# *From the laboratory to the real world:*

## Evaluating the impact of impostors, expertise and individual differences on human face matching performance

Dragana Calic

*A thesis submitted in fulfilment of the requirements*

*for the degree of Doctor of Philosophy*

School of Psychology

The University of Adelaide

11 January 2013

# Contents

# *List of Figures*

# *List of Tables*

# *Abstract*

Evaluating Human Operator face matching performance in applied settings, such as airports, surveillance and access control settings would not only be logistically difficult, but it may not be possible due to many unknowns, such as the presence of impostors. Consequently, Human Operator performance has most commonly been evaluated experimentally, in well controlled laboratory settings. However, the question is, do the results obtained in the well controlled laboratory settings sufficiently reflect, and can they explain what happens in the real world? This applied problem has motivated the principal aim of this research to evaluate the feasibility of extrapolating one-to-one face matching performance findings from laboratory to the real world access control setting, and, in the process, support the development of an ecologically motivated performance evaluation methodology that could be used for future performance assessments, beyond the research reported this thesis.

The approach taken to address this aim stemmed from the focus on identity verification or ***one-to-one face matching*** task, predominantly performed within access control settings. This focus helped identify numerous factors that may affect face matching performance within access control settings. As a result, this research evaluated the impact of ***impostor type and frequency***, ***Human Operator expertise*** and ***individual differences*** on one-to-one face matching performance. A preliminary evaluation (Experiment 1) provided important methodological input into subsequent experiments. To address the principal aim, Human Operator face matching performance was first assessed within a simulated *live* access control setting (Experiment 2) which was subsequently replicated within a laboratory setting (Experiment 3). Experiment 3 also assessed the performance of an automated FR system performance to evaluate the usability of the current methodology beyond only assessing Human Operator performance.

From a methodological perspective, this research emphasised the complexities associated with evaluating and understating applied face matching performance. Applied performance may be contingent on interplay of different factors, depending on the considered applied setting. Therefore, it may not be possible to assess and state one single "level" of Human Operator performance that would be relevant to all applied settings and tasks. Instead, Human Operator performance can be assessed in light of the different environmental and task constraints, with the focus on a set of factors. Applied claims need to be appropriately qualified by explaining the exact nature of the face matching task as well as any other factors that may have affected performance.

Finally, having considered the impact of frequency and type of impostors, Human Operator expertise and individual differences, the main finding of this research showed that while overall face matching performance in the live and laboratory settings was equivalent, in the live access control setting, Human Operators were more inclined to indicate that two presented stimuli were a match, suggesting a confirmation bias. These findings are discussed in light of previous work.

# *Declaration*

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution to Dragana Calic and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

……………………………………………

Dragana Calic

11 January 2013

# *Acknowledgments*

As this long journey comes to some form of a conclusion, I would like to take a few moments to reflect and thank the many people who have been there for me over the past many years.

Special thanks go to my supervisory panel. My principal supervisor John Dunn and co-supervisor Anna Ma-Wyatt from The University of Adelaide, School of Psychology, and my external supervisors Brett McLindin and Veneta MacLeod from National Security Systems Analysis Task, Defence Science and Technology Organisation.

John, you have always ensured I look at things differently. I have learnt an invaluable amount through our discussions and I still fail to realise the source of your many ideas – it has not ceased to amaze and at times frustrate me. I owe you my most sincere gratitude for your generous and continuous guidance and support

Anna, thank you for your help throughout this process – for your eagerness to listen and ensure that I was on track.

Brett, thank you for opening my eyes to a whole new world – the world of applied research that actually makes a difference (yes, I am still the young naïve student!) and systems engineering. Among many things, you have taught me to appreciate diagrammatic representations, shorter sentences, and realise that nothing is logistically impossible. Words cannot express thankfulness for you always being there at the shortest of notices, returning drafts in record time and always willing to discuss different approaches. I am forever grateful.

Veneta, you have been my determined and persistent supervisor that I have needed. What more can I say but to at least try to acknowledge your absolute commitment and determination to me as a student and a friend. I will never forget you insisting to see drafts and ensure that I was going well even while on maternity leave. Imogen is a lucky little girl to have a mum as wonderful, caring and dedicated as yourself.

I have been so lucky to conduct my research through DSTO and especially the NSSA task. To all members of the task – I would not have been able to do it without your help and assistance. From administrative help, constant Excel and Matlab assistance, help with various trials' preparations and conduct (ah, the days of vacuuming masses of dried up millipedes and getting rid of redbacks), photographic assistance, participant recruitment, constant FR and life related discussions and much, much more. I thank you all wholeheartedly.

To my Cog Lab group – Victoria, Rachel, Anastasia, Annemarie, Fernando and Nick. It has been wonderful to have friends to relate to during this time. You have been an invaluable support day in, day out, during all these years.

To my DSTO gang – Laura, Kendy, Piotr, Carolyn, Kathy, Maya. Our lunches, lively events (just think of Departure), many thrilling discussions about the special land of DSTO and life in many forms have enlightened my way of thinking about culture and life! My special thanks goes to Dan. Thank you for meticulously reading through my thesis and always patiently listening to my endless thesis related rants.

To my friends and family (outside of university and DSTO) who at times had to pick up the pieces and deal with my curt answers to how my research was progressing questions (I mean, seriously, who asks that ☺). I cannot thank you all enough for rescuing me from the constant Ph.D land that I was absorbed in. To my karate community – I no longer have an excuse, time to start preparing for second Dan grading!

Finally, I am forever indebted to my family – mum and Ivan. Your endless patience, encouragement and willingness to hear me out over and over and over again. Mama, thank you for bringing sister and I to this wonderful country. Where would we be without you? I love you. My darling little sis – I have so much to say but no words at all. You are my true and only rock. I love you sekice moja mala.

