate at matching prints when the prints are inverted as when upright, but that qualified fingerprint experts would be less accurate for inverted than for upright prints. I did not, however, find evidence for an inversion effect; experts were just as accurate for inverted prints as they were for upright prints, and novices were also just as accurate for inverted prints as they were for upright prints. This lack of effect is in contrast to Busey et al. (2005) who did find evidence for an expertise effect in the form of a delayed N170 of experts in response to inverted fingerprint stimuli.

It could be that, unlike our experience with faces, fingerprint examiners do have experience with inverted prints or, at least, prints that are often not rotated to be upright. My experience with examiners, however, suggests that they often rotate images upright as part of their regular workflow. It could also be that the prints were too noisy—and, therefore, too distant from the everyday experience of experts—for an inversion effect to manifest. More likely, however, is that the inversion effect that is seen in *memory* tasks might not hold for *matching* tasks in general, which makes sense in hindsight. In this experiment, the stimuli to be judged were presented side-by-side, whereas the stimuli in classic inversion effect tasks are spaced in time. Face recognition tasks involve a comparison of instances with those stored in memory, and people rely on the regularities in memory to make judgments. Matching tasks, however, involve a side-by-side comparison and so there is less reliance on memory. It could be that, by virtue of the matching task, there is little chance for an expertise inversion effect to

manifest. To the best of my knowledge, there have been no published demonstrations of the inversion effect for faces in a matching task. Therefore, as an avenue for further research, I predict that the classic inversion effect would not be present—or heavily attenuated, at least—in a two-alternative, forced choice face *matching* task, where the two faces are presented side-by-side. Similarly, if an expertise effect is evident in a fingerprint recognition memory task, then we could take that as evidence for a dissociation.

Surprisingly, the absolute level of accuracy for experts was much higher than anticipated. Experts' ability to discriminate similar nonmatches was on par with the results from Chapters 2 and 4. And the large accuracy difference between experts and novices with similar nonmatches was on par with previous results. These results come despite the fact that visual noise was added to the prints to a point where experts informally reported that there was not enough information present to make an identification. The level of noise was not directly manipulated in this experiment, but the high performance of experts—especially compared to previous experiments—is evidence against the notion that experts engage in careful, deliberate analysis of the minutiae in a fingerprint image in order to make accurate decisions. Indeed, minutiae such as bifurcations, forks, lakes, etc., were not visible in these highly noisy prints. These findings suggests that fingerprint experts are capable of making accurate decisions when the amount of visual information in the prints is decreased—with artificial visual noise, in this case. Future research could systematically examine accuracy as a function of the level of visual noise.

6.6 Experiment 2: Prints Spaced in Time

When fingerprint examiners compare fingerprints during casework, they compare the suspect prints and the candidate print side-by-side. Here I reduce the amount of information by separating, in time, the two fingerprints that are to be matched. That is, I present the first fingerprint image alone, remove it for a short time, and then present the second fingerprint image alone. In order to perform the task, participants clearly need to hold some information about the first image in memory and use that information to judge whether the second image—that is either from a different finger or is a new instance of a print from the same

finger—matches the first. The task can also be thought of as a test of short term memory for fingerprints. I expect that separating the two prints in time will reduce the accuracy of novices more than experts (Ericsson & Smith, 1991), but experts will still be capable of discriminating prints spaced in time.

6.6.1 Method

Participants

Experts were 16 qualified, court-practicing fingerprint experts with experience ranging from 4 to 34 years ($M = 14.56$, $SD = 7.31$) from the Netherlands Forensic Institute and four police organizations: The Australian Federal, New South Wales, Victoria, and Queensland Police. Novices were 42 undergraduates from The University of Queensland who participated for course credit and who had no experience with prints.

Procedure

Participants began by watching a video that explained the experimental task. Part of the video included an example of two fingerprint images, side-by-side, that are of the same finger from the same person, and two fingerprint images, side-by-side, that are from two different fingers. (The right side image was the same in both cases, i.e., only the left side image changed). When the experiment began, images of fingerprints were displayed on a computer screen and participants were asked to indicate whether two fingerprints are the same or different, i.e., whether the two fingerprints were from the same finger of the same person, or from two different fingers. The first image appeared on screen for 5 seconds, and there was a counter in the top left of the screen instructing participants to count up, out loud, from "One" to "Five," (to prevent them from verbally encoding the features in the first image); then a scrambled visual mask of the first image appeared for 100 milliseconds; then the same counter appeared on screen for 5 seconds with no images; and then the second image appeared on screen and remained until the participant responded by pressing the "same" or the "different" button. The first image was always a simulated crime scene latent fingerprint and the second image was always an exemplar. Each participant responded to a total of 36 trials.

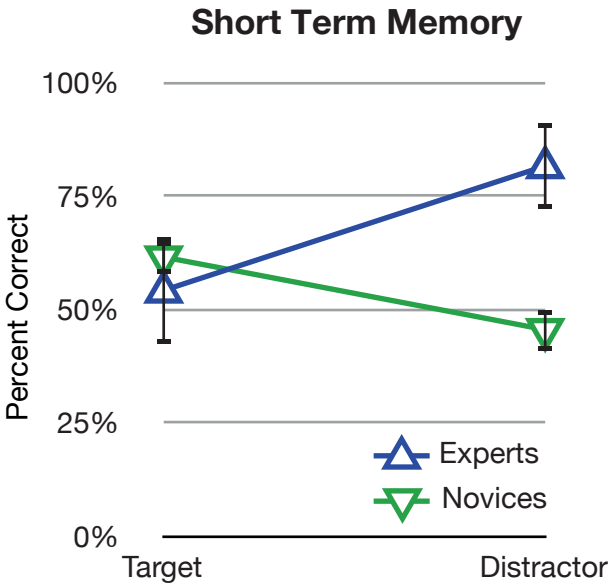

Figure 6.4: Results. Experts' and novices' mean percentage of correct responses for the two trial types (targets and distractors). Error bars represent 95% within-cell confidence intervals.

Stimuli

The stimuli were similar to those from Chapter 2 and consisted of 36 simulated crime-scene prints that were paired with fully rolled exemplar prints. Across participants, each simulated print was paired with a fully rolled print from the same individual (match), and with a nonmatching but similar exemplar (similar distractor). For each participant, each simulated print was randomly allocated to one of the two trial types (target and similar distractor), with the constraint that there were 18 prints in each condition. As described in Chapter 4, matching prints were sourced from the Forensic Informatics Biometric Repository (`FIB-R.com`), and similar nonmatching prints were obtained by searching the Australian National Automated Fingerprint Identification System.

6.6.2 Results

Figure 6.4 shows the percentage of correct responses (mean recognition accuracy) of experts and novices to target and distractor trials, which were subjected to a 2 (Expertise: expert

vs novice) × 2 (Trial: target vs distractor) mixed analysis of variance (ANOVA). Targets and distractors, in this case, can also be thought of as old and new items (in recognition memory parlance), and also as matches and similar nonmatches. Experts (67.9%) were more accurate than novices (53.5%) overall, $F(1, 56) = 29.455$, $p < .001$. Looking at the effect of Trial, participants overall were just as accurate with targets (57.8%) as with distractors (63.5%), $F(1, 56) = 2.034$, $p = .159$, *ns*. But it is clear from Figure 6.4 that this effect is driven mostly by experts on distractors. The interaction between Expertise and Trial was significant $F(1, 56) = 29.393$, $p < .001$, such that experts were more accurate for distractors and novices were more accurate for targets. Expert accuracy for target trials was 54.17%, and 81.60% for distractor trials, and novice accuracy for target trials was 61.51% and 45.50% for distractor trials. Experts show a conservative response bias—they said "different" to most trials—but their ability to discriminate prints in time was superior to novices.

6.6.3 Discussion

I set out to determine the relative performance of qualified, expert examiners and novices on a fingerprint matching task in which the two prints are separated in time. Experts were conservative but still able to discriminate pairs of matching and similar nonmatching prints that were separated by five seconds, even when the amount of information is reduced and when the task was different to their everyday experience. Yet again, the superior performance of experts seems to be driven largely by an expert's ability to discriminate similar, but nonmatching, prints. This experiment gives a preliminary indication of the short term memory capacity of fingerprint examiners for domain relevant stimuli, but future research could systematically vary the delay between prints, and determine whether short term memory for meaningful prints pairs—such as match or a nonmatch pairs presented simultaneously during training—is different from short term memory for less meaningful, individual prints (as in this experiment).

6.7 Experiment 3: Long-term Memory

Evans et al. (2010) found that expert cytologists and radiologists were no better than novices at recognising images of objects and scenes, but were better than novices at recognising images from their domain of expertise. As a further test of the matching ability of fingerprint examiners' and in further effort to reduce the amount of information available, I separate the prints further in time, thus adding a long term memory component. That is, I trained participants on large set of individual fingerprint images, followed by a five minute interval, followed by a recognition memory test. In order to perform the task, participants need to remember whether or not they had seen a print during training. The task can be thought of as a preliminary test of long term memory for fingerprints. I predict that experts will have better long term recognition memory for prints than do novices.

6.7.1 Method

Participants

Participants were the same as in Experiment 2 and they completed both experiments in the same session. (There were two fewer novices in this experiment because they failed to complete it.) Whether a participant completed Experiment 2 first or Experiment 3 first was counterbalanced.

Procedure

Participants were the same as those in Experiment 2. Experiments 2 and 3 were run back-to-back and so the order in which expert participants completed the two experiments was counterbalanced. The stimuli between the experiments were entirely independent; none of the prints used in Experiment 2 were used in Experiment 3. Participants watched a video explaining the task. Part of the video included two separate examples of two fingerprint images, side-by-side, that were of the same finger from the same person. In the learning phase 50 fingerprint images were displayed on screen, one-by-one, for 5 seconds each with a

500 millisecond blank screen between each. Participants were asked to learn the images as best they could and that they would be tested on their ability to recognise the images later.

Following the learning phase, participants completed a word-scramble filler task for five minutes. In order to avoid participants from getting stuck on one word, a 20 second time limit was set so that the correct word would automatically appear in the response field and the participant could move on. Following the filler task, participants watched a second video explaining that their task now was to recognise the fingerprints that they saw in the learning phase, and that some of the prints they will have seen before and some they will not have seen before. The video reiterated the examples of matching and nonmatching fingerprints. In the test phase, 100 fingerprint images were displayed on screen, one-by-one, with the question, "Have you seen this print before," along with "Yes" and "No" response buttons. Fifty of the fingerprint images were old (i.e., they had been presented in the learning phase) and 50 of the fingerprint images were new (i.e., they had not been presented in the learning phase). The old images in the test phase were not simply the same picture displayed again but, rather, a novel instance of an image from the same finger of the same person (i.e., two "matching" prints are two impressions from the same finger taken at different times). This concept was also explained to participants in the instructional video. For each participant, the 50 prints from the learning set were randomly chosen from the learning stimuli set of 100.

Stimuli

The stimuli consisted of a learning set and a testing set. The learning stimuli set consisted of 100 photographs of fully rolled individual fingerprint impressions made using a standard elimination pad and a 10-print card. Each card was scanned in color as a 600-dpi lossless Tagged Information File Format (TIFF), cropped to 750 $\times$ 750 pixels, and isolated in the frame. The testing set also consisted of 100 photographs of fully rolled individual fingerprint impressions made the same way and they "matched" those from the learning set. That is, each learning print had a corresponding match in the testing set, which was a novel photograph instance of the same finger from the same person. In most cases the two images were inked on two different occasions that were at least two weeks apart. The prints were sourced from the Forensic Informatics Biometric Repository.

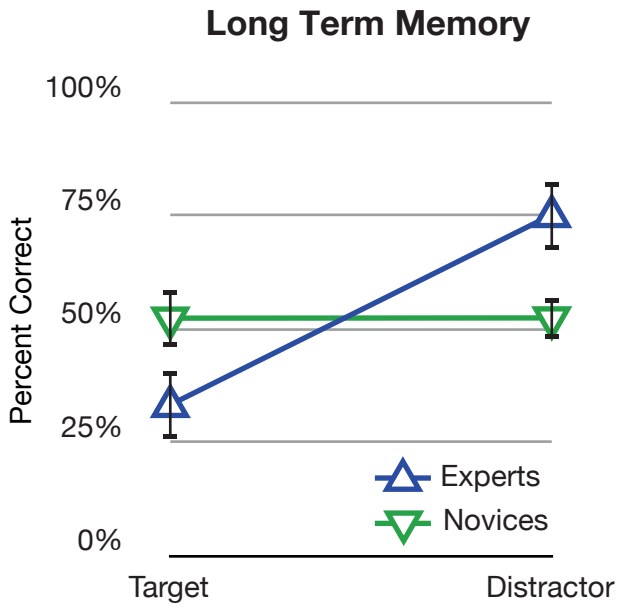

Figure 6.5: Results. Experts' and novices' mean percentage of correct responses for the two trial types (targets and distractors). Error bars represent 95% within-cell confidence intervals.

6.7.2 Results

Figure 6.5 shows the percentage of correct responses (mean recognition accuracy) of experts and novices to target and distractor trials, which were subjected to a 2 (Expertise: expert vs novice) × 2 (Trial: target vs distractor) mixed analysis of variance (ANOVA). Targets and distractors, in this case, can also be thought of as old and new items (in recognition memory parlance), and also as matches and similar nonmatches. Looking at the effect of Trial, everyone was more accurate with distractors (63.8%) than with targets (42.8%), $F(1, 54) = 22.971$, $p < .001$. But it is clear from Figure 6.5 that this effect is driven by experts on distractors. Unexpectedly, experts (54.2%) were no more accurate than novices (52.4%) overall, $F(1, 54) = 2.117$, $p = .151$, *ns*. The interaction between Expertise and Trial was significant $F(1, 54) = 22.861$, $p < .001$, such that the accuracy difference between targets and distractors for experts (41.8% difference) was larger than for novices (0%). Expert accuracy for target trials was 33.3%, and 75.1% for distractor trials, and novice accuracy for target trials was 52.4% and 52.4% for distractor trials. Experts show a conservative

response bias—they said "different" to many trials—and their ability to discriminate prints in long-term memory was not superior to novices.

6.7.3 Discussion

I set out to separate print pairs further in time and test the long term memory of experts relative to novices. Participants were asked to learn 50 fingerprint images and, after a five minute period, recall those they had seen before from a test set of 100 fingerprints (50 old, 50 new). Overall long term recognition memory for experts and novices was the same. Both experts and novices performed around the level of chance. Experts appear to be more conservative than novices, which could be diluting any genuine superior memory ability. Still, on the basis of previous research (K. K. Evans et al., 2010), I expected that experts' long-term memory for prints would be superior to novices. It could be that the regular task of fingerprint matching relies so little on long-term memory that even experts do not develop effective mechanisms for encoding to and retrieving prints. This experiment gives a preliminary indication of experts' long term memory for prints (or lack thereof), but, as with Experiment 2, future research could test long term memory for meaningful print *pairs* (such as match or a nonmatch pairs presented simultaneously during training) is different from memory for less meaningful, *individual* prints (as in this experiment).

6.8 Experiment 4: Short vs Long Exposure Duration

Making accurate decisions, quickly is an indicator of non-analytic cognition. Stimuli for which people have lots of experience, such as natural scenes, can be categorised accurately in around 100ms (Potter & Faulconer, 1975; Intraub, 1981; K. K. Evans, Horowitz, & Wolfe, 2011) and in as little as 20ms (Greene & Oliva, 2009; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007). Those with lots of experience in a domain generally perform better than those with little experience, if the regularities of the task are stable enough to learn from. In this experiment, I limit the amount of information available to participants by presenting fingerprint images

on screen for only a few seconds. Presenting prints quickly should provide little time for decision makers to engage in deliberate, analytic reasoning.

6.8.1 Method

Participants

Two distinct groups participated in the experiment: novices and experts. Novices were 33 undergraduates from The University of Queensland who participated for course credit and who had no experience with prints. Experts were 20 qualified, court-practicing fingerprint experts with an average 14 years ($SD = 8.1$) experience from four police organizations: The Australian Federal, New South Wales, Victoria, and Queensland Police.

6.8.2 Procedure

We presented the participants with pairs of prints displayed side by side on a computer screen. Participants were asked to judge whether the prints in each pair matched, using a confidence rating scale ranging from 1 (sure different) to 12 (sure same). Judgments were reported by moving a scroll bar to the left ("different") or right ("same"). The scale forced a "match" or "no match" decision, where ratings of 1 through 6 indicated "no match," whereas ratings of 7 through 12 indicated a "match." The experiment was a mixed 2 (Expertise: expert vs novice; between subjects) $\times$ 3 (Trial type: match vs similar nonmatch vs nonsimilar nonmatch; within subjects) $\times$ 2 (Deadline: 2 seconds vs 60 seconds; within subjects) design. A pair of prints was shown on screen for either 2 seconds or 60 seconds. Specifically, a fixation cross appeared (2 seconds), followed by a scrambled mask of the prints (2 seconds), followed by the prints to be judged (2 seconds or 60 seconds), followed by the same scrambled mask (2 seconds). A slider bar then appeared asking people to indicate confidence from 1 (sure different) to 12 (sure same).