# *Chapter 1*

## *Understanding Applied One-to-One Face Matching Performance*

> "Although there is a wealth of information about the relative strengths and weaknesses of individual algorithms for the problem of automatic face recognition, much less is known about how these algorithms compare to the system that many of them are designed to replace: the human perceiver." (O'Toole, Phillips, Cheng, Ross, & Wild, 2000, p. 552)

Although, this statement was made over a decade ago, it remains relevant today. World events such as September 11 and the Bali and London Bombings have served to intensify national and international security concerns. At the core of these security concerns are issues such as homeland security, border protection, surveillance, and access control. These issues have highlighted the importance of accurate, reliable, and timely identification and verification of individuals. Consequently, government, private, and commercial interests have been motivated to explore existing and develop new, security solutions. This has led to the consideration of, and the focus on, ***biometrics***,[1] and more specifically on ***face recognition*** (FR) as one of the most natural and publicly accepted forms of identity authentication (Bolle, et al., 2004; Bronstein, Bronstein, & Kimmel, 2006). Recent research initiatives have focused on the development, improvement, and implementation of a range of automated FR systems into applied settings (Australian Customs Service, 2007; Introna & Nissenbaum, 2009; Kirby, 2008). However, while consistent research efforts have increased theoretical and applied knowledge and resulted in implementations of automated systems, these implementations have occurred without appropriate consideration of the system that they have been designed to replace – the Human Operator.

Although recent FR research initiatives have predominantly focused on automated solutions, the Human Operator remains an important part of the applied environment. Popular television crime shows have many believe that FR systems are completely automated and that they can easily and accurately identify and verify people within seconds (known as the CSI effect) (Schweitzer & Saks,

[1] Biometrics constitutes unique measurable physiological (e.g., facial structure, fingerprints, hand geometry) and/or behavioural (e.g., keystroke dynamics, gait) characteristics that can be used to verify (one-to-one) and identify (one-to-many) an individual (Bolle, Connell, Pankanti, Ratha, & Senior, 2004; Wayman, 1999, 2001; Woodward Jr., Horn, Gatune, & Thomas, 2003).

2007). However, in reality, that is not the case. While research advancements have enabled complete automation of some FR systems (e.g., certain access control applications), the Human Operator continues to play an integral part in a variety of applications (MacLeod, 2010; Sunde et al., 2003). The majority of applied FR tasks are still predominantly conducted by Human Operators, even when an automated FR system is implemented (Australian Customs and Border Protection Service, 2010). This can be exemplified by considering SmartGate, an access control border processing system currently in operation across all major international airports in Australia (Australian Associated Press Pty Limited (AAP), 2010; Tay, 2010). Australian Customs 2010 Annual Report stated that of over 13 million travellers processed during the 2009-10 year, SmartGate processed 1.5 million, meaning that over 11 million travellers were processed by Customs Primary Line Officers (Australian Customs and Border Protection Service, 2010). The role of Human Operators within these settings extends beyond monitoring and maintenance of automated systems. Human Operators also work alongside automated FR systems conducting the same task or utilising automated FR systems to assist with their decision making (Kemp & Howard, 2007; MacLeod, 2010).

Consequently, implementation of automated FR systems without appropriate consideration of Human Operator abilities is concerning on two fronts. First, the newly developed technological solutions are often put into operation prior to comparing their performance with that of the Human Operator. This raises questions about the extent to which the new automated FR systems enhance security. Second, perhaps even more alarming, is that performance of Human Operators, required for the human-automated comparison, has not been appropriately evaluated and quantified.

Although, human visual abilities have traditionally been relied on within security settings, Human Operator FR performance within these settings remains largely unknown. This may be attributed to many complexities associated with appropriately evaluating applied performance. Applied Human Operator FR performance is contingent on the diversity of applied settings and many conditions and constraints associated with these settings. For example, the complexity of specific tasks (e.g., one-to-one or one-to-many); and the numerous environmental (e.g., lighting) and other factors (e.g., ethnicity, age, gender, etc., of Human Operators and presenting individuals) can differently affect performance. Therefore, it may not be possible to assess and state one single "level" of Human Operator performance that would be relevant to all applied FR settings and tasks. Instead, Human Operator FR performance can be assessed in light of different environmental and task constraints, with the focus on a set of factors.

On the one hand, evaluating Human Operator performance in airports, various surveillance, access control, and related applied settings is extremely difficult if not impossible due to numerous practical and logistical complexities. On the other hand, however, evaluating Human Operator performance in well controlled laboratory settings may produce results which do not reflect what happens within applied settings. This may be problematic when wanting to experimentally evaluate performance and, more importantly, determine the extent to which laboratory results can be used to inform real world applications. In an attempt to overcome this, researchers sometimes simulate applied settings to evaluate performance under conditions that as closely as possible resemble applied environments of interest. These evaluations are extremely time consuming and logistically difficult to organise and conduct. For that reason, only a few such evaluations have been conducted so far (Butavicius et al., 2008; Kemp, Towell, & Pike, 1997; Megreya & Burton, 2008) (Section 2.2.5).

This applied problem has motivated the principal aim of this research to evaluate the feasibility of extrapolating one-to-one face matching performance findings from a laboratory setting to the real world access control environment, and support the development of an ecologically motivated performance evaluation methodology that could be used for future performance assessments. The approach taken to address this aim stems from the focus on ***verification*** or the ***one-to-one face matching task,*** which is predominantly performed within ***access control settings***. The focus on access control has also brought to the forefront the many factors that may affect face matching performance within applied access control settings. This has resulted in this research also evaluating the impact of impostors, Human Operator expertise, and individual differences on one-to-one face matching performance. While considering the impact of these factors, this research first assesses performance within a simulated ***live access control setting*** and second, replicates the *live* access control evaluation to assess performance within an ***experimental laboratory setting***. Findings from these evaluations answer the following questions:

1. What is the performance on a one-to-one face matching task within a *live* access control environment?
2. What is the performance on a one-to-one face matching task within a laboratory environment when the same stimuli used in the *live* setting are presented?
3. How do findings from the *live* and laboratory settings compare, and to what extent can the findings from the laboratory experiment be extrapolated to the *live* access control evaluation?