Stimuli

Stimuli were the same set of prints we used in the experiment from Chapter 2, with a simulated crime scene print on the left and a candidate or "suspect" print on the right. There

were 3 types of trials: (1) matches, where the two images were prints of the same finger of the same person, and were separate instances; (2) similar nonmatches, where the two images were prints from two different people but were deemed similar by a database search algorithm; and (3) nonsimilar nonmatches, where the two images were prints from two different people, and the print on the right was randomly selected from the set. There were 36 left side prints in total, and each was paired with a matching print (i.e., a new instance of a print from the same finger of the same person) to create a match trial, or a similar nonmatching print (the result of a national database search as described below) to create a similar nonmatch trial, or a nonsimilar nonmatching print (the right side print was randomly selected from the set of new instance target images) to create a nonsimlar nonmatching trial. For each participant, each crime scene print was randomly allocated to one of the three trial types, with the constraint that there were 12 prints in each condition, and each pair of prints was randomly selected to one of the two Deadline conditions. Because of the three trial types, the total number of possible pairs in the stimulus set was 432 (144 $\times$ 3), but each participant responded to only 144 trials in total. In addition to the three trial types, each trial could be present for either 2 seconds or 60 seconds. Each participant saw 12 matching pairs (6 for 2 seconds and 6 for 60 seconds), 12 similar nonmatching pairs (6 for 2 seconds and 6 for 60 seconds), and 12 nonsimilar nonmatching pairs (6 for 2 seconds and 6 for 60 seconds). Allocation of prints to the Deadline conditions was random, but was counterbalanced by having either all 2 seconds viewings completed first or all 60 second viewings completed first for each participant.

6.8.3 Results

Thirteen novices were randomly selected and removed from the analysis in order to create two groups of equal size: 20 novices and 20 experts. For each participant, the percentage of trials that were responded to correctly in each condition was calculated, and subjected to a 3 (Expertise: expert vs novice) $\times$ 3 (Trial: target vs similar nonmatch vs nonsimilar nonmatch) $\times$ 2 (Deadline: 2 seconds vs 60 seconds) mixed analysis of variance (ANOVA).

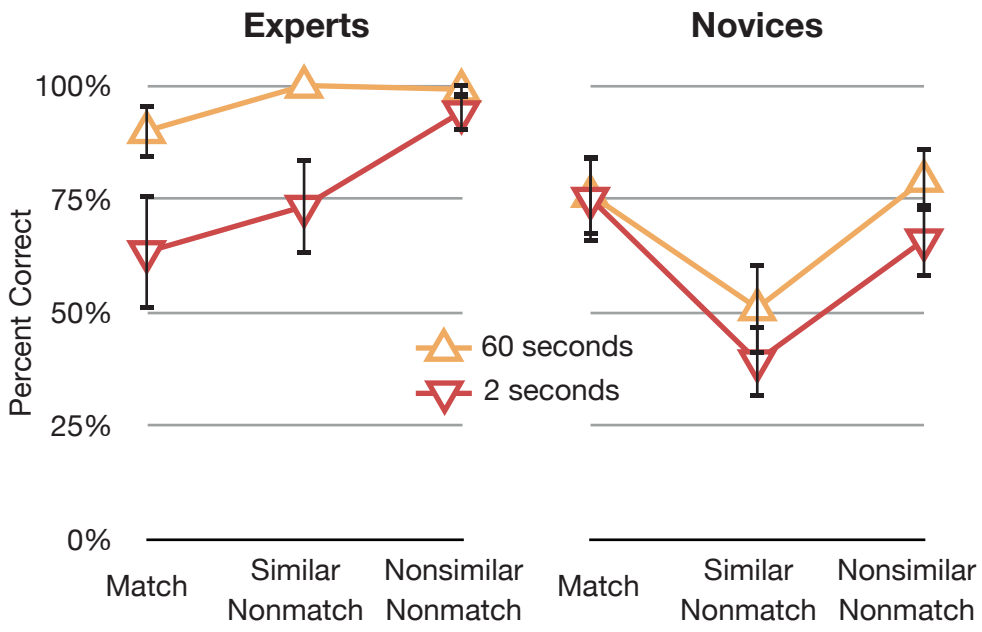

Figure 6.6: Results. Experts' and novices' mean percentage of correct responses for the three trial types (match, similar nonmatch, and nonsimilar nonmatch) and the two deadlines (60 seconds and 2 seconds). Error bars represent 95% within-cell confidence intervals.

As can be seen in Figure 6.6, experts appear to be more accurate than novices overall, and this was borne out by a significant main effect of Expertise, $F(1, 38) = 73.343$, $p < .001$, such that experts (86.7% correct) were more accurate than novices (64.3% correct). It is also clear that participants were more accurate when prints were presented for 60 seconds compared to 2 seconds, and this was borne out by a significant main effect of Deadline, $F(1, 38) = 42.144$, $p < .001$, such that participants were more accurate with a 60 second deadline (82.5% correct) than a 2 second deadline (68.5% correct). There was also a significant main effect of Trial, $F(2, 76) = 9.800$, $p < .001$, and follow-up contrasts revealed that nonsimilar nonmatch trials (84.6%) were significantly different from match trials (76.0%), $F(1, 38) = 8.01$, $p = .007$, and similar nonmatch trials (65.8%) were significantly different from match trials (76.0%), $F(1, 38) = 9.21$, $p = .004$.

It appears from Figure 6.6 that experts are responding to the Deadline condition differently than novices, and there was a significant interaction between Expertise and Deadline, $F(1, 38) = 6.284$, $p < .001$, suggesting that a 60 second Deadline increased accuracy more for

experts (a 19.5% increase) than it increased accuracy for novices (a 6.8% increase). It also appears that novices are particularly inaccurate with similar nonmatches, and this was borne out by a significant interaction between Trial and Expertise, $F(2, 76) = 23.076$, $p < .001$. Follow-up comparisons reveal that accuracy on nonsimilar nonmatch trials compared to match trials was greater for experts (a 21.3% increase) than for novices (a 2.9% decrease), $F(1, 38) = 14.42$, $p = .001$, and accuracy on similar nonmatch trials compared to match trials was greater for experts (an 11.3% increase) than for novices (a 30.4% decrease), $F(1, 38) = 36.08$, $p < .001$. The interaction between Deadline and Trial was nonsignificant, $F(2, 76) = 1.511$, $p = .227$, *ns*, so the effect of deadline did not differ depending on the type of trial. There was a significant interaction between Deadline × Trial × Expertise $F(2, 76) = 4.596$, $p = .013$, suggesting that the interaction between Trial and Expertise differed depending on Deadline. Within-subjects contrasts reveal that that experts perform relatively better with short, nonsimilar nonmatch trials compared to novices $F(1, 38) = 9.98$, $p = .003$ (2s vs 60s, nonsimilar nonmatch vs match, expert vs novice), but the other contrast (2s vs 60s, similar nonmatch vs match, expert vs novice) was nonsignificant, $F(1, 38) = .696$, $p = .409$, *ns*.

6.8.4 Discussion

I set out to promote non-analytic processing by presenting pairs of fingerprints quickly to experts and novices. I found that participants were more accurate when shown prints for 60 seconds than when shown the prints for 2 seconds, and experts were more accurate than novices overall. Both nonsimilar nonmatching and matching trials were more accurate than similar trials, and experts benefitted more from a longer viewing time than did novices. Expert were far more accurate for similar nonmatching prints that were presented for 60 seconds, which replicates results from Chapters 2 and 3. Crucially, experts were far more accurate for similar nonmatching prints that were presented for 2 seconds. This means that experts, unlike novices, could reliably discriminate similar nonmatches even when the prints were presented quickly. This finding is made more interesting by the fact that the vast majority of experts said, informally, that they disliked this experiment because they would not be able to match prints accurately in just two seconds; the evidence suggests otherwise.

Experts were more accurate than novices overall, but it again appears that their superior performance lies in discriminating highly similar, but nonmatching prints. Experts were generally more accurate when they had more time (except for nonsimilar nonmatches, on which experts were highly accurate regardless of presentation time). Presentation time affected experts more than it did novices, such that experts were less accurate with short presentation times than with a long presentation times, relative to novices who were only slightly less accurate with short presentation times than with long presentation times. Experts seem to benefit from more time, much more than novices do. Framed in terms of processing, we can think of a 2 second deadline as engaging non-analytic processing and a 60 second deadline as engaging analytic processing. Overall accuracy for experts, therefore, was 77% non-analytic and 19% analytic, and overall accuracy for novices was 60% non-analytic and 9% analytic. It appears that, although experts are more accurate than novices in the non-analytic condition, experts gain much more from analytic processing than do novices. It is clear experts can match prints accurately when there is little time to engage in deliberate reasoning, suggesting that non-analytic processing accounts for a substantial portion of the variance in superior expert accuracy.

6.9 Conclusion

I set out to determine the extent to which examiners rely on non-analytic cognition when identifying fingerprints, and to evaluate whether dual-process theory helps to explain the nature of expertise in fingerprint matching. In four experiments, I manipulated the amount of "information" available to decision makers in order to characterise the influence of non-analytic cognition. In Experiment 1, I reduced the amount of information available to experts by inverting fingerprint pairs and adding visual noise. There was no evidence for an inversion effect—experts were just as accurate for inverted prints as they were for upright prints–but expert performance with artificially noisy prints was impressive. In Experiment 2, I separated matching and nonmatching print pairs in time. Experts were conservative but still able to discriminate print pairs separated by five seconds, even though the task was quite different to their everyday experience. In Experiment 3, I separated print pairs further in time to

test the long term memory of experts relative to novices. Long term recognition memory for experts and novices was the same, with both performing around chance. In Experiment 4, I presented pairs of fingerprints quickly to experts and novices. Experts were more accurate than novices, particularly for similar nonmatching prints, and experts were generally more accurate when they had more time.

It is clear that experts can match prints accurately when there is reduced visual information, reduced opportunity for direct comparison, and reduced time to engage in deliberate reasoning. These findings suggest that non-analytic processing accounts for a substantial portion of the variance in superior expert fingerprint matching accuracy. Just as in other areas of genuine expertise, experts can get by with little information (Kahneman & Klein, 2009). The findings are also inconsistent with claims that fingerprint identification errors occur when examiners are careless, and with claims that the "method" process of fingerprint identification is based solely on the careful and deliberate analysis of tiny features (minutiae) in a print (Cole, 2009). Across these experiments, however, experts still performed at their best when the most information was available to them. This suggests that while rapid, effortless, non-analytic processing does a lot of work for experts, effortful, analytic processing still seems to play an important role. Here I did not aim to directly target analytic processing, but a complementary research approach with another series of experiments could attempt to isolate the effect of analytic processing in fingerprint identification, as have been conducted in medicine (Brooks et al., 1991; Norman & Eva, 2010).

The evidence presented here for non-analytic processing in expert fingerprint matching suggests that dual-process theory is good candidate to explain the superior performance of experts. Similarly, the large expert and novice differences in accuracy when information is limited seem best accounted for by an instance-based account of forensic reasoning (Brooks, 1978, 2005). The implications of these findings for training, testimony, dual-process models of reasoning are discussed in Chapter 9.

Chapter 7

The Gist of a Match: Fingerprint Expert Decision Making in the Blink of an Eye

7.1 Preface

This chapter is unpublished, but will eventually be submitted for publication in a basic psychology journal. As can be seen in Figure 6.1, this is the second of two chapters in PART 3 - NATURE OF EXPERTISE. This chapter is very much my own work, with Jason Tangen contributing to the overall conceptualisation and 30% to the design of the experiment. The manuscript that will eventually be submitted for publication will include content from Chapters 1 and 9, but I have tried to reduce redundancy by not repeating the content here.

In the previous chapter, we saw that experts can match prints accurately when there is little visual information available and little time to engage in deliberate, analytic processing. Here, I further promote non-analytic processing of fingerprints by presenting fingerprints on screen very briefly. I found that, compared to novices, experts had a much better idea about whether a pair of prints match or not within a very brief time. It seems that fingerprint experts are able to, very quickly, get the "gist" about whether the prints match or not, and

accords with similar findings from other expert domains of visual pattern recognition tasks, such as radiology.

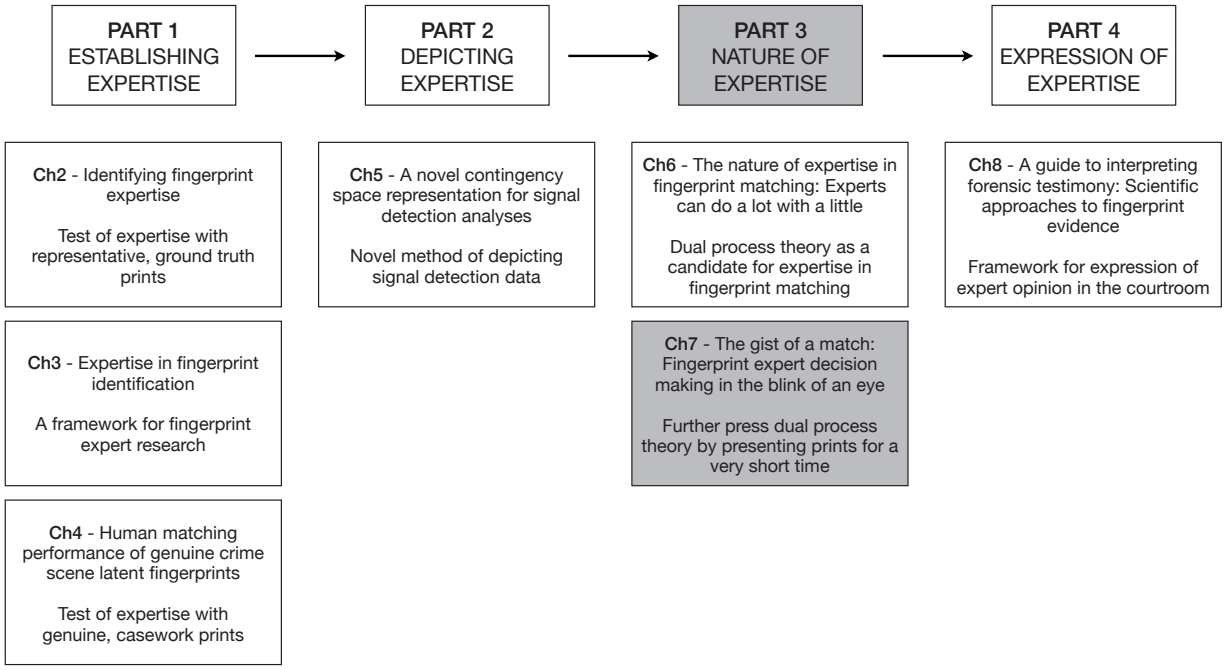

Figure 7.1: Conceptual diagram highlighting Chapter 7, Part 3 of the thesis: "The gist of a match: Fingerprint expert decision making in the blink of an eye."

7.2 Introduction

As discussed in Chapters 1 and 6, experts can perform more accurately and consistently than novices when the amount of information is limited, and when there is little time for careful, deliberate analysis. The ability to assess general meaning is sometimes referred to getting the "gist" of the stimuli (K. K. Evans et al., 2013). In Chapter 6, we saw that fingerprint experts can match prints more accurately than novices with only a two second viewing time. This ability is impressive, especially considering that examiners report that their work is highly analytic, and that the task of identifying fingerprints is based on finding minutiae that correspond. It seems that fingerprint experts are able to learn the regularities of the

task though exposure and feedback, and they accumulate task relevant instances and rapidly retrieve them from memory (Brooks, 2005; Norman et al., 2007).

Two *second* discrimination is impressive, but there is evidence that stimuli for which people have lots of experience can be categorised accurately within a few hundred *milliseconds*, such as natural scenes (Potter & Faulconer, 1975; Intraub, 1981; K. K. Evans et al., 2011; Joubert et al., 2007; Greene & Oliva, 2009), and medical images (Drew et al., 2013; K. K. Evans et al., 2013). In the experiment reported here, I test the limits what of expert fingerprint examiners can achieve with brief stimuli presentation times. Presenting prints for a few hundred milliseconds will give little opportunity for participants to make analytic, rule-based judgments. If experts can match prints more accurately than novices after seeing the prints for a few hundred milliseconds, then that would suggest that they are making use of instances of matching and nonmatching prints already stored in memory in order to judge new instances (Rouder & Ratcliff, 2006, 2004; Brooks, 1978; Norman et al., 2007).

In choosing the minimum presentation time I chose the time required to make a voluntary eye movement, considering that the two fingerprints are presented side-by-side (the typical latency for a voluntary saccade is around 200ms; Carpenter, 1988). The maximum presentation time of 2000ms was chosen as a replication of Experiment 4 in Chapter 6, and 500ms and 1000ms were chosen as midpoints. I predict that experts will be more accurate than novices with these brief presentation times, and that the difference in accuracy between experts and novices will increase as the presentation time increases.

7.3 Method

7.3.1 Participants

Novices were 31 undergraduates from The University of Queensland who participated for course credit and who had no experience with prints. Experts were 21 fingerprint examiners with an average 11.3 years ($SD = 8.4$) experience from two Australian police agencies.

7.3.2 Stimuli and Procedure

The experiment was a mixed 2 (Expertise: expert vs novice; between subjects) x 3 (Trial type: match vs similar nonmatch vs nonsimilar nonmatch; within subjects) x 4 (Deadline: 250ms vs 500ms vs 1000ms vs 2000ms; within subjects) design. All stimuli were individual, 'fully-rolled' exemplar prints scanned and extracted from 10-print cards. Each trial was made up of two print images, one on the left and one on the right. There were 3 types of trials: (1) matches, where the two images were prints of the same finger of the same person, and were separate instances; (2) similar nonmatches, where the two images were prints from two different people but were deemed similar by a database search algorithm; and (3) nonsimilar nonmatches, where the two images were prints from two different people, and the print on the right was randomly selected from the set.

There were 144 left side prints in total, and each was paired with a matching print (i.e., a new instance of a print from the same finger of the same person) to create a match trial, or a similar nonmatching print (the result of a national database search as described below) to create a similar nonmatch trial, or a nonsimilar nonmatching print (the right side print was randomly selected from the set of new instance target images) to create a nonsimlar nonmatching trial. For each participant, each left side print was randomly allocated to one of the three trial types (with the constraint that there were 48 prints for each trial type), and each pair of prints was randomly selected to one of the four Deadline conditions (with the constraint that there were 36 prints for each Deadline type). The total number of possible pairs across the experiment was $144 \times 3 = 432$, but each participant responded to only 144 trials in total. In addition to the three trial types, each trial could be present for either 250ms, 500ms, 1000ms or 2000ms. For each participant, allocation of prints to the Deadline and Trial conditions was random, and the trial and deadline conditions were randomly distributed throughout the experiment.

As described in Chapter 4, matching and nonsimilar nonmatching prints were sourced from the Forensic Informatics Biometric Repository (`FIB-R.com`), and similar nonmatching prints were obtained by searching the Australian National Automated Fingerprint Identification System. Participants were presented with pairs of prints displayed side by side on a computer

screen. They were asked to judge whether the prints in each pair matched, using a confidence rating scale ranging from 1 (sure different) to 12 (sure same). Judgments were reported by moving a scroll bar to the left ("different") or right ("same"). The scale forced a "match" or "no match" decision, where ratings of 1 through 6 indicated "no match," whereas ratings of 7 through 12 indicated a "match." A fixation cross appeared on screen (2000ms), followed by a scrambled mask of the pair of prints (2000ms), followed by the pair of prints to be judged (250ms, 500ms, 1000ms, or 2000ms), followed by the same scrambled mask (2000 ms). A slider bar then appeared asking people to indicate confidence from 1 (sure different) to 12 (sure same).

7.4 Results

Figure 7.2 shows the ROCs for each of the four deadline conditions where the distractors were similar non matches. Figure 7.3 shows the ROCs for each of the four deadline conditions where the distractors were nonsimilar non matches. These figures are for the reader's convince, but, for reasons discussed in Chapter 5, I will use a contingency space representation of the data to describe the results (Figure 7.4).

7.4.1 Accuracy

As a measure of accuracy, the average A′ was calculated for each of the experimental conditions. A′ is the area under the ROC curve and a measure of discriminability which ranges from 0.0 to 1.0, with .5 indicating chance performance and indicating 1.0 perfect performance. A correction was applied to the A′ values by adding .5 to the number of hits (or false alarms) and dividing by the number of trials plus 1 (Williams & Simons, 1999). For example, if a participant correctly responded "match" to 10 of 12 trials, then the corrected hit rate would be $10.5 \div 13 = .8076$. Average A′ values were subjected to a 2 (Expertise: experts, novices) $\times$ 2 (Trial: similar nonmatch, nonsimilar nonmatch) $\times$ 4 (Deadline: 250ms, 500ms, 1000ms, 2000ms) mixed analysis of variance (ANOVA).

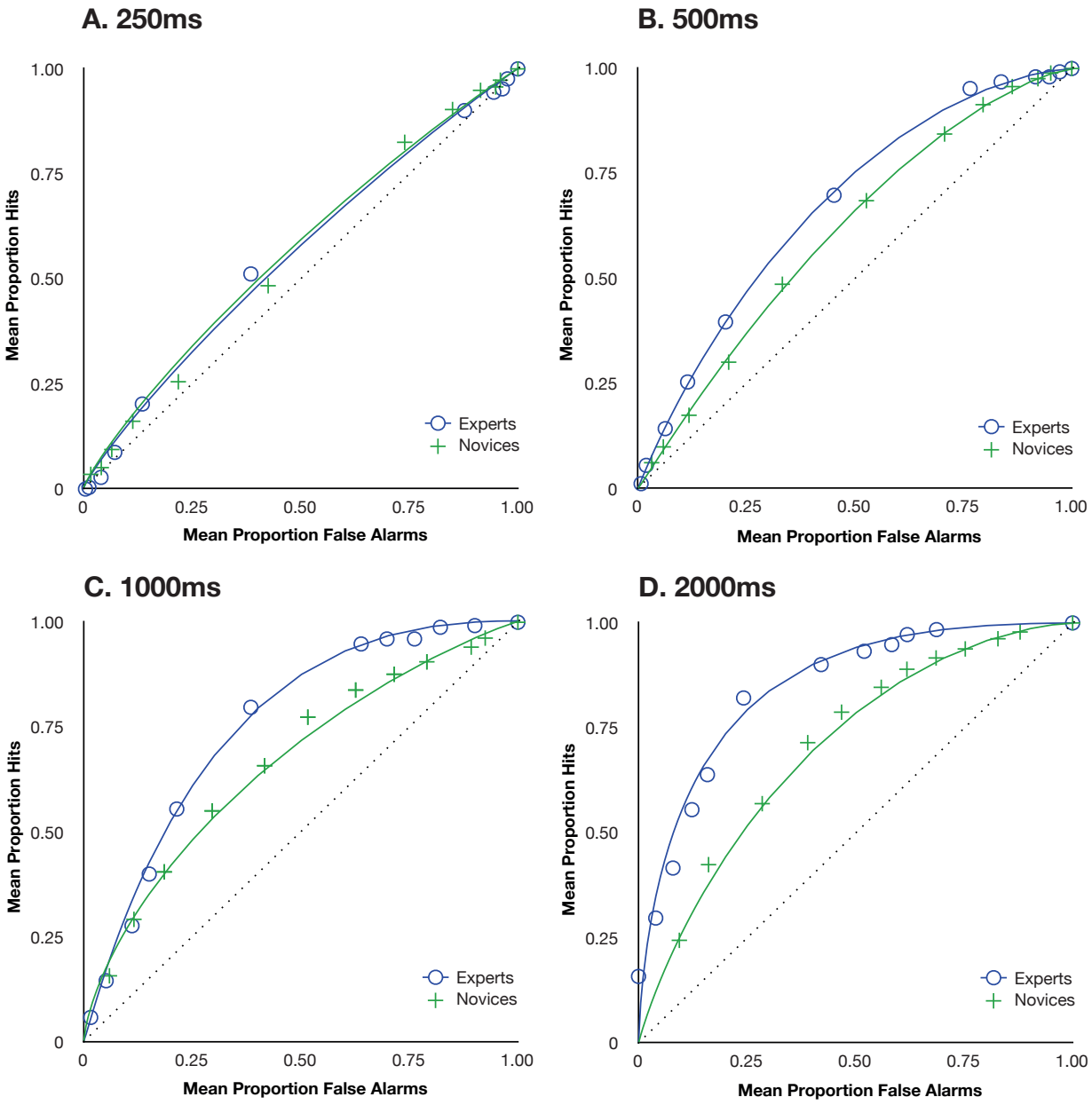

Figure 7.2: Results. Experts' and novices' Receiver Operator Characteristics for the four deadlines (250ms, 500ms, 1000ms, and 2000ms) for matches and similar nonmatches.

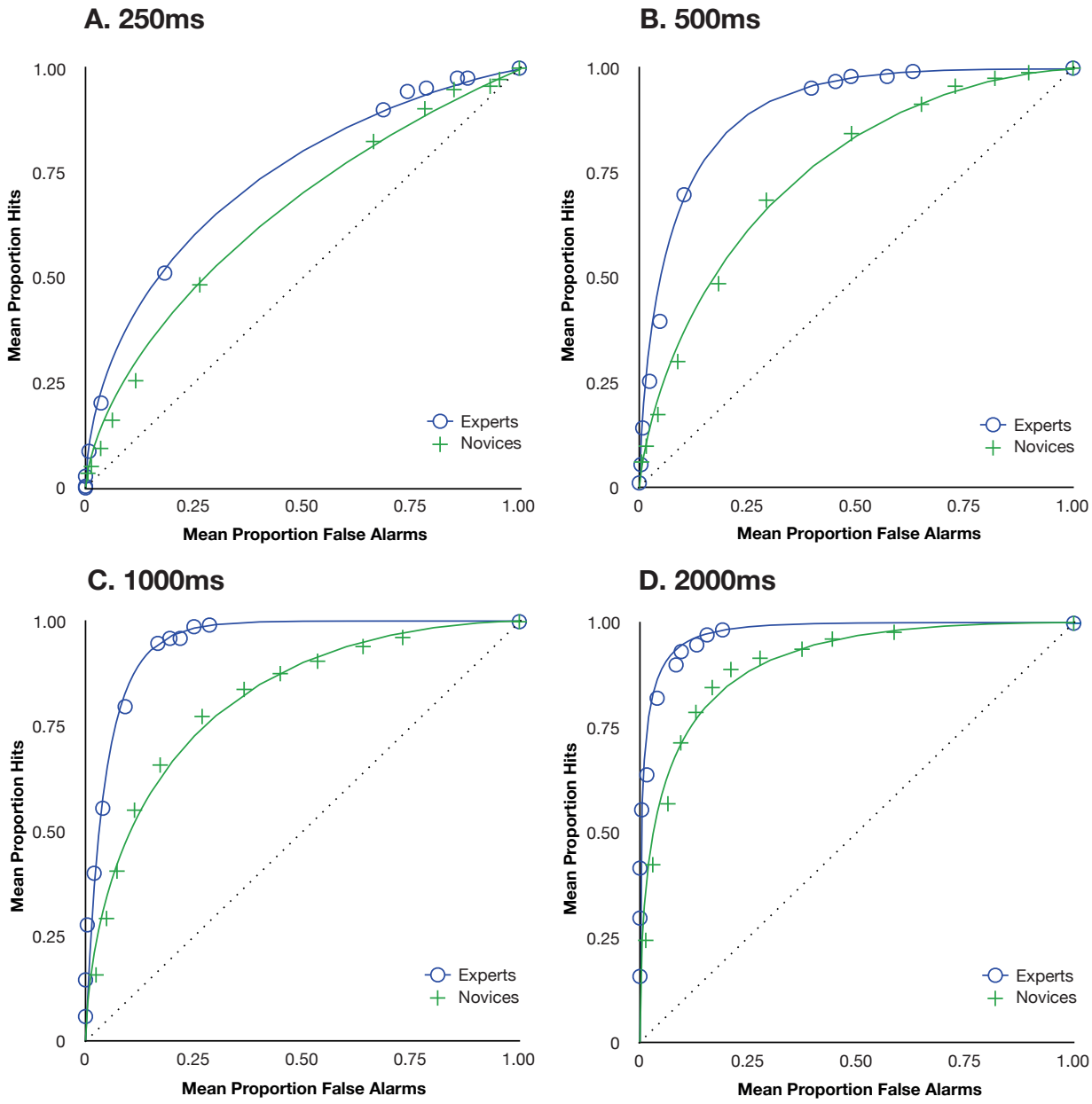

Figure 7.3: Results. Experts' and novices' Receiver Operator Characteristics for the four deadlines (250ms, 500ms, 1000ms, and 2000ms) for matches and nonsimilar nonmatches.

In Figure 7.4, it appears that the responses of experts generally lie towards the top of the figure (more accurate) while the responses of novices generally lie towards the bottom of the figure (less accurate). This was borne out in a significant main effect of Expertise, $F(1, 50) = 21.577$, $p < .001$, such that experts ($A' = .788$) were more accurate than novices ($A' = .712$). It also seems that nonsimilar nonmatches appear higher in the space than do similar nonmatches, a this was borne out in a significant main effect of Trial, $F(1, 50) = 215.205$, $p < .001$; such that nonsimilar nonmatches (.815) were more accurate than similar nonmatches (.686). As deadline values increase (i.e., from 250ms to 2000ms), it appears that they generally climb up the figure (increasing accuracy), and this was borne out in a significant main effect of Deadline $F(3, 150) = 39.146$, $p < .001$; 250ms = .636, 500ms = .726, 1000ms = .794; 2000ms = .845.

There was no significant interaction between Deadline and Expertise, $F(3, 150) = .085$, $p = .968$, *ns*; no significant interaction between Deadline and Trial $F(3, 150) = 1.629$, $p = .185$, *ns*; and no significant interaction between Trial and Expertise, $F(1, 50) = .036$, $p = .851$, *ns*. There was a significant interaction between Expertise, Trial, and Deadline, $F(3, 150) = 3.564$, $p = .016$, and it appears that this effect is driven by the accuracy of all conditions increasing as the Deadline time increases, except for novices looking at similar nonmatches (which are clustered in the bottom left).

7.4.2 Bias

As a measure of bias, the average B''_D was calculated for each of the experimental conditions, with the same correction above applied. B''_D is a convenient statistic because it varies from -1.0 to $+1.0$, with positive numbers indicating a bias to respond "No Match," negative numbers indicating a bias to respond "Match," and 0.0 indicating no bias (see Donaldson, 1992, for discussion). Average B''_D values were subjected to a 2 (Expertise: experts, novices) $\times$ 2 (Trial: similar nonmatch, nonsimilar nonmatch) $\times$ 4 (Deadline: 250ms, 500ms, 1000ms, 2000ms) mixed analysis of variance (ANOVA).

In Figure 7.4, it appears that the responses of experts generally lie more towards the right of the figure (more conservative) while the responses of novices lie more towards the left side

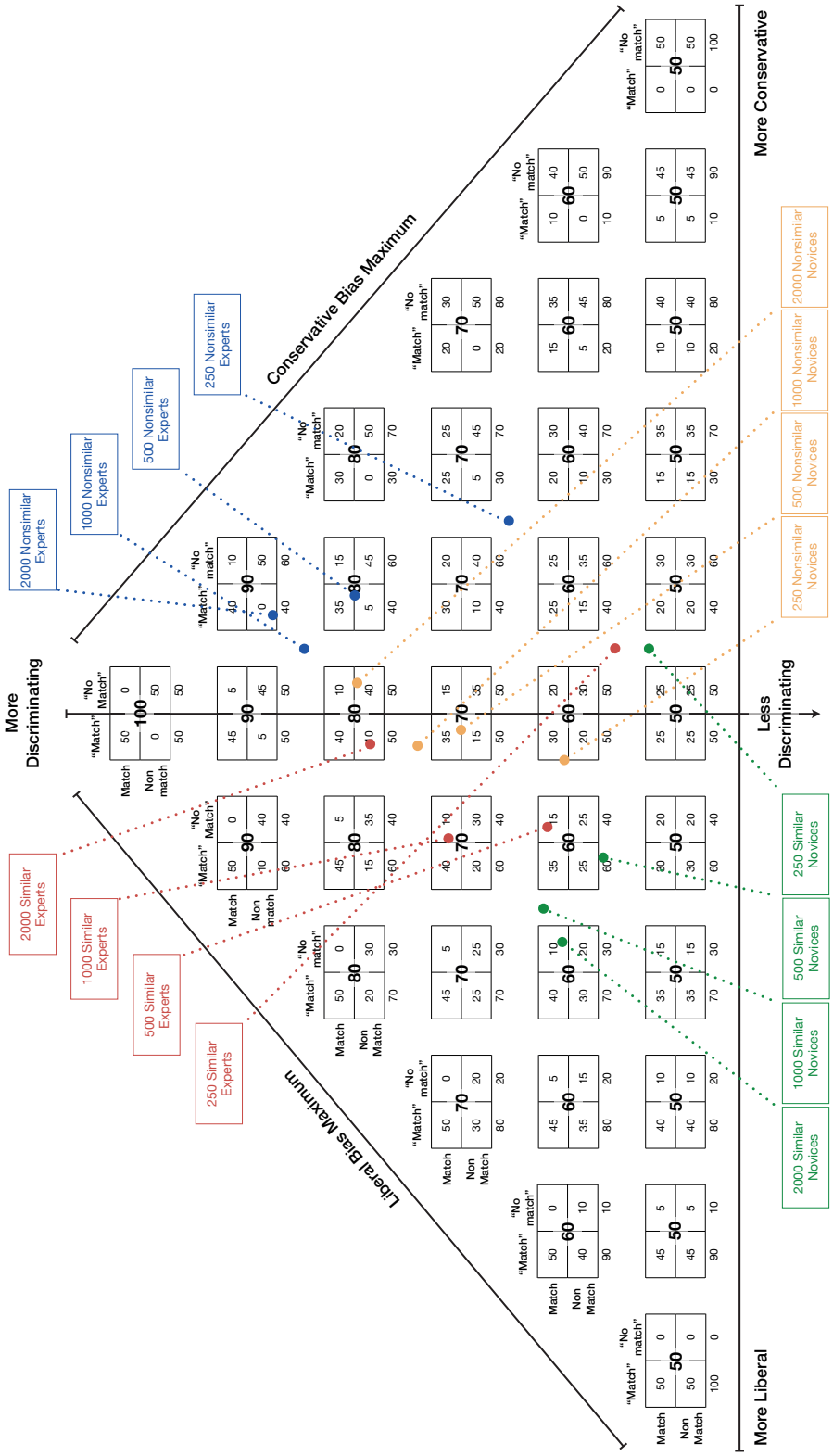

Figure 7.4: Contingency space. The space represents all possible performance results from a discrimination task and the relationship between discrimination and response bias. Each of the tables that comprise the figure is a 2×2 contingency table. Each filled circle represents the center of the 2×2 contingency table based on the data from each of the sixteen conditions, and the number of trials is scaled to give a total of 100.

of the figure (more liberal), and this was borne out in a significant main effect of Expertise, $F(1, 50) = 6.975$, $p = .011$, such that experts (+.128) were more conservative than novices (−.140). It also appears that nonsimilar nonmatches lie towards the right of the figure (more conservative) while similar nonmatches lie towards the left (more liberal), and this was borne out in a significant main effect of Trial, $F(1, 50) = 183.927$, $p < .001$; such that nonsimilar nonmatches (+.228) were responded to more conservatively than similar nonmatches (−.240).