In addition to evaluating the feasibility of extrapolating one-to-one face matching performance findings from laboratory to real world access control settings, this research also aims to, support the development of an ecologically motivated performance evaluation methodology. In order to evaluate the applicability of laboratory performance findings to the real world, Human Operator performance will be evaluated within multiple laboratory experiments and a simulated real world access control setting (Section 1.3). In addition to evaluating Human Operator performance, this methodology will also be used to assess the performance of an automated FR system. This will be done to assess the usability of this methodology beyond only evaluating Human Operator performance. Consequently, a comparative performance assessment of Human Operator and an automated FR system will be conducted. It is hoped that this methodology could demonstrate how to appropriately prepare and conduct face matching performance evaluations which consider applied settings of interest and multiple factors that may affect performance.

What follows is a generic overview of FR and an introduction of the relevant terminology. The distinction between face matching and recognition, and identification and verification is also presented. Finally, provided is an overview of the main factors considered as part of the current research, along with an outline of what each of the following chapters contains.

## *1.1 Face Recognition: An Overview*

In discussing FR in general, it is important to consider its applicability beyond everyday social settings and to investigate its function within access control and, more broadly, security.

The remarkable ability of the human visual system is fundamental to everyday life. Human FR abilities are constantly relied upon and have also been applied to various security situations. It has long been commonplace for Human Operators to authenticate individuals' identities. As such, humans are comfortable with presenting their face and having it viewed and inspected by others, especially for the purposes of authentication. Consequently, the majority of identification documents available today contains a photograph of its bearer (Bolle, et al., 2004; Bronstein, et al., 2006; Introna & Nissenbaum, 2009). Probably the most prominent operational example of this task is international traveller border processing. Other examples may include Police Officers verifying individuals' identities, or security personnel in situations where authenticated access to facilities is required.

These characteristics of FR compared to other biometrics serve to reinforce its use within diverse security applications (Bolle, et al., 2004; Bronstein, et al., 2006; iTWire, 2011; Liu & Wechsler, 2005). Unlike many other biometrics, FR is *passive* as it normally does not require an individual's cooperation and physical proximity, thus allowing clandestine collection (Introna & Nissenbaum, 2009). It does not require the contact between the individual and the sensor as is the case with fingerprints and palm prints, for example. It also does not require a sample of the body to be acquired, such as with DNA. Finally, it does not require observation of the individual's behaviour such as with keystroke dynamics or gait biometrics. It is one of the least intrusive biometrics, causing minimal inconvenience or discomfort to its users. These aspects of FR, coupled with the fact that it has historically been used to authenticate people, serve to reinforce that FR is a natural, less intimidating, and widely accepted biometric (Bolle, et al., 2004; Bronstein, et al., 2006). Therefore, it is not surprising that FR is one of the most widely used and commonly discussed biometrics (Bolle, et al., 2004; iTWire, 2011).

The combination of these characteristics of FR coupled with the increase in national and international security concerns have led to an intensified interest in FR, and especially the possibility of its automation. The increase in security concerns is consequent to numerous world events such as, September 11, the Bali and London Bombings, and more recently the London Riots. Those events have, to an extent, been used to promote the usability of FR. For example, in relation to September 11, a video of Mohammad Atta and Abdulaziz Alomari passing through airport security at the Maine airport on the day of the attack is often shown and coupled with claims that FR technology would have identified the men as wanted terrorists (Sinha, Balas, Ostrovsky, & Russell, 2006a). Similar claims are made in relation to the London 7/7 Bombings (National Academy of Sciences, 2010).

Although it is not possible to know if the presence of automated FR technologies would have prevented those events, such claims have most certainly inspired numerous research initiatives to explore the possibility of FR automation. This has probably occurred because FR has historically been used in security applications and is thus widely accepted for identity authentication (National Academy of Sciences, 2010; Woodward Jr., et al., 2003). Also, its applicability and usability with existing image databases and surveillance footage from closed circuit television (CCTV) have made it an especially viable operational option (Woodward Jr., et al., 2003). Consequently, FR is appropriate for a wide range of security applications and has been easily and popularly implemented within various security contexts (Bolle, et al., 2004; Bronstein, et al., 2006; iTWire, 2011; Liu & Wechsler, 2005). To sufficiently understand the extent of FR security applications for both, human and automated implementations, it is first important to define and distinguish between face matching and recognition.

## *1.2 Face Matching: Definition and Application*

The term FR overarchingly refers to a number of distinct, but related tasks, performed by Human Operators and automated FR systems (Adler & Schuckers, 2007; Bruce, Henderson, Newman, & Burton, 2001; Burton, Miller, Bruce, Hancock, & Henderson, 2001; Burton, Wilson, Cowan, & Bruce, 1999; O'Toole & Tistarelli, 2010). To understand these tasks it is important to distinguish between ***face matching*** and ***face recognition***.

***Face matching*** involves comparison of two or more faces (*live* or presented in imagery) with the aim to make a decision about if they are of the same or of a different individual. A face matching task is not associated with memory load as face stimuli are presented simultaneously. ***Face recognition***, although to some extent still involves face matching, is different because it is a memory based task. It involves retrieval of relevant facial information in order to make a decision about whether a face had been seen before and whose it was. As such, while recognition is mainly associated with perceiving faces that have been seen before or familiar faces, face matching can be thought of as processing and comparing unfamiliar faces (Bruce, et al., 2001; Burton, et al., 1999; Havard, 2007). As recognition relies on accessing previously stored information, it can be thought of as a task that is internalised to the individual performing it. In some ways it may be unconsciously performed. During face matching, however, an individual is predominantly presented with faces that had not previously been encountered and is required to appropriately process the novel stimuli. Therefore, contrary to recognition, it may be useful to think about face matching as an externalised task that requires conscious effort and thought. It has been suggested that unfamiliar face matching relies on pictorial, or image-based processing; whereas recognition of familiar faces requires a more specialised and robust type of processing (Hancock, Bruce, & Burton, 2000).

In considering current and potential applications of face matching, it is also important to distinguish between ***surveillance*** and ***access control***. The following sections discuss the generic surveillance and access control applications as they relate to human and automated face matching, and make a distinction between ***identification*** (i.e., one-to-many) and ***verification*** (i.e., one-to-one face matching).

### *1.2.1 Surveillance: Identification or One-to-Many Face Matching*

Surveillance applications are most commonly concerned with the detection of persons of interest. An example may include airport surveillance where the aim is to notify security about the presence of a person of interest (e.g., a terrorist). As such, surveillance applications are concerned with one-to-many (i.e., 1:n) or ***identification*** tasks which attempt to identify one individual among a database of many other individuals. The process of identification asks and attempts to answer, "Is this person present in the database?" and "Who is this individual?".

Identification, like any other FR task, has traditionally been conducted by Human Operators. However, more recently with the uptake of technology the conduct of identification, perhaps, provides the best example of how identity authentication is conducted jointly by automated FR systems and Human Operators. Figure 1 depicts a generic surveillance scenario and demonstrates how this may occur.

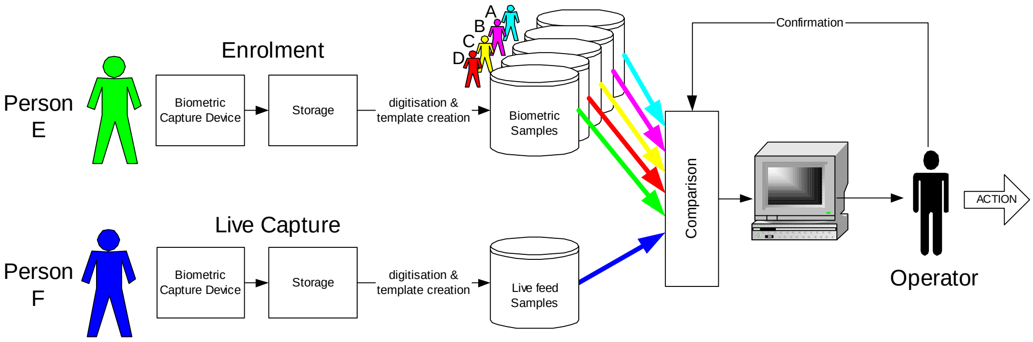


*Figure 1: A generic surveillance application, adapted from McLindin (2005)*

First, individuals (A to E) are enrolled into the database. To subsequently identify these individuals, *live* face data (by imaging individuals, with or without their knowledge) or a template (if a *live* individual is not present) is presented (F) and compared against the previously enrolled database of templates. This comparison results in a collection of images (determined by the algorithm to most closely resemble the target person) which are ranked based on their similarity scores (Introna & Nissenbaum, 2009; Woodward Jr., et al., 2003).[2] These images are then presented to the Human Operator who makes the final identity decision. Therefore, for surveillance applications, automated FR is a tool assisting human decision making (Kemp & Howard, 2007; MacLeod, 2010; McLindin, 2005).

[2] It should be noted that this is a complex task and that the description provided here is very simplified. However, the aims of this research are not concerned with identification. Therefore, it is described sufficiently to enable distinction with verification, which is the focus of this research.

### *1.2.2 Access Control: Verification or One-to-One Face Matching*

Access control applications involve situations where an individual requires access to a location or information (e.g., at international airports, a secure area, bank details). To obtain access, the individual needs to provide certain information (e.g., password, name) or present a portable token (e.g., identity card, passport). Essentially, the individual makes a claim to an identity which requires a comparison between the *live* person and the provided information. Hence, access control applications are concerned with one-to-one (i.e., 1:1) matching or ***verification*** of an individual's claim to identity. Verification can be understood in terms of answering the question "Is this individual who they claim to be?" (Introna & Nissenbaum, 2009; Woodward Jr., et al., 2003). Traditionally, this task has been conducted by Human Operators. In various security access control applications (e.g., border control, access to secure premises), Human Operators would commonly be required to compare a *live* individual's face to a facial photograph contained in a token presented by the individual claiming the identity.

An example of an automated access control system is SmartGate. SmartGate is Australia's automated border processing system, which relies on the passenger to provide an ePassport[3] which contains a previously enrolled facial image encoded on an embedded chip within the passport. When the passport is presented to the system, a comparison is made between the enrolled template created from the face on the ePassport and that of its holder (Australian Customs Service, 2004, 2007; Fraser, 2004; Introna & Nissenbaum, 2009; Wayman, 2008). Figure 2 shows a general access control application process as conducted by automated FR systems.

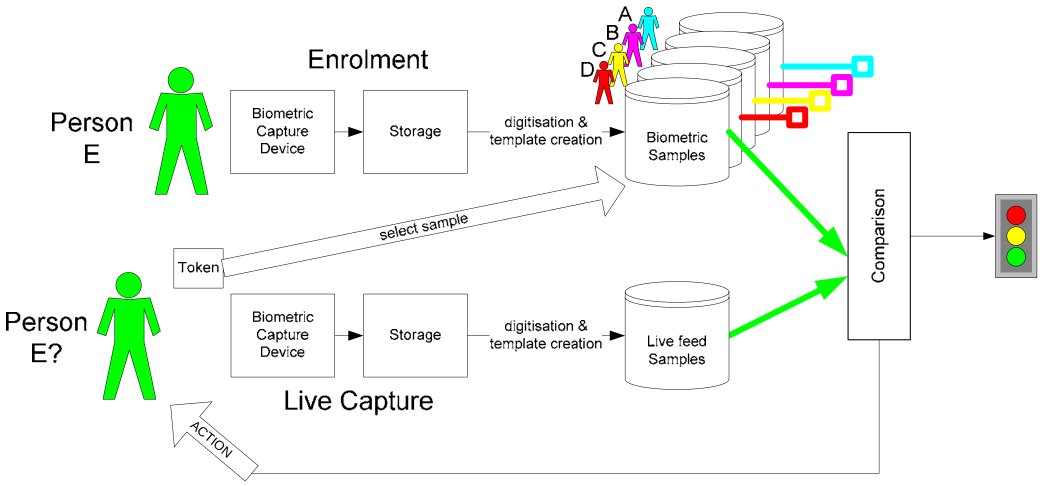


*Figure 2: A generic access control application, adapted from McLindin (2005)*

Similarly to surveillance, access control applications follow an enrolment and acquisition process. However, unlike surveillance applications, where *all* enrolled

[3] An e-Passport has an embedded microchip which stores the information printed on its data page (e.g., name, date of birth). The microchip also stores the holder's digitised biometric identifier (e.g., a photograph or other, depending on the issuing country) (Australian Government: Department of Immigration and Citizenship, 2009; U.S. Department of Homeland Security, 2011).

templates are compared to the *live* template (i.e., one-to-many), access control applications involve comparisons of only one 'claimed' template with a *live* presenting face (i.e., one-to-one).