It appears that participants' responses move from right to left in the space (more liberal) as Deadline increases, and this was borne out in a significant main effect of Deadline $F(3, 150) = 12.504$, $p < .001$; 250ms = +.257, 500ms = −.058, 1000ms = −.152; 2000ms = −.071—people become more liberal as Deadline increases. There was a significant interaction between Deadline and Expertise, $F(3, 150) = 5.021$, $p = .002$, suggesting that novices become more liberal as deadline increases but experts do not. The other interactions were not of particular interest, but there was a significant interaction between Deadline and Trial $F(3, 150) = 5.299$, $p = .002$; no significant interaction between Trial and Expertise, $F(1, 50) = 1.811$, $p = .184$, *ns*; and a significant interaction between Expertise, Trial, and Deadline, $F(3, 150) = 2.682$, $p = .049$.

7.5 Discussion

A hallmark of expertise is the ability to accurately perform a domain relevant task quickly. I set out to test the limits of what expert fingerprint examiners can achieve with brief stimuli presentation times. Experts were more accurate than novices overall, and, compared to novices, they have a much better idea about whether a pair of prints match or not within rapid time. Experts were more conservative than novices and participants were generally conservative with nonsimilar nonmatches and generally liberal with similar nonmatches.

It is clear that, through their experience, experts can learn the regularities of matching and nonmatching prints and, because presentation times were so short (from 250ms to 2000ms), there was little time to engage in careful, deliberate analysis of the minutiae in a fingerprint image. These findings suggest that fingerprint experts are capable of making accurate decisions when the amount of visual information in the prints is dramatically decreased,

and is evidence that experts are processing prints non-analytically. Experts appear to have learned the regularities in matching and nonmatching prints, probably on a global rather than featural level. On the other hand, expert accuracy increased as the presentation time increased, suggesting that analytic processing is interacting with with non-analytic processing to produce maximum accuracy. These findings are similar to those of Evans et al. (2013) who found that expert radiologists can discriminate between normal and abnormal medical images that are presented in the blink of an eye.

It is possible, however, that participants are picking up something tangential in the stimuli that is helping them to perform accurately. It could be that the matching (and the nonsimilar nonmatching) prints that we manufactured have something about them—the lighting, the way the photograph was taken, the ink type, etc.—that makes them easy to tell apart from similar nonmatches (that were sourced from police archives). This possibility, however would not account for the accuracy difference between experts and novices—especially for those presented for 500ms—which is why a novice comparison group, rather than simply comparing to chance, is crucial. It seems that fingerprint experts are able to, very quickly, get the "gist" about whether two fingerprints match or not. These findings contradict claims that fingerprint identification is a wholly "scientific process" that requires careful, thorough analysis in order for judgments to be accurate. It seems that experts have developed a fast and accurate method of fingerprint matching based on their vast experience and familiarity with fingerprints, a finding that supports an instance-based account of forensic reasoning (Brooks, 1978, 2005). The implications of these findings are discussed in Chapter 9.

Chapter 8

A Guide to Interpreting Forensic Testimony: Scientific Approaches to Fingerprint Evidence

8.1 Preface

This chapter is extracted from a published article in the journal *Law, Probability and Risk*. As can be seen in Figure 4.1, this is the first and only chapter of PART 4 - EXPRESSION OF EXPERTISE. The published article included several sections, but I have included in this thesis only from the abstract to section four, which I contributed to heavily. I am entirely responsible for the language in *The Guide* itself and I am very much responsible for the *Insights from Medicine: The Diagnostic Model* section, with Jason Tangen contributing 30% to the section's conception and writing. Reference:

Edmond, G., **Thompson, M. B.**, & Tangen, J. M. (2013). A guide to interpreting forensic testimony: Scientific approaches to fingerprint evidence. Law, Probability and Risk. doi:10.1093/lpr/mgt011

In PART 1, I explored expert-novice differences in fingerprint matching and, in PART 3, I investigated the nature of finerprint expertise. Further empirical evidence on the performance of fingerprint experts will continue to emerge, but we are still left with the problem of what experts can legitimately claim in court. Claims of individualisation and zero-error are, given the epistemic problems and the results of recent experiments, unsustainable. The question is, now that we have an indication of the performance of qualified fingerprint experts, what claims can they reasonably make in court and how should the latest empirical evidence be presented to triers of fact? Based on an approach from diagnostic medicine, I propose a way to express information about data from the aggregation of many controlled experiments in order make inferences about the particular case. Broadly, the diagnostic model is an approach that offers information about similar situations in order to help decision-makers reason about the particular case.

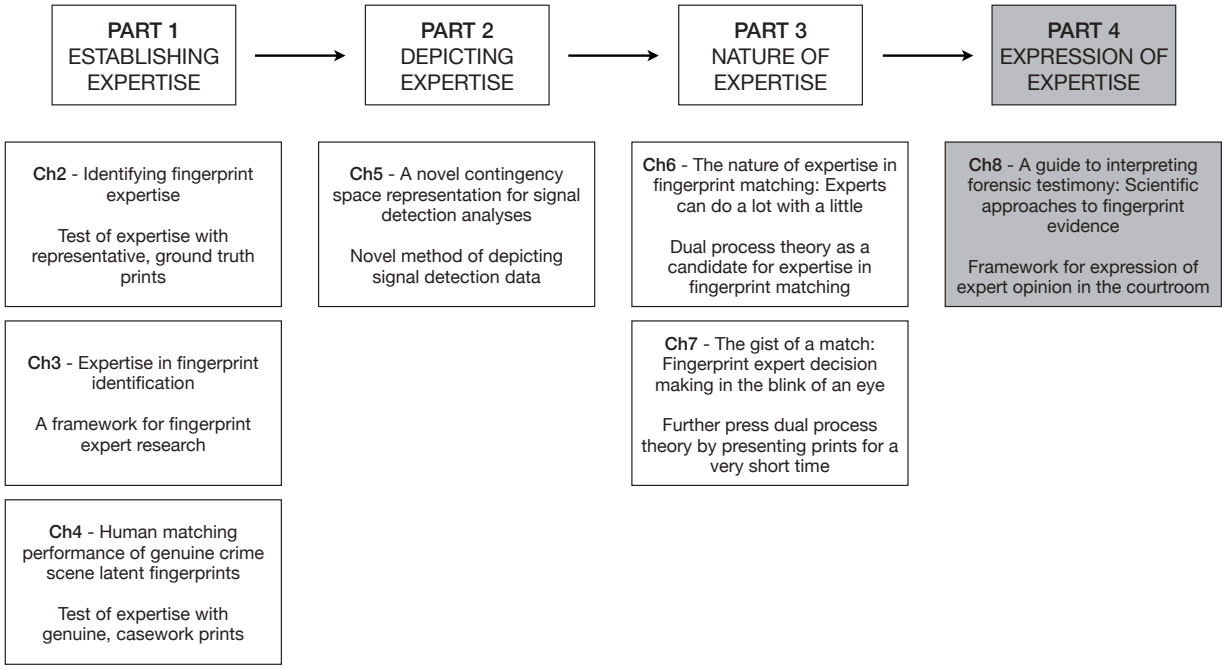

Figure 8.1: Conceptual diagram highlighting Chapter 7, Part 3 of the thesis: "A guide to interpreting forensic testimony: Scientific approaches to fingerprint evidence."

8.2 Abstract

In response to criticism of latent fingerprint evidence from a variety of authoritative extra-legal inquiries and reports, this essay describes the first iteration of a guide designed to assist with the reporting and interpretation of latent fingerprint evidence. Sensitive to the recommendations of these reports, we have endeavoured to incorporate emerging empirical evidence about the matching performance of fingerprint examiners (i.e., indicative error rates) into their testimony. We outline a way of approaching fingerprint evidence that provides a more accurate—in the sense of empirically and theoretically justified—indication of the value of fingerprint evidence than existing practice. It is an approach that could be introduced immediately. The proposal is intended to help non-experts understand the value of the evidence and improve its presentation and assessment in criminal investigations and proceedings. This first iteration accommodates existing empirical evidence and draws attention to the gap between the declaration of a match and positive identification (or individualization). Represented in this way, fingerprint evidence will be more consistent with its known value as well as the aims and conduct of the accusatorial trial.

8.3 Reforming the Presentation of Comparison Evidence

Fingerprint examiners have been active in investigations and presented "identification" evidence in criminal courts for more than a century (Cole, 2002). Notwithstanding increasing automation, examiners continue to play a central role in the comparison of prints, the interpretation of prints, and in attributing significance to apparent matches. When confronted with an unknown print, usually a part (or fragment) of a fingermark recovered from a crime scene (known as a "latent"), it is the examiner who decides if the latent print provides sufficient information to interpret and, if so, whether it matches a known (i.e., reference) print (Dror & Mnookin, 2010). Where the examiner is satisfied about the sufficiency of the print and declares a "match," this is conventionally understood by examiners, and represented

to others, as positive identification of the person who supplied the reference print to the exclusion of all other persons (Mnookin, 2008b).

Remarkably, given the interpretive (i.e., subjective) dimensions of comparison and the considerable gap between declaring a match and positive identification (so-called individualization; Cole, 2009; Cole et al., 2008; Saks & Koehler, 2007; Kaye, 2010), there have been few scientific investigations of the human capacity to correctly match fingerprints, let alone attach significance to apparent similarities (National Research Council, 2009; Loftus & Cole, 2004). Nevertheless, for more than a hundred years, and in the absence of experimental support, fingerprint examiners have claimed that fingerprint evidence is basically infallible (Cole, 2005; Federal Bureau of Investigation, 1984). These assertions are typically justified by reference to training and experience, and the use of a method such as ACE-V: Analysis, Comparison, Evaluation and Verification (Koehler, 2012; Cole et al., 2008; Haber & Haber, 2007; Vokey et al., 2009; Tversky & Kahneman, 1974), assumptions about the uniqueness of fingerprints, along with legal acceptance and the effectiveness of fingerprint evidence in securing confessions and convictions (Alpert & Noble, 2009; B. L. Garrett, 2011). In recent decades, however, commentators have questioned uniqueness (and its significance) and dismissed claims about error-free, positive identification as scientifically implausible. In recent years, these doubts have materialized in notorious mistakes, and scholarly criticisms endorsed in independent inquiries and reports (Cole, 2010).

This article presents a first iteration of what we envisage could be an evolving response to the vexed issue of the reporting (or expression) of forensic comparison evidence (Koehler, 2008, 2012). Conceived as a practical aid to assist with the presentation and interpretation of forensic science evidence, the guide to interpreting forensic science testimony (or Guide) is intended to embody the current state of relevant scientific research in relation to a particular technique (or set of techniques). This empirically predicated guide is designed to assist with the evaluation of evidence by highlighting areas of demonstrated expertise and incorporating an indicative error rate to assist with the assessment of expert opinion (Edmond, 2012a). Using the example of fingerprints, it would enable a fingerprint examiner to express an opinion about whether two prints match (or do not match) against the backdrop of an empirically derived error rate and other indicators of expertise and its limitations. Introducing an error

rate into the provision of comparison evidence assists with the evaluation of opinions and in delimiting the scope of expertise (Gieryn, 1999).

8.4 Background to the Guide: Evidence, Expertise, Error and Advice

8.4.1 Authoritative Reports and Recommendations

In a landmark report issued in 2009, a committee of the National Research Council (NRC) of the US National Academy of Sciences (NAS) drew attention to questionable practices and the lack of research in many areas of forensic science. The committee was surprised to discover that many forensic science disciplines are typically not supported by scientific research and that analysts are not necessarily bound by experimentally derived standards to ensure the evidence offered in courts is valid and reliable (Edwards, 2009a; Saks & Koehler, 2005).

In confronting language, the report itself states:

> Often in criminal prosecutions and civil litigation, forensic evidence is offered to support conclusions about "individualization" (sometimes referred to as "matching" a specimen to a particular individual or other source) or about classification of the source of the specimen into one of several categories. With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source.

In relation to latent fingerprint comparison, the NRC report explicitly challenged the dominant method—that is, ACE-V (Huber, 1959).

> ACE-V provides a broadly stated framework for conducting friction ridge analyses. However, this framework is not specific enough to qualify as a validated method for this type of analysis. ACE-V does not guard against bias; is too broad to ensure repeatability and transparency; and does not guarantee that two analysts

> following it will obtain the same results. For these reasons, merely following the steps of ACE-V does not imply that one is proceeding in a scientific manner or producing reliable results (Haber & Haber, 2007; National Research Council, 2009).

The Committee also confronted and rejected the idea that fingerprint comparisons are free from error.

> Errors can occur with any judgment-based method, especially when the factors that lead to the ultimate judgment are not documented. Some in the latent print community argue that the method itself, if followed correctly (i.e., by well-trained examiners properly using the method), has a zero error rate. Clearly, this assertion is unrealistic, and, moreover, it does not lead to a process of method improvement. The method, and the performance of those who use it, are inextricably linked, and both involve multiple sources of error (e.g., errors in executing the process steps, as well as errors in human judgment).

The NRC report highlighted the absence of experiments on human expertise in forensic comparison (or pattern matching): "The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine its validity." Going further, it concluded that: "[t]his is a serious problem." The Committee recommended that US Congress fund basic research to help the forensic community strengthen their research foundations, develop valid and reliable measures of performance, and establish evidence-based standards for analyzing and reporting results. The Committee placed emphasis on addressing the limited research base, determining error rates, as well as understanding and reducing the effects of bias and human error. The NRC is not alone in its expressed concerns about forensic science used for the purposes of identification. Subsequent to the NRC report, two prominent inquiries in the US and the UK have released reports focused directly on fingerprint evidence. One report was produced by the Expert Working Group on Human Factors in Latent Print Analysis (EWGHF)—a large multidisciplinary collective jointly sponsored by the United States National Institute of Standards and Technology (NIST) and National Institute of

Justice (NIJ)—*Latent Print Examination and Human Factors*(2012). The other emerged from an inquiry conducted by Lord Campbell into problems with fingerprint evidence following the controversial McKie case in Scotland—*The Fingerprint Inquiry (SFI)* (Campbell, 2011). These reports, from jurisdictions generally regarded as leading forensic science providers, are again surprisingly critical in their responses to widely accepted identification practices.