Human Operator and automated FR system verification performance can, in part, be assessed and understood in terms of error rates that can be calculated. The following section defines these in terms of the verification task.

#### *1.2.2.1 Understanding Performance*

Applied access control settings require a binary, match or non-match decision to be made. Based on this decision, an individual's identity is either confirmed or not. This decision, however, produces four different outcomes, summarised in Table 1 and explained further. It should be noted that the decision outcomes presented here have become conventional terminology, with some discipline specific variation. However, to explain the results of this work, this terminology is defined differently. The reasoning behind this is outlined in Section 4.2. For the purposes of this section, however, the aim is to explain the meaning behind match and non-match decisions and this is done using conventional terminology.

*Table 1: Decision outcomes*

| | ***Decision*** | |
|---|---|---|
| | ***Accepts***<br>(Same) | ***Rejects***<br>(Different) |
| True Stimuli<br>(i.e., matching stimuli or same individuals) | Correct Accept | False Negative |
| Impostor Stimuli<br>(i.e., mismatching stimuli or different individuals) | False Positive | Correct Reject |

Confirmed identity can be understood in terms of two outcomes. If the identity is correctly accepted, that is referred to as a ***correct accept***. If, however, a Human Operator or an FR system make a mistake and incorrectly accept an impostor, this is referred to as a ***false positive***. A false positive is a Type II error which can be used to determine the system's False Accept Rate (FAR) (also referred to as False Match Rate (FMR)). In applied settings, a false positive may be difficult to detect as it requires the identification of individuals who are falsely presenting as legitimate users (i.e., impostors) (Graves, Johnson, & McLindin, 2003; Introna & Nissenbaum, 2009; National Science and Technology Council (NSTC), 2006; Wayman, et al., 2005; Woodward Jr., et al., 2003). An example of this could be a person using another similar looking individual's passport and incorrectly being granted access.

If an identity claim is rejected, this can also be understood in terms of two outcomes. First, this may occur because the human or the system made a mistake and incorrectly rejected a true presenter. This is referred to as a ***false negative*** which is a Type I error, and can be used to calculate the system's False Reject Rate (FRR) (also referred to as False Non-Match Rate (FNMR)). Operationally, this error rate is considerably easier to assess as it occurs when a true presenter is incorrectly rejected. Second, this may be because the individual is an impostor and is making an illegitimate identity claim. This is referred to as ***correct reject*** (Graves, et al., 2003; Introna & Nissenbaum, 2009; National Science and Technology Council (NSTC), 2006; Wayman, et al., 2005; Woodward Jr., et al., 2003).

Having considered performance measures, it is important to turn attention to how those decisions are made by Human Operators and automated FR systems. Although in applied settings, a simple binary, match or non-match decision is made, the process behind the final decision is much more complex and is best explained using Signal Detection Theory (SDT). SDT is a method of modelling human decision making processes that can be applied to almost any task which requires a discrimination of two possible stimuli – ***signal*** and ***noise*** – to help understand decision making in the presence of uncertainty (Macmillan & Creelman, 2005; McNicol, 2005; Stanislaw & Todorov, 1999; WISE (Web Interface for Statistics Education)). As such, SDT is concerned with decisions based on evidence which does not allow for an unequivocal selection of only one stimulus against other options. Therefore, decisions made may be variable, and are even likely to change when the same stimuli are presented at a different point in time. This difference in response to the same stimuli implies that the threshold values change. Consequently, decision making in these situations allows for a distinction of two aspects of human decision making – ***sensitivity*** and ***decision criterion*** (Macmillan & Creelman, 2005; McNicol, 2005; Stanislaw & Todorov, 1999; WISE (Web Interface for Statistics Education)).

An observer's ***sensitivity*** or ***discriminability*** refers to the extent to which the individual is able to make correct judgements and avoid incorrect ones. Human decision making can also be affected by a combination of non-sensory factors (e.g., attention, motivation, certain preconceived ideas about the task, etc.) – collectively referred to as a ***decision criterion*** or ***response bias*** (Macmillan & Creelman, 2005; McNicol, 2005; Stanislaw & Todorov, 1999; WISE (Web Interface for Statistics Education)). To explain the impact that a decision criterion can have, it may be useful to consider an example relating to how an individual's criterion can be affected by perceived consequences of their decision. For instance, a Human Operator working in an international airport's arrivals area suspects that an arriving passenger is not who they claim to be. Deciding to act on this suspicion could lead to disruptions of the airport's daily work flow. The Human Operator may be required to justify their claim, and the passenger would probably be questioned. If the traveller is not found to be fraudulent, this may have negative consequences for the reputation of the border processing services and even potentially, for the Human Operator. If, however, consequences of not making this decision are perceived by the Human Operator to be more costly (e.g., country's national security may be in jeopardy) then they would be more willing to make the claim that the arriving individual may not be who they claim to be.

To some extent "decisions" made by automated FR systems are similar to those made by Human Operators. Comparing a presented face with an enrolled template in a database produces a similarity score which indicates the degree of similarity between the presented face and the enrolled template. The system operating point or ***threshold*** is programmed into the system – this is equivalent to a decision criterion for humans. The two templates are defined as a match if the similarity score meets the threshold criteria and rejected if it does not (Ashbourn, 2005; Introna & Nissenbaum, 2009; Wayman, et al., 2005).