Convened in 2008, in the shadow of the NRC inquiry, the EWGHF was tasked with undertaking a scientific assessment of the effects of human factors on latent print analysis (Woods et al., 2010). Specifically, the group was directed to evaluate current practices, to explain how human factors contribute to errors, and to offer guidance and recommendations. One way to obtain an impression of the Report's main thrust is to consider its recommendations, particularly those relating to the comparison of fingerprints and the expression of results. Relevant recommendations include:

> Recommendation 3.3: Procedures should be implemented to protect examiners from exposure to extraneous (domain-irrelevant) information in a case. Recommendation 3.7: Because empirical evidence and statistical reasoning do not support a source attribution to the exclusion of all other individuals in the world, latent print examiners should not report or testify, directly or by implication, to a source attribution to the exclusion of all others in the world. Recommendation 3.9: The federal government should support a research program that aims to: (a) Develop measures and metrics relevant to the analysis of latent prints; (b) Use such metrics to assess the reproducibility, reliability, and validity of various interpretive stages of latent print analysis; and (c) Identify key factors related to variations in performance of latent print examiners during the interpretation process. Recommendation 6.3: A testifying expert should be familiar with the literature related to error rates. A testifying expert should be prepared to describe the steps taken in the examination process to reduce the risk of observational and judgmental error. The expert should not state that errors are inherently impossible or that a method inherently has a zero error rate. Recommendation 9.1: Management should foster a culture in which it is understood that some

> human error is inevitable and that openness about errors leads to improvements in practice. Recommendation 9.5: The latent print community should develop and implement a comprehensive testing program that includes competency testing, certification testing, and proficiency testing.

The Scottish inquiry into the controversy surrounding the mistaken attribution of a latent print collected from a crime scene to Shirley McKie (a police officer) also generated a large, though perhaps less systemically oriented, report along with a series of recommendations (Cole & Roberts, 2012). Under the heading *The subjective nature of fingerprint evidence*, recommendations from the SFI included:

> Recommendation 1: Fingerprint evidence should be recognised as opinion evidence, not fact, and those involved in the criminal justice system need to assess it as such on its merits. Recommendation 2: Examiners should receive training which emphasises that their findings are based on their personal opinion; and that this opinion is influenced by the quality of the materials that are examined, their ability to observe detail in mark and print reliably, the subjective interpretation of observed characteristics, the cogency of explanations for any differences and the subjective view of "sufficiency." Recommendation 3: Examiners should discontinue reporting conclusions on identification or exclusion with a claim to 100% certainty or on any other basis suggesting that fingerprint evidence is infallible.

Beneath the heading "Fingerprint methodology," recommendations were particularly concerned with contextual bias:

> Recommendation 6: The SPSA [Scottish Police Services Authority] should review its procedures to reduce the risk of contextual bias.

From these independent inquiries, supported by a range of scientific studies and pre-existing scholarly critiques, several consensus themes emerge that are consistent with this proposal (Haber & Haber, 2007; Saks & Koehler, 2005; Saks & Faigman, 2008; Cole, 2005,

2010; Dror & Cole, 2010; Haber & Haber, 2009; Edmond & Roach, 2011) Most prominent are: confirmation about the lack of scientific support for contemporary fingerprint comparison practice and underlying assumptions (NRC Rec. 3, LPEHF 3.9 and SFI 2, respectively); the rejection of claims about an infallible method and a zero error rate (NRC p.142, LPEHF 6.3 and SFI 3); and, concern about equating a declared "match" with positive identification (NRC 3, LPEHF 3.7 and SFI 3).30 The reports place considerable emphasis on the need for research, standards derived from research (NRC 1, 7, 8, LPEHF 3.4, 3.6, 3.8, 3.9, 8.1), the need to attend to a range of potential biases, and the possibility of shielding analysts from some kinds of information (NRC 5, LPEHF 3.3 and SFI 6, 7, 8). The reports also recognize the need to present opinions derived from fingerprints in a manner that embodies their value and is simultaneously comprehensible to the tribunal of fact (NRC Rec. 2, LPEHF 5.1, SFI 64). The LPEHF Report (Rec. 4.3, 6.3, 9.1, and 9.2) directs attention to the need for examiners to be familiar with error rates, probabilities and statistics and the SFI (Rec. 82, 83) advocates the development of probabilities.

For the average reader—whether lawyer, judge or potential juror—all of this might come as something of a surprise. For, notwithstanding long reliance on fingerprint evidence, relatively little is known about the performance of fingerprint examiners or the value of their opinions. Contemporary investigative practices and reporting appear to fall well short of the recommendations and advice outlined in the independent reports. Concerns expressed by the NRC, EWGHF, and Lord Campbell, along with several notorious cases of fingerprint misattribution, raise serious (and as yet unresolved) questions about the forensic use of fingerprint evidence. There is, however, an indisputable need to reform the way fingerprint examiners work as well as the manner in which they express their opinions. Currently, there is a dearth of research. The necessary studies are often beyond the capabilities and competence of fingerprint examiners (and yet to be undertaken, or completed). Understandably, the training of fingerprint examiners is primarily oriented toward comparing fingerprints. Most do not have the methodological skills, funding, time, infrastructure, or experience with research techniques to mount scientific studies of human performance. Moreover, few fingerprint examiners, lawyers, or judges have the time, resources or expertise to track and evaluate extant studies, inquiries and reports, or respond to research as it emerges (Mnookin et al.,

2010). Consequently, changes to practices and reporting will require the ongoing assistance of research scientists. Research into expertise and complex sociotechnical systems is the domain of cognitive science and human factors. Researchers in these areas already have the infrastructure in place to conduct the requisite studies, and are well positioned to work with examiners to strengthen the field.

8.4.2 Emerging Studies

Scholarly criticisms and recent inquiries have already spawned a range of studies. The first studies focused on consistency and bias in expert decisions, but it is difficult to glean indicators on human matching performance from them (Wertheim et al., 2006b; Dror et al., 2011; Langenberg, 2009; Haber & Haber, 2006). Most of the research is currently in progress, though three studies have recently been published. In a controlled fingerprint matching experiment, Tangen et al. (2011) found that examiners incorrectly declared 0.68% of similar nonmatching prints as "match" (false positive errors)—compared to 55.18% for lay persons—and 7.88% of matching prints as "nonmatching" (false negative errors M. B. Thompson et al., 2013a; Tangen, 2013; Tangen et al., 2011).

In a similar experiment that made use of genuine crime scene prints, where the ground truth is uncertain, they found that examiners incorrectly declared 1.65% of similar nonmatching prints as "matching" (false positive errors)—compared to 55.73% for lay persons—and 27.81% of matching prints as "nonmatching" (false negative errors; M. B. Thompson et al., 2013b). In another controlled fingerprint matching experiment, Ulery et al. (2011) found that examiners incorrectly declared 0.1% of similar nonmatching prints as "matching" (false positive errors) and 7.5% of matching prints as "nonmatching" (false negative errors). These results demonstrate that qualified, court-practicing fingerprint examiners were far more accurate (and more conservative) than laypersons, and that the rate of false positive errors (i.e., incorrectly reporting that nonmatching fingerprints match) in these experimental matching tasks was around 1% and the rate of false negative errors (i.e., incorrectly reporting that matching fingerprints do not match) ranged from 8% to 28%. For criminal justice systems that have routinely relied upon fingerprint evidence for convictions and pleas, these

preliminary results should come as a great, if necessarily partial, relief. They suggest that fingerprint examiners have genuine expertise in *discriminating between prints that match and those that do not.*

In conjunction with the findings and recommendations in the various reports, these studies provide a platform upon which to begin reforming the way opinions about fingerprints are represented and used in legal settings. Our proposed guide to interpreting forensic science testimony begins to address some of the conspicuous deficiencies in contemporary fingerprint practice, especially in the reporting and explanation of results. This proposal attempts to take seriously some of the destabilizing epistemic and organizational problems raised in scholarly critiques and the recent authoritative and independent inquiries and reports.

We aim, with this proposal, to enhance legal performance by directing attention toward actual abilities, based on emerging evidence. It is clear that fingerprint identification cannot be regarded as an infallible "methodology" that is detached from human judgment (Cole, 2005; Tangen, 2013). Given the long history of claims about uniqueness, individualization, and a disembodied identification processes, examiners and their institutions should now begin to replace traditional practices and reporting with evidence-based claims that reflect actual capabilities (R. Garrett, 2009; Scientific Working Group on Friction Ridge Analysis Study and Technology, 2011a). Regardless of what forensic scientists do, criminal courts have a principled obligation to truth and justice (Ho, 2008). Courts, particularly those jurisdictions with a reliability-based admissibility standard, have an obligation to require forensic scientists to present their evidence in ways that embody actual capabilities. This requires evaluating reliability and conveying limitations clearly to the tribunal of fact.

In addition, we take seriously concerns, such as those recently voiced by the Law Commission of England and Wales, about the criminal trial and its limitations with expert opinion evidence (Edmond, 2012b; Law Commission, 2011). The historically accommodating response to fingerprint evidence, the few substantial challenges, and the vanishingly small number of appellate decisions suggest a legal reluctance (or inability) to appreciate the significance of problems with fingerprint evidence.

Through the provision of a guide, it is our intention to integrate some of the recommendations and emerging research to produce a serviceable tool to assist the legal regulation and use

of fingerprint evidence. The Guide is intended to help with the expression and interpretation of opinions about fingerprints by locating them within the appropriate research matrix. We envisage that a version of the Guide would be appended to expert reports prepared by fingerprint examiners, although we also envisage an updated Guide available through a publicly accessible repository. The Guide represents a pragmatic attempt to acknowledge and explain actual abilities as well as non-trivial limitations with fingerprint evidence. It is intended to recognize the existence of genuine expertise in comparison work, expose the weak decision-making framework and problem of extrapolation (i.e., the "leap" from match to identification), as well as address the historical reluctance to make appropriate concessions in reports and testimony.

8.5 Insights from Medicine: The Diagnostic Model

In modern diagnostic medicine, the accuracy of a test is inferred from controlled experiments. In home pregnancy testing, for example, a pregnancy test produces a result that reads "pregnant" or "not pregnant" based on the level of human chorionic gonadotropin (hCG) in urine used as a marker for pregnancy—which may or may not agree with the true state of the world. The validity, reliability, and accuracy of the test come from the aggregation of many controlled experiments. So, in a particular case (i.e., when a woman takes a pregnancy test) we can use this aggregated information to infer something about her true state.

For example, Figure 8.2A depicts the results from an experiment by Tomlinson et al. (2008) comparing the accuracy of six home pregnancy tests available over-the-counter. The numbers in Figure 8.2A represent groups of women who were pregnant or not and who took the Answer® home pregnancy test, which either resulted in a reading of "pregnant" or "not pregnant." One hundred and twenty pregnant women were given the Answer® home pregnancy test, 98 of the tests correctly read "pregnant" and 22 incorrectly read "non pregnant." Similarly, 120 women who were not pregnant were given the Answer® home pregnancy test, 2 of the tests incorrectly read "pregnant" and 118 correctly read "not pregnant."

A.

	Test said "pregnant"	Test said "not pregnant"	
Pregnant	98	22	120
Not pregnant	2	118	120
	100	140	

B.

	Expert said "match"	Expert said "no match"	
Match	409	35	444
No match	3	441	444
	412	476	

Figure 8.2: Pregnancy test results from Tomlinson et al. (2008; Panel A) and expert fingerprint matching results from Chapter 2 (Panel B).

If a woman purchases an Answer® home pregnancy test from the chemist, and tests herself, what could she conclude on the basis of the experiment by Tomlinson and colleagues? If the home pregnancy test read "pregnant" in this particular case, whether the woman is in fact pregnant (like the 98 for whom the test produced the correct reading) or whether she is not (like the 2 for whom the test produced the incorrect reading), we do not know. Similarly, if the home pregnancy test read "not pregnant" in this particular case, whether the woman is in fact pregnant (like the 22 for whom the test produced the incorrect reading) or whether she is not (like the 118 for whom the test produced the correct reading), we do not know. This uncertainty does not render the woman helpless; rather, the information could inform her interpretation of the test result (and the question of pregnancy).

We can apply the same diagnostic model to the experiment in Chapter 2 on the matching performance of court practicing fingerprint examiners. The results from this experiment are depicted in Figure 8.2B. A group of 37 qualified fingerprint examiners examined 444 pairs of fingerprints from the same person, 409 were correctly declared as a "match," and 35 were incorrectly declared "no match." Similarly, the examiners examined 444 pairs of fingerprints from different people, 3 were incorrectly declared as a "match," and 441 were correctly declared "no match."

If a juror hears an examiner give an opinion about whether two prints match (or not) in a criminal case, what could the juror conclude on the basis of the experiment in Chapter 2? If the examiner declared a "match" we do not know in this particular instance whether the prints are from the same source (like the 409 for which a "match" opinion was correct) or whether the prints are not from the same source (like the 3 for which a "match" opinion was

incorrect). Similarly, if the examiner said "no match" we do not know whether the prints are from the same source (like the 35 for which a "no match" opinion was incorrect) or whether the prints are not from the same source (like the 441 for which a "no match" opinion was correct). This uncertainty does not render the juror helpless; rather, the information could inform his interpretation of the examiner's opinion (and the question of source).

We suggest that an indication of performance (and error) in previous situations, (reasonably) similar to the particular analysis, provides potentially valuable information to those obliged to evaluate fingerprint testimony. This "statistical base rate" is general information. The juror can reason with this information to infer something about the particular case—to deduce the particular from the general. The juror can also use information about the particular case ("causal base rates") to temper these judgments, if they think the information is relevant. Judgments can be anchored to a plausible base rate and tuned by reasoning, informally, about the information specific to the particular case (Kahneman, 2011).

Broadly, the diagnostic model is an approach that offers information about similar situations in order to help decision-makers reason about the present case. The goal for a diagnostic model applied to forensic testimony is to give information to the legal participants to assist their decision-making around admissibility, challenges to evidence, instructions and warnings, and for the jury around the value of evidence and the ultimate conclusion. The goal is to provide information in a way that will help the jury to weigh the evidence, evaluate the arguments, and to judge the degree of belief warranted by the information presented (Pinker, 2003). Often this will involve information about general, or indicative, error rates and practical limitations. As in the diagnostic model, much of the information (i.e., scientific evidence) that can be presented will be based on general data from previous studies (i.e., from beyond the instant case) and the legal participants and fact-finder must reason and make inferences from the general to the particular case (Faigman, Monahan, & Slobogin, 2013).

Little is currently known about the types and forms of information that will assist triers of fact to make optimal decisions. The Guide is presented as a pragmatic intervention: an evolving compromise that endeavours to provide a diagnostic-style framework to improve forensic reasoning. An in-depth treatment of the diagnostic model applied to forensic

testimony is forthcoming, but our goal, in the first instance, is to help to ensure that expert evidence is presented in ways that are scientifically tenable. This involves embodying its known value and disclosing limitations in ways that help triers of fact to make sensible decisions about the forensic science evidence in a particular case.

8.6 A Guide to Forensic Testimony: Fingerprints

This section provides an example of what a guide for fingerprint evidence proffered for identification might look like at this stage. That is, the kind of information or caveats that ought to be included with the fingerprint examiner's report and testimony. It is an intentionally short document that places emphasis on brevity, comprehensibility and the goal of capturing both the considerable evidentiary potential as well as known limitations. Requiring ongoing revision—at least until there are sufficient studies to support a stable consensus—this preliminary version is based on the few scientific studies that have assessed the performance of fingerprint examiners in circumstances where conditions were deliberately controlled. In the remainder of this article we endeavour to unpack the Guide and some of the implications of the recent reports and emerging studies in ways that are sensitive to the criminal justice milieu, and especially the criminal trial.

A Guide to Forensic Testimony: Fingerprints

A decision about whether two fingerprints match or not is based on the judgment of a human examiner, not a computer.

There are several documented cases where an examiner has incorrectly said that two prints "match" when they actually came from two different people.

Laboratory-based experiments suggest that errors of this sort happen infrequently (around 1% of the time).

In practice, however, it is unknown how often examiners say that two fingerprints match when they actually come from two different people.

Without specific evidence, it cannot be known whether an error has occurred in a particular case.

For further information see `www.InterpretingForensicTestimony.com`

8.7 Conclusion

The proposed Guide represents a pragmatic attempt to address criticisms of fingerprint methodology and the way most comparisons are currently reported and explained. It is intended to begin to address and recognize the existence of genuine expertise in undertaking comparison work, the weak underlying decision-making framework, and the historical reluctance among examiners to make appropriate concessions in reports and testimony. If this approach is conceived as excessively empirical (or requires too much evidentiary support), we would remind the reader that for a century courts have allowed techniques to be misrepresented even though they could have been studied and improved. Along with the National Research Council of the United States National Academy of Sciences, the National Institute of Justice and the National Institute of Standards and Technology (US), Lord Campbell (Scottish Fingerprint Inquiry), and others, we are supporting the introduction of accountability mechanisms in forensic science reporting and testimony.

At the base of this proposal is a growing chorus of criticism about the kinds of studies and evidence that should underpin the forensic sciences—at least, when they are relied upon in criminal proceedings. We, along with others, believe that courts have an obligation to require evidence of ability—actual expertise and rates of accuracy—particularly about forensic science and medicine techniques in routine use. Empirical studies provide evidence of ability, they enable those evaluating the evidence to have a clearer idea of the value of evidence than is usually provided through cross-examination or judicial cautions, and the studies will often inform the manner in which experts should be allowed to express their opinions as well as the scope of their testimony. All of these can be observed in the recent experiments on the abilities and accuracy of fingerprint examiners.

The Guide, and particularly the focus on error rates, offers a useful means of assessing and regulating a range of comparison practices, especially those where the likelihood of generating useful probability rates seems remote or even unlikely. Even if fingerprint examiners and others develop probabilistic approaches, there will be a need to consider how error rates should be incorporated into the results, as the various reports recommend. For many other types of comparison practices (e.g., images, voices, ballistics, tool marks, bites, footprints, tire prints, pattern marks and so on), error rates will provide important insights into the value of the evidence—by highlighting the abilities of the witnesses—that enable judges (and lawyers) to determine the admissibility of opinions and their weight should they be sufficiently reliable for admission. Preliminary studies suggest that not all comparison and identification techniques will be as accurate as fingerprint identification.

We accept that, notwithstanding its empirical sensitivities, this is a pragmatic response or compromise. We also accept that others may have alternative, perhaps better, ideas about how we can respond to the frailties of latent fingerprint evidence. Our proposal is based on what we currently know, empirically, about latent fingerprint evidence in combination with the realization that investigators and courts are unlikely, and perhaps unable, to respond unilaterally to latent fingerprint evidence. No doubt many examiners, and others, will be concerned about even these modest, empirically inflected impositions. While acknowledging that the Guide and the imposition of an indicative error rate represents a considerable departure from historical assumptions and practices, there is no reason to continue to admit

and accept opinions about latent fingerprints in the conventional accommodating manner. Our proposal accepts that latent fingerprint evidence is potentially powerful evidence of identity, and we believe that the Guide provides a compromise that reflects current knowledge and abilities, as well as what we now know about the limitations of fingerprint evidence (as well as the limitations of trials and appeals).

The legal system should not avert its eyes to mainstream scientific consensus around frailties and errors in the way comparison sciences are practiced and reported. Disregarding scientific consensus threatens the legitimacy of legal institutions and undermines their ability to deliver accurate verdicts and, simultaneously, justice (Edmond & Roach, 2011).

Chapter 9

Conclusion

9.1 Preface

Before I embarked on this program of research, little was known about the accuracy of fingerprint experts, the performance of experts relative to novices, the cognitive processes underlying fingerprint expertise, or the ways in which experts could legitimately testify in court. I approached these research questions from the perspective of cognitive science. The task of crime scene fingerprint identification is clearly a problem of human judgment and decision making. Unlike other areas of forensic science, such as analytic chemistry, it is the *human* who is the instrument of analysis. The calls for error rates and indications of accuracy from authoritative sources equated to calls better understand human performance—a task that falls squarely in the domain of psychology—and forensic practitioners were not well positioned to address the problem.

Although there was little research on fingerprint expertise in itself, there was an array of disparate literature from which to draw, such as general expertise and skill acquisition, exemplar theory of categorisation, familiar and unfamiliar face recognition, signal detection theory, and dual-process models of cognition. In the series of experiments presented here, consistent themes emerged about the factors that affect matching accuracy, and about the ways in which experts process prints differently to novices. As can be seen in Figure 9.1, I have explored four aspects of fingerprint matching expertise: **PART 1** - ESTABLISHING

EXPERTISE, **PART 2** - DEPICTING EXPERTISE, **PART 3** - NATURE OF EXPERTISE, and **PART 4** - EXPRESSION OF EXPERTISE. I will first speak generally about the each of the parts and close by discussing the thesis as a whole.

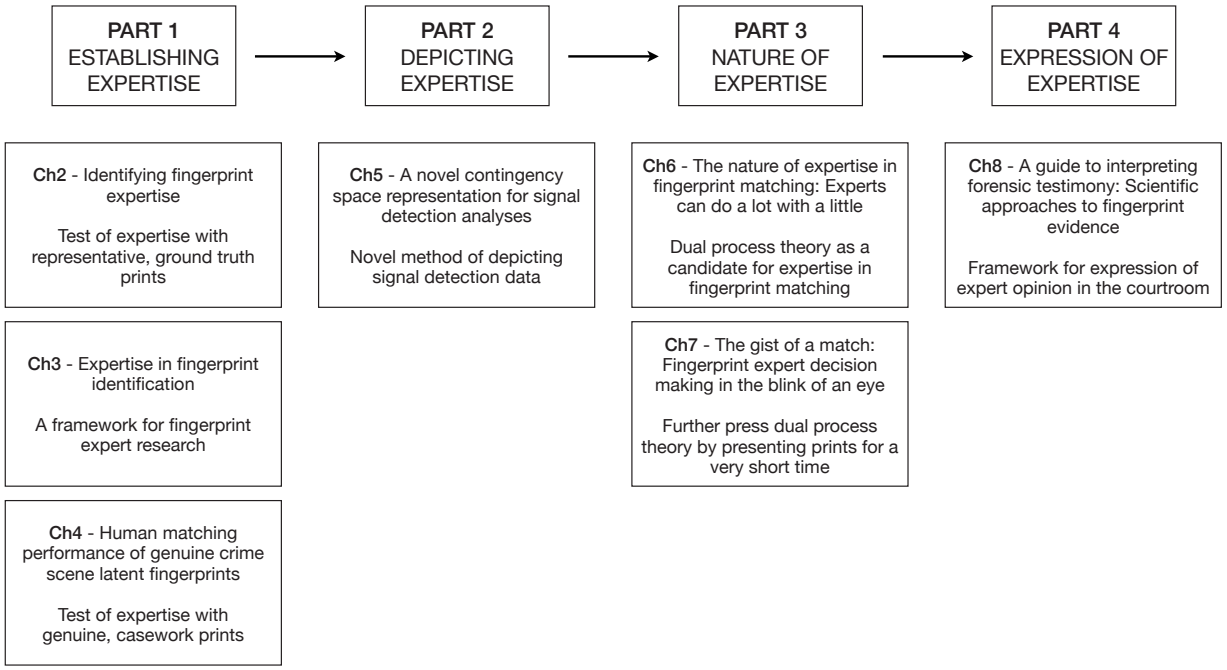

Figure 9.1: Conceptual diagram of the thesis in four parts: (1) Establishing expertise, (2) Depicting expertise, (3) Nature of expertise, and (4) Expression of expertise. Each of the four parts includes at least one thesis chapter.

9.2 PART 1 - ESTABLISHING EXPERTISE

In PART 1 of this thesis—ESTABLISHING EXPERTISE—I explored expertise in fingerprint identification. I attempted to find evidence for expert-novice performance differences in fingerprint matching, and to see where those differences lie. In Chapter 2—Identifying Fingerprint Expertise—I tested the matching accuracy of expert and novice examiners using pairs of representative, ground truth prints that either match, don't match, or don't match but are highly similar. I set out to determine whether fingerprint experts are any more accurate at matching prints than novices, and to get an idea of how often experts make errors of the sort that could allow a guilty person to escape detection compared with how often

they make errors of the sort that could falsely incriminate an innocent person. Qualified court-practicing fingerprint experts were exceedingly accurate compared with novices, and they tended to err on the side of caution by making errors of the kind that would fail to identify a criminal rather than provide incorrect evidence to the court. Further, an examiner's expertise seems to be situated, not in their ability to match prints per se, but in their superior ability to identify highly similar, but nonmatching fingerprints as such.

In Chapter 3—Expertise in Fingerprint Identification—I expanded on the results and interpretation of first experiment and presented a framework for fingerprint expertise research. I argued that fidelity, generalizability, and control must be balanced to answer important research questions; that validity, proficiency, and the competence of fingerprint examiners are best determined when experiments include highly similar print pairs where the ground truth is known; that a signal detection paradigm can be employed to separate the two ways of being right and the two ways of being wrong, and to isolate accuracy from response bias; that the most appropriate comparison group to demonstrate expertise should be novices who have no training with fingerprints whatsoever; that determining error rates with black box studies may be unnecessary at best and ineffective and inefficient at worst, and unless one can demonstrate that a particular qualifier will systematically affect accuracy, the default should be to report accuracy at the broader level.

In Chapter 4—Human Matching Performance of Genuine Crime Scene Latent Fingerprints—I tested the matching accuracy of expert, trainee, and novice examiners using pairs of genuine, casework prints that either match, don't match, or don't match but are highly similar. I increased the fidelity of the discrimination task (i.e., the resemblance of the discrimination task to actual casework) by using genuine crime-scene latents (and their matched exemplars) from police training materials, compiled from casework. Again, qualified examiners were more accurate (and more conservative) than novices and were better as discriminating prints that look highly similar but come from two different people. Intermediate trainees—despite their lack of qualification and average 3.5 years experience—performed about as accurately as qualified experts who had an average 17.5 years experience. New trainees—despite their 5-week, full-time training course or their 6 months experience—were not any better than

novices at discriminating matching and similar nonmatching prints, they were just more conservative.

Taken together, these experiments show that, compared to novices, experts are better able to discriminate matches from nonmatches, particularly when the nonmatches are highly similar. Why might this be? A theoretical explanation is made easier if we consider the task as a *categorisation* task, rather than an *identification* task. Experts need to look at a pair of prints and classify them as "Match" or "No match." So, in this case, there are only two categories in which the prints can fall. (In practice, however, experts have make various classifications which vary from department to department. Other categories include, for example, "Insufficient" [there is not enough information to make a judgment either way], and "Inconclusive" [I am unwilling to make a judgment either way]). Similarly, people categorise everyday situations, scenes, and objects as "Chair," "Dog," "Male," etc., with ease. And experts categorise chest x-rays, skin disorders, and aircraft as "Abnormal," "Measles," and "Friendly," etc., also with ease. The exemplar theory of categorisation posits that categorisation is easy for people who have acquired a large number of exemplars from the various categories, because this experience allows them to categorise new items based on the similarity of the new item to the previously encountered exemplars (Brooks, 1978, 2005). Learning the structure and acceptable variation of categories can develop without intention or formal training, especially in the presence of accurate and timely feedback (Hogarth, 2001; Kahneman & Klein, 2009).

Again, in framing fingerprint identification as a classification task, fingerprint experts have learned many of the ways that matching and nonmatching prints can vary—the acceptable variation between and within the two categories. They have seen many instances of prints that look the same, but come from two different people, and many instances of prints that look different, but come from the same person. Put simply, experts have lots of experience with prints that match and prints that don't match, and they receive feedback from extraneous case information (e.g., "The guy later confessed."), and from colleagues.

This vast experience allows experts to resolve information in a print: to correctly regard ambiguous information that is more consistent with *within-source* variability as a "match," and correctly regard ambiguous information that is more consistent with *between-source*

variability as a "nonmatch." An ambiguous mark on a fingerprint, for example, can be regarded as signal (i.e., as evidence of a "match"), or it can be disregarded as noise (i.e., as evidence of a "non-match"). This kind of process is undoubtedly operating in novices too, but the ambiguity cannot be sufficiently resolved unless the examiner has accumulated enough matching and nonmatching exemplars in memory to point to one direction or the other. One clear result of this vast experience is the experts' capacity to disregard, to 'see through', the ambiguity and surface structure of similar prints and discriminate them accurately.

It seems that some combination of training and the daily comparison of untold numbers of fingerprints leads to an uncanny ability to match fingerprints to their source. Experts are drawing on an entire career of experiences in making their decisions, as well as their training in fundamentals of fingerprint impressions to understand the "behavior" of minutiae. Experts, likely implicitly, understand the structure, regularities, and acceptable variation of fingerprint impressions. Viewing hundreds of prints, day in day out, gives opportunities to appreciate the multidimensional space in which prints can appear. During training and casework, examiners spend their days submitting crime scene prints to a computer algorithm that searches enormous print databases, and returns a list of several candidates (prints from different people that, according to the algorithm, are similar to the crime scene print). This process gives examiners the opportunity to see prints that come from the same person but that appear quite different, and the opportunity to see two prints that come from two different people but that appear very similar. This experience might give them a feel for the variation in pairs of prints that match and those that don't, and could be one explanation for the conservative bias of fingerprint experts. When they have an appreciation for how similar two nonmatching prints can look, they may adjust their decision criterion accordingly (to be more conservative).

Then again, new trainees (in Chapter 4) displayed a comparably high conservative response bias to qualified experts. An alternative explanation for the conservative bias could be the "error-free" culture of the discipline—the policy in many departments is that if you make a false alarm error you will lose your job (Woods et al., 2010). The extent of this culture was made clear from a recent case in which a fingerprint examiner testified about a fingerprint match and then spoke about errors:

> "It's interesting to note that any fingerprint individualization that's made, whether it be at the Canadian Police College or throughout my apprenticeship, if there is any errors made on a fingerprint, it's immediate withdrawal or removal from the program. There's no errors allowed in fingerprint identification. That continues today. There is no errors permitted in fingerprint identification. . . I've never made an error." (R. v. Bornyk, 2013 BCSC 1927).

Further, examiners are keenly aware of the weight that fingerprint evidence holds in court, and they may be conscious of the fact that a false alarm is unlikely to be picked up by the criminal justice system. The knowledge that committing a false alarm at work could easily result in falsely convicting an innocent person, may contribute to the conservative bias of trainees and experts.

9.3 PART 2 - DEPICTING EXPERTISE

In PART 2 of this thesis—DEPICTING EXPERTISE—I explore alternative methods for communicating and illustrating sets of signal detection data, and the results of experiments from PART 1 in particular. In Chapter 5—A Novel Contingency Space Representation for Signal Detection Analyses—I describe a method of depicting signal detection data. I present a methodology of plotting data from a two-alternative, forced choice matching task against a background of contingency tables that represent all possible outcomes from such an experiment. This contingency space representation allows one to depict sensitivity, response bias, chance, and relative performance. I argue that the method makes the theoretical independence between sensitivity and response bias clearer, makes experimental results easier to interpret through the use of natural frequencies, and is especially useful for comparing results between experimental conditions. I suggest that traditional representations can be difficult to interpret for those not well versed in Signal Detection Theory, and that there may be a complementary representation that better depicts the relationship between sensitivity and response bias. I hope that thinking carefully about the most effective way to present scientific results to non-scientists will go some way to fostering good working relationships

between researchers and forensic professionals and help promote a research culture in forensic science (Mnookin et al., 2010).

9.4 PART 3 - NATURE OF EXPERTISE

In PART 3 of this thesis—NATURE OF EXPERTISE—I explored the cognitive processes that might account for the superior performance of expert fingerprint examiners. In Chapter 6—The Nature of Expertise in Fingerprint Matching: Experts Can Do a Lot with a Little—I evaluated dual-process theory as a candidate to explain the nature of expertise in fingerprint matching. In Experiment 1, I tested experts and novices to see if the classic face inversion effect was also present in fingerprint matching expertise. I did not find evidence for an inversion effect, but expert performance with highly artificially noisy prints, however, was impressive. In Experiment 2, I tested the short term memory of experts and novices by separating fingerprint pairs in time by a few seconds. Experts were better than novices at discriminating prints that were spaced in time and, again, experts were far better as discriminating similar, nonmatching prints. In Experiment 3, I tested the long term memory of experts and novices by asking them to learn a set of fingerprints to be recognised a few minutes later. There was no difference in long term memory accuracy between experts and novices, and both groups performed around the level of chance. In Experiment 4, I tested the ability of experts and novices to discriminate prints by presenting them on screen for only a few seconds. I found that experts could accurately discriminate prints when presented for two seconds, and the largest difference between experts and novices was on similar, nonmatching prints.

A hallmark of expertise is the ability to accurately perform a domain relevant task quickly. In Chapter 7—The Gist of a Match: Fingerprint Expert Decision Making in the Blink of an Eye—I explored the limits of rapid expert decision making. I found that experts were more accurate than novices overall, and experts had a much better idea about whether a pair of prints match or not in a rapid period of time. With such short presentation times (i.e., from 250 to 2000ms), there is little time to engage in careful, deliberate analysis of the minutiae in a fingerprint image in order to make accurate decisions. These findings suggests

that fingerprint experts are capable of making accurate decisions when the amount of visual information in the prints is dramatically decreased. It is clear that, through experience, experts can learn the regularities of matching and nonmatching prints, and rapidly compare new prints to memory in order to make accurate judgments.

The findings above are in stark contrast to the common and consistent claims in formal training, textbooks, and courtroom testimony: that fingerprint identification is a "scientific process" that requires careful, thorough analysis in order for judgments to be accurate (Busey & Parada, 2010; Busey & Vanderkolk, 2005; Cole et al., 2008). I found, however, that expert performance is impressive when the amount of information is severely limited. Experts can accurately discriminate matching and nonmatching prints that are artificially noisy, that are spaced in time, and that are flashed on screen only briefly. These findings suggest that non-analytic processing plays a key role in expert fingerprint matching. It seems that experts develop a fast and accurate method of fingerprint matching based on their vast experience and familiarity with fingerprints. Emerging evidence supports this view. For example, the size of experts' pupils—an indicator of the extent of analytic cognitive effort (Kahneman & Beatty, 1966) are smaller than novices' when they are identifying fingerprints (Laukkonen, 2012), and expert accuracy is not affected by high cognitive load (Laukkonen, Tangen, Baird, & Eva, 2013).