To implement an FR system it is important to set an appropriate threshold. The threshold is the trade-off between the two main types of errors – type II (i.e., false positives) and type I (i.e., false negatives). For example, when threshold settings are high, false positives are generally reduced and false negatives increased. This means that fewer impostors would be able to get through the system, but also, that more true presenters would be rejected. Conversely, when threshold settings are low, false positives are increased which may result in more impostors being incorrectly accepted, and also, false negatives are reduced meaning that more true presenters would be correctly accepted. Threshold settings are dictated by the application and security and business requirements (Ashbourn, 2005; Introna & Nissenbaum, 2009; Wayman, et al., 2005). Although, an automated FR system threshold is strictly defined and potentially justifiable, it is equivalent to Human Operator decision criterion. Directly or indirectly, both are based on the applied settings and the consequences that certain decisions may have on that setting.

The following section outlines main aims of the current research.

## *1.3 Current Research Aims: Factors Considered*

The principal aim of this research is to evaluate the feasibility of extrapolating one-to-one face matching performance findings from laboratory to the real world access control settings, and, in the process, to support the development of an ecologically motivated performance evaluation methodology. As the approach taken to address this aim stems from the focus on access control applications, various factors that may affect one-to-one face matching performance within an access control setting are considered. Consequently, this research evaluates the impact of ***impostor frequency*** and ***type***, Human Operator ***expertise***, and ***individual differences*** on one-to-one face matching performance. In assessing the impact of these factors, this research first evaluates performance within a *live* access control setting and, second, replicates the *live* access control evaluation to assess performance in a laboratory experiment. Finally, findings from the *live* and laboratory evaluation are compared to answer the question about the extent to which the findings from the laboratory can be extrapolated to the *live* access control setting.

The focus on examining the impact of ***impostors***, ***expertise***, and ***individual differences*** on one-to-one face matching performance is motivated by the applied nature of this research. Essentially, performance cannot be appropriately assessed without considering the environment and the interplay of many different factors that can affect it. Therefore, this research focuses on these three factors because they have not been extensively assessed, and because they have greater applicability than just to access control settings.

The impact of impostors is considered from two perspectives. The first focuses on assessing the impact of presenting different ***frequencies of impostors***, and the second on the impact of ***different types of impostors***. The impact of different frequencies of impostors on performance is assessed because in the majority of applied settings the occurrence of impostors is not known, yet to allow for neat experimental designs researchers predominantly present 50% impostor stimuli (Section 2.2.4). It is therefore important to assess if performance would differ when Human Operators are presented with different frequencies of impostors. Assessing the impact of different types of impostors is similarly motivated. The aim is to evaluate if differently created impostors affect performance in different ways. The focus here is on generating impostor types based on operationally plausible scenarios (Section 3.3).

Human Operator ***expertise*** incorporates aspects of training and/or experience that a Human Operator may have. The impact of expertise is highly relevant to face matching research as various face matching and recognition tasks are performed as part of daily social interactions. That is why it is often assumed that people are very good at face matching and recognition tasks. While that is true for familiar FR, face matching tasks in applied settings in which Human Operators are predominantly comparing unfamiliar faces prove challenging to the human perceiver (Bruce, et al., 2001; Hancock, et al., 2000; Megreya & Burton, 2006, 2007, 2008). While it may make sense to assume that expertise would improve performance, it is first important to know what the current performance levels of individuals who conduct face matching as part of their employment are. Therefore, the approach taken here is to assess one-to-one face matching performance of individuals who currently conduct different face matching tasks as part of their employment, and compare their performance to that of lay people. It is hoped that a greater understanding of the current Human Operator face matching expertise combined with other outcomes from this work could inform development of appropriate training regimes to improve face matching performance within different security applications. This also ties in with identifying individuals who may be best suited to performing face matching tasks. It is therefore hoped that the focus on ***individual differences*** will increase our understanding of human verification performance and provide insights about if any individuals are predisposed to better performing this task. As such, this work could have important practical implications in terms of selection and recruitment of individuals within customs, immigration, military, security settings, and other personnel who are required to perform verification tasks.

The impact of impostors, expertise, and individual differences is considered as part of the two main evaluations of this research – Experiment 2 (Chapter 5) and Experiment 3 (Chapter 6). These evaluations jointly input into the assessment of the extent to which results obtained in a laboratory can be extrapolated to applied access control settings. This focus stems from the need to better understand Human Operator and automated face matching performance within applied settings, as well as to more appropriately assess applied face matching performance. However, assessing face matching performance within applied settings may not be plausible and/or possible. Even experimental research which simulates applied settings is rare (Section 2.2.5). Overwhelming logistical complexities associated with conducting such evaluations has meant that the majority of research has been conducted in tightly controlled laboratory experiments. Therefore, the aim of the current work is to conduct a simulated *live* evaluation and an equivalent laboratory evaluation to explore the feasibility of applying the results obtained in the laboratory to applied settings. If performance differences between the *live* and laboratory assessments are minimal, it may mean that appropriately prepared and conducted laboratory evaluations are adequate to provide an accurate estimation of applied performance, thus reducing much time and effort associated with *live* performance evaluations. With that, it is anticipated that the methodology used for the assessment of Human Operator face matching performance would serve as a prototype (an example) of how to develop an ecologically motivated methodology for the assessment of face matching performance within access control settings and more broadly.

The remainder of this thesis is divided into six chapters. ***Chapter 2*** provides a review of relevant literature, focusing on Human Operator face matching abilities and the many different factors that can affect performance. ***Chapter 3*** – Image Preparation – involved the preparation of stimuli which were used during performance experiments. ***Chapter 4*** – Experiment 1 – was a scoping study which aimed to examine the impact of impostor frequency and type on one-to-one face matching performance. Results from this study determined the impostor frequency that was used throughout the following two experiments. ***Chapter 5*** – Experiment 2 – aimed to recreate an applied access control setting in an attempt to evaluate *live* Human Operator one-to-one face matching performance. This experiment also assessed the impact of four impostor types on face matching performance. Human Operators also completed a battery of individual difference tests which were used to assess if certain characteristics (e.g., cognitive, personality, etc.,) are able to predict face matching performance. ***Chapter 6*** – Experiment 3 – was a laboratory replication of Experiment 2. Using video stimuli of participants acquired during Experiment 2, the performance of trained Human Operators was assessed in a laboratory experiment. Additionally, in line with Experiment 2 methodology, the impact of different impostor types and Human Operator individual differences were assessed. Finally, this study also assessed the performance of an automated FR system using the same stimuli that was presented to Human Operators. Finally, ***Chapter*** *7* combines, summarises and discusses the key findings, then grounds them within relevant literature and suggests further work that can be done within the area.

# *Chapter 2*

# *The Human Operator*

Chapter 1 provided the basis for the current research and established a clear need to focus on the Human Operator as an important part of the applied access control setting. This chapter reports on relevant literature relating to Human Operator performance in general, and more specifically, provides a review of different factors which can affect Human Operator one-to-one face matching performance.

## *2.1 Human Operator Performance*

Humans rely on their abilities to accurately recognise (a memory based task) and distinguish (a matching based task) between faces to adequately function in everyday society. It is therefore often assumed that they are highly skilled in those tasks. This has been shown to be correct for ***familiar faces***, even under varying conditions (e.g., lighting, pose, motion, facial expression, etc.) (Burton, et al., 1999; Hancock, et al., 2000; Pike, Kemp, & Brace, 2000; Vast & Butavicius, 2005). This may be because the individual has had the opportunity to view the face from different angles, under variable lighting, still and/or in motion, as well as under various other optimal and non-optimal conditions. Therefore, the individual has been exposed to, and provided with, an abundant amount of information about a particular face to build a rich memory representation of the person and their face which is subsequently used to assist with recognition.

However, this is not the case for ***unfamiliar faces***. When required to match unfamiliar faces, performance is found to be very poor (Bruce, et al., 2001; Burton, et al., 1999; Hancock, et al., 2000; Megreya & Burton, 2006, 2007, 2008; Pike, et al., 2000). People are generally not provided with sufficient information (i.e., they do not have the opportunity to view the face from different angles, under different lighting conditions, with different expressions, etc.) that would enable them to perform face matching accurately and reliably, and this seems to be reflected in their performance. Bruce and Young (1986) argued that memory for unfamiliar faces is based on situation specific conditions, whereas memory for familiar faces is associated with abstract information allowing generalisation beyond specific situations. Furthermore, Hancock, et al., (2000) have suggested that processing of unfamiliar faces relies on image or pictorial matching, rather than on more sophisticated face matching strategies that are used when recognising familiar faces.

Illumination conditions, pose of the person, ethnicity, image quality, motion, intentional or unintentional appearance changes, as well as many other factors may all be found to affect human recognition, and especially face matching performance. The way and the extent to which these factors affect human face matching and recognition performance are discussed in Section 2.2. Prior to turning attention to those factors, face processing is briefly considered.

### *2.1.1 Face Processing*

To better understand how people recognise and match faces, three ways in which they perceive and process faces are considered – featural, configural and holistic processing.

Firstly, faces can be processed in terms of the distinct facial features (e.g., eyes, nose, mouth, etc.,) that make them up. This type of processing is known as ***featural*** or ***piecemeal*** (Bruce, 1988; Carey & Diamond, 1977; O'Toole & Tistarelli, 2010; Schwaninger, Carbon, & Leder, 2003; Schwaninger, Lobmaier, Wallraven, & Collishaw, 2009). Research has shown that a single facial feature can provide sufficient information for recognition of famous (i.e., familiar) faces (Fraser, Craig, & Parker, 1990; Sadr, Jarudi, & Sinha, 2003). Sadr, et al., (2003) digitally erased eyes or eyebrows from a set of celebrity faces. Perhaps surprisingly, they found a more significant detriment in recognition performance for faces that had no eyebrows compared to those which had no eyes. These findings may suggest that facial features alone are important and sometimes sufficient for accurate recognition of familiar faces.

Having considered the importance of facial features for face processing, a distinction between internal and external features should be made. Internal features refer to the central region of the face which constitutes eyebrows, eyes, nose and mouth, while external features are hair, beard, ears and jaw line (Megreya, Memon, & Havard, 2011; Sinha, et al., 2006a). Face perception research has consistently shown that when a face is familiar, internal features assist with face matching and recognition performance (Bonner, Burton, & Bruce, 2003; Clutterbuck & Johnson, 2002; Ellis, Shepherd, & Davies, 1979). However, when a face is unfamiliar, focus on external features has been found to aid performance (Bonner, et al., 2003; Bruce et al., 1999; Frowd, Bruce, McIntyre, & Hancock, 2007). Some researchers have therefore suggested that the dissociation between familiar and unfamiliar face processing may be attributed to the way that internal features are encoded (Clutterbuck & Johnson, 2002). However, Megreya and Bindemann (2009) found that while British participants displayed an external-feature advantage for processing unfamiliar faces consistent with previous findings (Bonner, et al., 2003; Clutterbuck & Johnson, 2002; Ellis, et al., 1979), Egyptian participants however, showed an internal-feature advantage which was consistent across a number of different face matching tasks, both genders and when viewing both Egyptian and British faces. Megreya and Bindemann (2009) attributed this to Egyptian participants' long term exposure to female faces with headscarfs, which cover the external features. This finding was replicated in a recent study by Megreya, et al., (2011).

Secondly, faces can be perceived in terms of ***configural*** or ***relational*** information which is based on the relationships between facial features (Carey & Diamond, 1977; Schwaninger, et al., 2003). Configural information can be further divided into ***first-order relational information*** which refers to basic "eyes above nose, which is above mouth" arrangements; and ***second-order relational information*** which refers to specific metric relations between features (e.g., distances between the eyes, or nose and mouth) (Maurer, Le Grand, & Mondloch, 2002; Schwaninger, et al., 2003, p. 82).