These findings are in opposition to the notion that careful and deliberate analysis is the basis of the work that expert fingerprint examiners do. Pattern recognition is clearly important and accounts for a significant portion of the variance in superior expert performance. Fingerprint experts through experience—just as in many other areas of genuine expertise—have built up a repository of instances that they can draw on to make judgments about novel instances. This conclusion is at odds with general wisdom in fingerprint identification practice, and at odds with formal training and the claims and explanations that are offered in court during expert testimony. The implications are far reaching.

If it is the case that experience with prints, that vary in a multitude of ways, is what leads to superior expert performance, then are examiners being trained most effectively? Current fingerprint identification training programs often focus on formal classification and identification rules that date back several decades. Examiners are trained to classify prints

into categories (e.g., loops, arches, and whorls) and use minutiae (e.g, forks, ridge endings, and lakes) to individualize a fingerprint. The findings above, however, suggest that training focused on exposure to many varied instances of matching and nonmatching prints (i.e., to the full range of between and within variation among fingerprints), coupled with accurate and corrective feedback, would be more efficient than training based on formal rules. The training program required to become a qualified expert in an Australian police department is currently five years, minimum. Training focused, at least partly, on the non-analytic basis of fingerprint expertise could conceivably increase the efficiency of training programs without compromising performance standards—we could turn novices into experts more quickly and efficiently. These systems ought to be designed to make it easy for examiners to gain lots of experience with varied prints, and ought to promote a blame-free culture in which people can receive immediate, corrective, and accurate feedback about their performance. The best way to design a system that promotes the accumulation of a multitude of instances with corrective feedback, is an open question. Relatedly, given that much of fingerprint expert processing is automatic and unconscious, should they have to justify the basis of their decisions in court? It is unlikely that experts will be able to accurately articulate the basis of a decision that is largely non-analytic.

Although non-analytic processing is important for fingerprint matching, my results also indicate that non-analytic processing alone is not sufficient to achieve maximum performance. I found, for example, that experts generally do better when they have a chance to see the prints for longer. (This is true for short presentation times at least; I would expect performance gains to plateau rather quickly). Put simply, I found that overall expert accuracy moves from 70% with a short presentation time to 90% with a long presentation time, and similar performance gains with clear versus noisy prints. It is likely that slow, analytic processing is also important in fingerprint matching, and both kinds of processing will interact, in some way, to drive superior expert performance. Doing the flip of this thesis (promoting analytic processing rather than non-analytic processing), could form the basis of a complementary program of research. One could tease apart the unique contributions of analytic and non-analytic processing, again to inform training, testimony, and basic theory about the development of expertise.

Further, and taking a systems-level approach, relying solely on non-analytic processing is unlikely to be optimal. Human factors research in areas such as healthcare and aviation, shows that systems that are designed to make errors difficult—by, for example, making use of checklists (Degani & Wiener, 1993), and through thoughtful team communication (Kanki, Helmreich, & Anca, 2010)—is important for building resilient systems. The practice of fingerprint identification often involves the "verification" of decisions by a fellow examiner, and processes for documenting the basis of decisions and the steps taken from crime scene to courtroom. Even though experts could rely on non-analytic processing for accurate decisions, it would be unwise to neglect the physical, individual, team, organisational, and societal factors that will influence the resilience or fragility of a system (Expert Working Group on Human Factors in Latent Print Analysis, 2012; Woods et al., 2010; Vicente, 2004). But we now have an indication of the cognitive processes and mechanisms that are operating in fingerprint identification and the factors that affect matching accuracy, and we can use this understanding to provide a foundation for evidence-based testimony.

More generally, fingerprints are a complex pattern for which people have lots of experience, like faces. Dimension reduction has been proposed as a model for how people learn the structure of various complex categories (e.g., Burton, Bruce, & Hancock, 1999). In this model, experience with a particular class of stimuli leads people to extract the primary dimensions of variation in a category and use these dimensions to categorise subsequent stimuli. In other words, when people gain experience with a particular class of stimuli they become sensitive to the main dimensions of variation that are important for distinguishing the stimuli from one another. I have suggested that superior expert fingerprint matching performance arises from the accumulation of instances with feedback, but it is still not clear how those instances are encoded and stored (Brooks, Squire-Graydon, & Wood, 2007). Given its success with faces, dimension reduction could provide a useful starting point for understanding the structure of fingerprints.

9.5 PART 4 - EXPRESSION OF EXPERTISE

What forensic scientists should say, or should be allowed to say, about the results of forensic comparisons has been a contentious issue in the legal and forensic science communities. When, if ever, should forensic scientists claim to have determined with certainty that two items have a common source? When, if ever, should they express an opinion about the probability that two items share a common source? Should forensic scientists be allowed, or even required, to present statistics on the probability of coincidental matches? Should they be allowed or required to present statistics about error rates? And if statistical evidence is presented, how should such statistics be presented? Should they be expressed as frequencies, as likelihood ratios, or in some other format? And may, or should, an expert witness provide information or express opinions about the value of such statistics for drawing inferences regarding a contested matter at trial?

In PART 4 of this thesis—EXPRESSION OF EXPERTISE—I explored ways that fingerprint examiners can communicate (express) their opinions in ways that are responsive to recent epistemological critiques and recent empirical findings. In Chapter 8—A Guide to Interpreting Forensic Testimony: Scientific Approaches to Fingerprint Evidence—I proposed a framework for the expression of expert opinion in the courtroom. The framework is based on the medical diagnostic model where the validity, reliability, and accuracy of the test come from the aggregation of many controlled experiments, and which offers information about similar situations in order to help decision-makers reason about the present case. I suggest that an indication of performance (and error) in previous situations, (reasonably) similar to the particular analysis, provides potentially valuable information to those obliged to evaluate fingerprint testimony.

Part of the difficulty we face in addressing the questions above, is that the various ways forensic scientists might characterize their findings, and their strengths and weaknesses, have not been evaluated. There is a range of possible approaches to presenting forensic evidence: individualization, match, non-exclusion + random match probability, non-exclusion + diagnostics, likelihood of source, and likelihood ratio. “Non-exclusion + diagnostics” is the approach that I advocated for the various reasons described in Chapter 8. But we are still

left with the problem of how to decide amongst these options. How does one choose among these approaches for the evaluation and expression of evidence? Will there be a universal best framework that should be applied to all kinds of forensic evidence? Will the appropriate framework depend on the level of current knowledge and empirical support for each kind of forensic evidence/discipline?

A possible solution could be to design a framework for evaluating the various approaches to forensic expression, that could help determine the most advantageous approach—that is, the approach that currently best facilitates reliable, accurate, and effective communication of forensic evidence to a trier of fact. I can think of five ways in which expressions of forensic evidence could be evaluated: epistemic support, scientific/empirical support, legal support, explanatory power, intelligibility. Each of these areas could have high or low, strong or weak support.

- First, is the approach epistemically supported (Cole, 2009)? Is it possible to determine whether the claim is true? Can it be validated? Epistemology is the theory of knowledge, especially with regard to its methods, validity, and scope, and the distinction between justified belief and opinion. Epistemology dictates the claims one can and cannot make, or the claims that cannot be supported even before empirical or statistical data. Is it possible to know the answer to this question? Is the knowledge required to make a certain claim so burdensome that it renders the claim practically impossible? To understand the epistemic support for an approach, we can ask if it is testable, if hypothesis be generated and will lead to obtaining analyzable results, and if it is falsifiable.

- Second, is the approach scientifically supported (Mnookin, 2008a)? Has the approach been validated? Does the approach measure what it purports to measure; can the output of a framework be regarded as accurately describing ground truth? Is the approach reliable? Does the approach yield the same or compatible results across repeated applications and under different conditions? Measures of validity and reliability can give an indication of the scientific strength of an approach.

- Third, is the approach legally supported? In assessing legal support, we can ask if an approach satisfies Daubert, the confrontation clause, evidentiary standards Rule 702, and fits within a trial framework.

- Fourth, does the approach offer powerful explanations? Explanatory power is the ability of a theory to effectively explain the subject matter it pertains to. One theory has greater explanatory power than another if it offers more details about we should expect to see (i.e., whether the defendant is the source of the specimen) and what we should not, and more facts and details of causal relations are accounted for. In assessing explanatory power, we can ask if an approach captures what the jury needs to know, reduces uncertainty in the evidence, and is helpful.

- Fifth, is the approach intelligible? Can the trier of fact make sense of what is said (B. L. Garrett & Mitchell, 2013; W. C. Thompson, Kaasa, & Peterson, in press)? Is the approach strong against subtle variations in particular instances of the testimony (i.e., do people still understand the testimony if it is pitched in different ways by different people)? To measure intelligibility, we could consider whether people's beliefs about the testimony are consistent and robust across situations, and the extent to which their beliefs correspond with normative models (Martire, Kemp, Watkins, Sayle, & Newell, 2013). Expressions that are robust to intelligibility despite incidental changes in expression, such as which expert testifies and what explanations they give could be favoured.

Taken together, an approach could be epistemically, scientifically, and legally supportable but arguably incomplete if it does not tell the trier of fact what they need to know, and every permutation in between. We could find, through philosophical inquiry, that individualization expressions have weak epistemic support and likelihood ratio expressions have strong epistemic support. We could find, through empirical research, that people make good decisions (more reliable, accurate, in line with normative models, etc.) when given individualization expressions but make bad decisions (unreliable, inaccurate, in opposition to normative models, etc.) when given likelihood ratio expressions. Here, individualization expressions are more intelligible but not epistemically supported, and likelihood ratio expressions are less intelligible

but are epistemically supported. If we—the field, the legal system, society, etc.—had to decide between these two options, we would have to trade-off intelligibility and epistemic support. One may argue that, if there is no epistemic support then whether the expression is intelligible is of no consequence. Another may argue that an expression with strong epistemic support is practically useless if it is unintelligible. Of course, I think that the diagnostic approach proposed in Chapter 8 is the most appropriate model for testifying to forensic evidence, based on what is currently known (and knowable), but more research into legal support and intelligibility is needed (I am currently doing empirical research on the latter). Where the responsibility lies for deciding the "best" approach—whether it be the courts, the academy, individual experts, professional societies, etc.—is an open and difficult question.

9.6 Final Thoughts

This program of research was about determining the factors that affect matching accuracy, to better understand the development of expert forensic identification, to inform training, and to provide an empirical basis for expert testimony in the courtroom. Little was known previously about the nature and development of fingerprint expertise and, therefore, the best way to turn novices into experts. Little was known previously about the factors that affect matching accuracy and, therefore, what experts can legitimately testify to in court.

It is tempting to think that when hearing the testimony of a fingerprint examiner in court, we need only be concerned with evidence about the particular prints in question, or with a particular opinion expressed by the examiner. Valid inferences about the true state of affairs, however, are difficult when only a single data point is available. That is, the accuracy of a particular fingerprint identification cannot be known. As an alternative, we can look for evidence in the aggregate, about performance in the past and about the factors that might affect the strength of the evidence in the particular case. We can think of a fingerprint comparison as a diagnostic test by a human (assisted by tools, technology, etc.) that produces an opinion. In order to make a judgment about whether to "believe" the examiner or not, or how much weight to give their opinion, we need to know something about performance in the past and about factors that affect performance.

In this thesis, I have attempted to generate *general* evidence about the matching accuracy of fingerprint examiners and the factors that affect their performance. I have gathered empirical evidence about the relative performance of experts and novices, the source of identification errors, the factors that influence performance, and the nature of expertise in fingerprint matching. We have strong evidence that qualified examiners are more accurate and more conservative than novices; we have good evidence that errors are more likely to occur on prints from database searches, which are highly similar but nonmatching; we have a reasonable idea about how performance changes as people move from novice to expert; and we know that experts can discriminate prints in noise, when spaced by a short time, when presented for two seconds and even when presented in the blink of an eye. These findings indicate that experts make use of non-analytic processing when identifying prints and can get by (i.e., can perform more accurately than novices) when information is sparse—experts can do a lot with a little. In designing the experiments of this thesis, I have tried to balance fidelity, generalisability, and control in order to simultaneously inform basic theory, training, and expert testimony.

This general empirical evidence could form the basis of legitimate expert testimony (Cole, 2007). What we now know about expert fingerprint matching performance, and the factors that influence performance, could be used to make inferences about performance in the wild, and to help triers of fact reason about forensic evidence. As such, I have proposed a guide in which empirical evidence from highly controlled, but representative experiments can be used by a trier of fact to make inferences about a particular case. Further evidence that comes to light about the factors that influence performance could be used to support the claims of expert witnesses in court. For example, if empirical evidence from aggregate experiments shows that fingerprint examiners can gauge the intention of a perpetrator based on the orientation and placement of their prints (as has been claimed), then examiners could legitimately testify on this capability in a particular case. If, on the other hand, the evidence shows that examiners cannot gauge intent, then examiners could not legitimately testify on this point in any particular case.

The list of factors that could potentially influence expert matching performance is long. Fatigue, lifting agents, distractions, interruptions, sleep deprivation, time constraints,

verification, surface types, lifting agents, rotation, distortion, algorithms, etc., may or may not affect performance. Research on eyewitnesses testimony, for example, is decades ahead, and we now have a very good idea about the factors that affect recognition of unfamiliar faces. The depth of our understanding is made clear by a recently proposed model of eyewitness identification jury instructions (Loftus, Francis, & Turgeon, 2012). The detailed list of factors known to affect accuracy ranges from general (given in all cases) to specific, and include length of observation time, distance, lighting, stress, time elapsed, confidence, and line up procedure, to name but a few. Not surprisingly, the proposed model spans several pages and is in stark contrast to the equivalent for fingerprint identification (*The Guide* - Chapter 8) which is four *lines* long. Further programs of research, like this one, on the factors that affect fingerprint matching accuracy and performance, will serve to increase our confidence about the legitimacy of claims made by expert witnesses in court, if, of course, those claims are based on the best available empirical evidence.

References

Alpert, G. P., & Noble, J. J. (2009). Lies, true lies, and conscious deception police officers and the truth. *Police Quarterly*. doi: 10.1177/1098611108327315

Berner, E. S. E., & Graber, M. L. M. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5).

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science*, *23*(10), 1208–1214.

Brinberg, D., & McGrath, J. E. (1985). *Validity and the research process*. California: Sage Publications.

Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorisation*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Brooks, L. R. (2005). The blossoms and the weeds. *Canadian journal of experimental psychology = Revue canadienne de psychologie expérimentale*, *59*(1), 62–74.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, *120*(3), 278.

Brooks, L. R., Squire-Graydon, R., & Wood, T. J. (2007). Diversion of attention in everyday concept learning: Identification in the service of use. *Memory & Cognition*, *31*(1), 1–14.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. Univ of California Press.

Bukach, C. M., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: the power

of an expertise framework. *Trends in Cognitive Sciences*, *10*(4), 159–166.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, *23*(1), 1–31.

Busey, T. A., & Dror, I. E. (2010). Special Abilities and Vulnerabilities in Forensic Expertise. In *Fingerprint sourcebook.* Washington DC: National Institute of Justice Press.

Busey, T. A., & Parada, F. J. (2010). The nature of expertise in fingerprint examiners. *Psychonomic Bulletin & Review*, *17*(2), 155–160.

Busey, T. A., & Vanderkolk, J. R. (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research*, *45*(4), 431–448.

Busey, T. A., Yu, C., Wyatte, D., Vanderkolk, J., Parada, F., & Akavipat, R. (2011). Consistency and variability among latent print examiners as revealed by eye tracking methodologies. *Journal of Forensic Identification*, *60*(1), 61–91.

Campbell, A. (2011). *The Fingerprint Inquiry Report.* Edinburgh, Scotland: APS Group Scotland.

Carle, D. (2011, January). *Leahy Proposes Landmark Forensics Reform Legislation.* Retrieved from `http://www.leahy.senate.gov/press/leahy-proposes-landmark-forensics-reform-legislation/`

Carpenter, R. H. S. (1988). *Movements of the eyes.* London: Pion Ltd.

Champod, C., & Evett, I. W. (2001). A probabilistic approach to fingerprint evidence. *Journal of Forensic Identification*.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. Ericsson, N. Charness, & R. R. Hoffman (Eds.), *The cambridge handbook of expertise and expert performance.* New York, NY: Cambridge University Press.

Cho, A. (2002a). Fingerprinting doesn't hold up as a science in court. *Science*, *295*(5554), 418–418.

Cho, A. (2002b, March). Forensic science. Judge reverses decision on fingerprint evidence. *Science*, *295*(5563), 2195–2197.

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological

science and public policy. *Perspectives on Psychological Science*, *7*(3), 238–259.

Cole, S. A. (2002). *Suspect Identities: A History of Fingerprinting and Criminal Identification*. Harvard University Press.

Cole, S. A. (2005). More than zero: Accounting for error in latent fingerprint identification. *Journal of Criminal Law & Criminology*, *95*, 985–1078.

Cole, S. A. (2007). Toward evidence-based evidence: supporting forensic knowledge claims in the post-Daubert era. *Tulsa Law Review*, *43*(2), 263–284.

Cole, S. A. (2009). Forensics without uniqueness, conclusions without individualization: the new epistemology of forensic identification. *Law, Probability and Risk*, *8*(3), 233–255.

Cole, S. A. (2010). Who speaks for science? A response to the National Academy of Sciences Report on forensic science. *Law, Probability and Risk*, *9*(1), 25–46.

Cole, S. A., & Roberts, A. (2012). Certainty, individualisation, and the subjective nature of expert fingerprint evidence. *Criminal Law Review*, *11*, 824–849.

Cole, S. A., Welling, M., Dioso-Villa, R., & Carpenter, R. (2008). Beyond the individuality of fingerprints: a measure of simulated computer latent print source attribution accuracy. *Law, Probability and Risk*, *7*(3), 165–189.

Degani, A., & Wiener, E. L. (1993). Cockpit checklists: Concepts, design, and use. *Human Factors*, *35*(2), 345–359.

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General*, *121*(3), 275–277.

Drew, T., Evans, K. K., Vo, M. L. H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images? *Radiographics*, *33*(1), 263–274.

Dreyfus, H. L., & Dreyfus, S. E. (2005). Peripheral vision expertise in real world contexts. *Organization studies*, *26*(5), 779–792.

Dror, I. E. (2011). The paradox of human expertise: Why experts get it wrong. In N. Kapur (Ed.), *The paradoxical brain* (p. 177). Cambridge University Press.

Dror, I. E. (2012). Letter to the editor-combating bias: The next step in fighting cognitive and psychological contamination. *Journal of Forensic Sciences*, *57*(1), 276–277.

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.

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification*, *56*(4), 600–616.

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.

Dror, I. E., & Cole, S. A. (2010). The vision in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review*, *17*(2), 161–167.

Dror, I. E., & Mnookin, J. L. (2010). The use of technology in human expert domains: Challenges and risks arising from the use of automated fingerprint identification systems in forensic science. *Law, Probability and Risk*, *9*(1), 47–67.

Dror, I. E., Peron, A. E., & Hind, S. L. (2005). When emotions get the better of us: the effect of contextual topdown processing on matching fingerprints. *Applied Cognitive Psychology*, *19*(6), 799–809.

Dror, I. E., & Rosenthal, R. (2008). Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences*, *53*(4), 900–903.

Dror, I. E., Wertheim, K., Fraser-Mackenzie, P., & Walajtys, J. (2012). The impact of human-technology cooperation and distributed cognition in forensic science: Biasing effects of AFIS contextual information on human experts. *Journal of Forensic Sciences*, *57*(2), 343–352.

Edmond, G. (2008). Specialised knowledge, the exclusionary discretions and reliability: Reassessing incriminating expert opinion evidence. *University of New South Wales Law Journal*, *43*(2), 1–54.

Edmond, G. (2011). Actual innocents? Legal limitations and their implications for forensic science and medicine. *Australian Journal of Forensic Sciences*, *43*(2), 177–212.

Edmond, G. (2012a). Advice for the courts? Sufficiently reliable assistance with forensic science and medicine (Part 2). *The International Journal of Evidence & Proof*, *16*(3), 263–297.

Edmond, G. (2012b). Is reliability sufficient? The Law Commission and expert evidence in international and interdisciplinary perspective (Part 1). *The International Journal of Evidence & Proof*, *16*(1), 30–65.

Edmond, G., & Roach, K. (2011). A contextual approach to the admissibility of the state's forensic science and medical evidence. *University of Toronto Law Journal*, *61*, 343–409.

Edmond, G., Thompson, M. B., & Tangen, J. M. (2013). A guide to interpreting forensic testimony: Scientific approaches to fingerprint evidence. *Law, Probability and Risk*. doi: 10.1093/lpr/mgt011

Edwards, H. T. (2009a). Solving the problems that plague the forensic science community. *Jurimetrics*, *50*(1), 5.

Edwards, H. T. (2009b). Statement of the honorable Harry T. Edwards: Strengthening forensic science in the United States: A path forward. *United States Senate Committee on the Judiciary*, *47*, 1–12.

Ericsson, K. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*, 273–305.

Ericsson, K. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. Ericsson, N. Charness, & R. R. Hoffman (Eds.), *The cambridge handbook of expertise and expert performance.* New York, NY: Cambridge University Press.

Ericsson, K., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*(8), 725–747.

Ericsson, K., Krampe, R., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363.

Ericsson, K., & Smith, J. (1991). *Toward a general theory of expertise: Prospects and limits.* Cambridge: Cambridge University Press.

Eva, K. W. (2002). The aging physician: Changes in cognitive processing and their impact on medical practice. *Academic Medicine*, *77*(10), S1–S6.

Eva, K. W., Link, C. L., Lutfey, K. E., & McKinlay, J. B. (2010). Swapping horses midstream: Factors related to physicians' changing their minds about a diagnosis. *Academic Medicine*, *85*(7), 1112–1117.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: a reformulation and research agenda. *Academic Medicine*, *80*(10), S46–S54.

Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Evans, J. S. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, *18*(1), 5–31.

Evans, K. K., Cohen, M. A., Tambouret, R., Horowitz, T., Kreindel, E., & Wolfe, J. M. (2010). Does visual expertise improve visual recognition memory? *Attention, Perception, & Psychophysics*, *73*(1), 30–35.

Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin & Review*, *20*(6), 1170–1175.

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *Psychological Science*, *22*(6), 739–746.

Expert Working Group on Human Factors in Latent Print Analysis. (2012). *Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach.* Washington, DC: U.S. Government Printing Office.

Faigman, D. L., Monahan, J., & Slobogin, C. (2013). Group to individual (G2i) inference in scientific expert testimony. *Available at SSRN 2298909*.

Federal Bureau of Investigation. (1984). *The science of fingerprints: Classification and uses.* Washington, DC: U.S. Government Printing Office.

Galton, F. (1893). *Decipherment of Blurred Finger Prints.*

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong.* Cambridge, MA: Harvard University Press.

Garrett, B. L., & Mitchell, G. (2013). How jurors evaluate fingerprint evidence: The relative importance of match language, method information, and Eeror acknowledgment. *Journal of Empirical Legal Studies*.

Garrett, R. (2009, February). Memorandum from the President of the International Association for Identification. *International Association for Identification*.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training 'greeble' experts: A framework for studying expert object recognition processes. *Vision Research*, *38*(15-16), 2401–2428.

Gieryn, T. F. (1999). *Cultural Boundaries of Science*. University of Chicago Press.

Gigerenzer, G., Mata, J., & Frank, R. (2009). Public knowledge of benefits of breast and prostate cancer screening in Europe. *Journal of the National Cancer Institute*, *101*(17), 1216–1220.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176.

Haber, L., & Haber, R. N. (2004). Error rates for human latent fingerprint examiners. In N. Ratha & R. Bolle (Eds.), *Automatic fingerprint recognition systems*. New York, NY: Springer.

Haber, L., & Haber, R. N. (2006). Letter to the editor. Re: A Report of latent print examiner accuracy during comparison training exercises. *Journal of Forensic Identification*, *56*(4), 493–499.

Haber, L., & Haber, R. N. (2007). Scientific validation of fingerprint evidence under Daubert. *Law, Probability and Risk*, *7*(2), 87–109.

Haber, L., & Haber, R. N. (2009). *Challenges to Fingerprints*. Tucson, Arizona: Lawyers & Judges Publishing Company.

Hayward, R. A., & Hofer, T. P. (2001). Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. *JAMA*, *286*(4), 415–420.

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 57–80.

Ho, H. L. (2008). *A Philosophy of Evidence Law: Justice in the Search for Truth*. Oxford: Oxford: Oxford University Press.

Hogarth, R. M. (2001). *Educating Intuition*. University of Chicago Press.

Hollnagel, E., Woods, D. D., & Leveson, N. (2012). *Resilience Engineering*. Ashgate

Publishing, Ltd.

Huber, R. A. (1959). Expert Witnesses. *The Criminal Law Quarterly*, *2*(3).

International Association for Identification. (2007, December). *IAI position concerning latent fingerprint identification.* Retrieved from `http://www.onin.com/fp/IAI_Position_Statement_11-29-07.pdf`

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(3), 604–610.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577–596.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*(26), 3286–3297.

Kahneman, D. (2011). *Thinking, Fast and Slow.* New York, NY: Farrar Straus & Giroux.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585.

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

Kanki, B. G., Helmreich, R. L., & Anca, J. (2010). *Crew Resource Management.* Academic Press.

Kaye, D. (2010). Probability, individualization, and uniqueness in forensic science evidence: Listening to the academies. *Brooklyn Law Review*, *75*(4), 1163–1185.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*(3), 211–222.

Klein, G. A. (1998). *Sources of Power.* The MIT Press.

Koehler, J. (2008). Fingerprint error rates and proficiency tests: What they are and why they matter. *Hastings Law Journal*, *59*, 1077–1100.

Koehler, J. (2012). Proficiency tests to estimate error rates in the forensic sciences. *Law, Probability and Risk*, *12*(1), 89–98. doi: 10.1093/lpr/mgs013

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (2000). *To Err Is Human: Building a Safer Health System.* National Academies Press.

Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology*, *116*(3), 527–532.

Langenberg, G. (2009). Performance study of the ACE-V process: A pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ACE-V process. *Journal of Forensic Identification*, *59*(2), 219–257.

Langenburg, G., Champod, C., & Wertheim, P. (2009). Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. *Journal of Forensic Sciences*, *54*(3), 571–582.

Laukkonen, R. (2012). *Pupil Dilation as a Physiological Indicator of Perceptual Expertise.* Unpublished doctoral dissertation, The University of Queensland.

Laukkonen, R., Tangen, J. M., Baird, J., & Eva, K. W. (2013). Distracting fingerprint experts: An insight into the nature of fingerprint identification. *Unpublished Manuscript*.

Law Commission. (2011). *Expert evidence in criminal proceedings in England and Wales.* The Stationery Office.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing x-ray pictures. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Loftus, E. F., & Cole, S. A. (2004). Contaminated evidence. *Science*, *304*(5673), 959.

Loftus, E. F., Francis, E., & Turgeon, J. (2012). Model eyewitness identification jury instructions. *Dauphin County Court of Common Pleas*.

Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A., & Newell, B. R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior*, *37*(3), 197–207.

Maxmen, A. (2012, July). *Proposed bill calls for better forensic science.* Retrieved from `http://blogs.nature.com/news/2012/07/proposed-bill-calls-for-better-forensic-science.html`

Megreya, A. M. A., & Burton, A. M. A. (2008). Matching faces to photographs: poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364–372.

Mnookin, J. L. (2008a). Of black boxes, instruments, and experts: Testing the validity of

forensic science. *Episteme*, *5*(3), 343–358.

Mnookin, J. L. (2008b). The validity of latent fingerprint identification: Confessions of a fingerprinting moderate. *Law, Probability and Risk*, 127–141.

Mnookin, J. L., Cole, S. A., Dror, I. E., & Fisher, B. A. J. (2010). The need for a research culture in the forensic sciences. *UCLA Law Review*, *58*, 725.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 379–387.

Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 553.

National Research Council. (2009). *Strengthening forensic science in the United States: A Path Forward.* The National Academies Press.

Neumann, C. (2012). Fingerprints at the crimescene: Statistically certain, or probable? *Significance*, *9*(1), 21–25.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., & Bromage-Griffiths, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences*, *52*(1), 54–64.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., Meuwly, D., & Bromage-Griffiths, A. (2006). Computation of likelihood ratios in fingerprint identification for configurations of three minutiae. *Journal of Forensic Sciences*, *51*(6), 1255–1266.

Newman-Toker, D. E. D., & Pronovost, P. J. P. (2009). Diagnostic errors – The next frontier for patient safety. *JAMA*, *301*(10), 1060–1062.

Norman, G. R., & Brooks, L. R. (1997). The non-analytical basis of clinical reasoning. *Advances in Health Sciences Education*, *2*(2), 173–184.

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education*, *44*(1), 94–100.

Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of dermatology*, *125*(8), 1063–1068.

Norman, G. R., Young, M., & Brooks, L. R. (2007). Non-analytical models of clinical reasoning: the role of experience. *Medical Education*, *41*, 1140–1145.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin*

& Review, *11*(6), 988–1010.

Phillips, V. L., Saks, M. J., & Peterson, J. L. (2001). The application of signal detection theory to decision-making in forensic science. *Journal of Forensic Sciences*, *46*(2), 294–308.

Pinker, S. (2003). *How the Mind Works*. London: Penguin UK.

Potter, M. C., & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature*, *253*(5491), 437–438.

Proctor, R., & Dutta, A. (1995). *Skill Acquisition and Human Performance*. Thousand Oaks, CA: Sage Publications, Inc.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive Systems Engineering*. Wiley-Interscience.

Risinger, D., Saks, M., Thompson, W., & Rosenthal, R. (2002). The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review*, *90*, 1–56.

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, *133*(1), 63-82.

Rouder, J. N., & Ratcliff, R. (2006). Comparing exemplar- and rule-based theories of categorization. *Current Directions in Psychological Science*, *15*(1), 9-13. doi: 10.1111/j.0963-7214.2006.00397.x

Saks, M. J., & Faigman, D. L. (2008). Failed forensics: How forensic science lost its way and how it might yet find it. *Annual Review of Law and Social Science*, *4*(1), 149–171.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892–895.

Saks, M. J., & Koehler, J. J. (2007). The individualization fallacy in forensic science evidence. *Vanderbilt Law Review*, *61*(1), 199-219.

Sanderson, P. M. (2008). Capturing the essentials: Simulator-based research in aviation and healthcare. In *Eighth international symposium of the australian aviation psychology association*. Sydney, Australia.

Sanderson, P. M., Liu, D., Jenkins, S., Watson, M., & Russell, W. J. (2010). *Summative display evaluation with advanced patient simulation: Fidelity, control, and generalizability* (Tech.

Rep. No. CERG-2010-01). Brisbane, Australia.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. (1990). A cognitive perspective on medical expertise: Theory and implication. *Academic Medicine*, *65*(10), 611–621.

Scientific Working Group on Friction Ridge Analysis Study and Technology. (2011a). Standard for the definition and measurement of rates of errors and non-consensus decisions in friction ridge examination (latent/tenprint). Ver. 1.1.

Scientific Working Group on Friction Ridge Analysis Study and Technology. (2011b). Standard terminology of friction ridge examination. Ver. 3.

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica*, *68*(1), 203–215.

Spinney, L. (2010a). Forensic science braces for change. *Nature*. doi: 10.1038/news.2010.369

Spinney, L. (2010b). Science in court: The fine print. *Nature*, *464*(7287), 344–346.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–726.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*(4), 522.

Tangen, J. M. (2013). Identification personified. *Australian Journal of Forensic Sciences*, 315–322. doi: 10.1080/00450618.2013.782339

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science*, *22*(8), 995–997.

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013a). Expertise in fingerprint identification. *Journal of Forensic Sciences*, *58*(6), 1519–1530.

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013b). Human matching performance of genuine crime scene latent fingerprints. *Law and Human Behavior*. doi: 10.1037/lhb0000051

Thompson, W. C., Kaasa, S. O., & Peterson, T. (in press). Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies*.

Tomlinson, C., Marshall, J., & Ellis, J. E. (2008). Comparison of accuracy and certainty of results of six home pregnancy tests available over-the-counter. *Current Medical Research and Opinion*, *24*(6), 1645–1649.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7733–7738.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE*, *7*(3), e32800.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, *79*, 471–491.

Vicente, K. J. (2004). *The Human Factor: Revolutionizing the Way People Live with Technology*. New York, NY: Taylor & Francis.

Vokey, J. R., Tangen, J. M., & Cole, S. A. (2009). On the preliminary psychophysics of fingerprint identification. *The Quarterly Journal of Experimental Psychology*, *62*(5), 1023–1040.

Wertheim, K., Langenburg, G., & Moenssens, A. (2006a). Authors' response to letter: Letter to the editor. Re: A report of latent print examiner accuracy during training exercises. *Journal of Forensic Identification*, *56*(4), 500–510.

Wertheim, K., Langenburg, G., & Moenssens, A. (2006b). A report of latent print examiner accuracy during comparison training exercises. *Journal of Forensic Identification*, *56*(4), 55–93.

Williams, P., & Simons, D. J. (1999). Detecting changes in novel, complex three-dimensional objects. *Visual Cognition*, *7*(1), 297–322.

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, *7*(3), 275–278.

Woods, D. (1985). The observation problem in psychology. *Westinghouse Technical Report*.

Woods, D., Johannesen, L., Dekker, S., Cook, R., & Sarter, N. (2010). *Behind Human Error* (2nd ed.). Burlington, TV: Ashgate Publishing, Ltd.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of experimental psychology*, *81*(1), 141.