The importance of configural information for face processing is often illustrated by considering the 'Thatcher illusion' shown in Figure 3 (Thompson, 1980). Figure 3, shows that when a face is inverted (top right), distortion of its facial features is barely noticeable. However, once the same face is presented in the upright orientation (bottom right), the grotesqueness of the face is clearly visible. This example is often used to demonstrate the importance of face configuration in face processing and perception. However, it can also be used to illustrate that people are sensitive to configural changes only when faces are presented upright.

This can be seen from the Thatcher illusion example where the inverted altered image (top right) appears almost unchanged. This may suggest that configural information in faces can only be perceived when faces are presented in the upright orientation (Maurer, et al., 2002; O'Toole, 2004; O'Toole & Tistarelli, 2010).

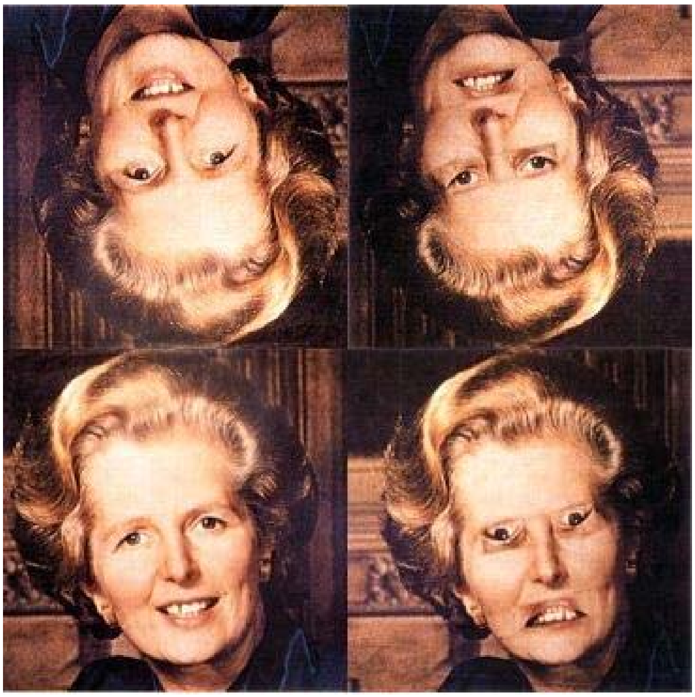


*Figure 3: The Thatcher illusion, adapted from Thompson (1980)*

Sadr, et al., (2003) presented facial images of famous faces where the top and bottom half of the face belonged to different individuals. Participants found that it was difficult to decipher the identity of either individual. However, when the two face halves were misaligned (so that they do not present as one face), the two distinct identities were easily recognised (Hole, 1994; Young, Hellawell, & Hay, 1987). This demonstrates the usefulness of relational information between facial features in assisting recognition of familiar individuals.

Finally, a face can be processed as a whole, with the focus not being on specific facial features or their relationships. This type of face processing is referred to as ***holistic*** (O'Toole & Tistarelli, 2010; Schwaninger, et al., 2003; Schwaninger, et al., 2009; Tanaka & Farah, 1993).[4] Some research has demonstrated the advantage of processing full faces over only specific facial features. For example, on a one-to-one face matching task, Kemp (2009) presented participants with hard and easy face pairs which contained whole face-whole face; whole face-internal features only; and, internal features-internal features. Overall, accuracy was significantly better when two whole faces were compared (i.e., whole face-whole face condition). However, for hard face pairs, only a small statistically significant difference was found for internal-internal condition compared to whole face-whole face condition. The three types of face processing discussed here are dependent on a number of different factors (e.g., familiarity, the task, stimuli presentation). Better understanding the extent to which face processing is affected by any of these factors may help improve applied face matching performance.

[4] Please note that there has been some terminology related contention, especially in relation to configural and holistic face processing often being used interchangably (Maurer, et al., 2002).

## *2.2 Factors that Impact Human Operator Face Matching Performance*

Face matching is performed in various access control settings, retail outlets, airports, banks, and other similar situations. The task may also be performed for long periods with restricted time limits, under strict processing guidelines, and often along with additional related (e.g., confirming biographical information), or even unrelated tasks (e.g., confirming flight information in an airport setting). In spite of this, quick, accurate and efficient performance is crucial. In such settings face matching performance can be affected by various factors. Some of these factors and the extent to which they can affect face matching (and recognition) are considered next.

To organise the many different factors that can impact human face matching performance, Hancock, et al., (2000) usefully divided the factors into those that are image-based and those that are individual specific (i.e., inherent to the facial structure). Using this distinction, the current overview is divided into five sections. ***Image and environment specific factors*** (e.g., illumination/lighting, viewpoint/pose, image quality) are discussed in Section 2.2.1. ***Individual specific factors*** (e.g., gender, age, ethnicity/race) that may be relevant to both the individual being inspected (e.g., client, user, target) and the Human Operator are covered in Section 2.2.2. ***Human Operator specific factors*** (e.g., expertise, individual differences, and stress and fatigue) relate specifically to the individual performing the face matching task are discussed in Section 2.2.3. Current research also seeks to understand two operationally relevant factors which are of specific relevance to the aims of this research. One relates to understanding ***the impact of impostors*** (Section 2.2.4), and the second relates to understanding ***Human Operator performance in live versus experimental settings*** (Section 2.2.5).

Finally, two key issues must be clarified. First, it should be noted that the list of factors considered here is not exhaustive. With the applied nature of this work, this list has aimed to incorporate the most operationally prevalent and relevant factors. Second, in the absence of research relating specifically to face matching, relevant studies focusing on face recognition research are reported.

### *2.2.1 Image and Environment Specific Factors*

#### *2.2.1.1 Image Quality*

The impact of image quality is of particular relevance in applied settings where it may vary substantially and is likely to be poor. The majority of imagery from everyday surveillance equipment is of poor quality in terms of image resolution, the acquisition angle, and the distance from which the face of an individual may have been acquired (Bruce, et al., 2001; Heyer, MacLeod, Calic, Kuester, & McLindin, 2009; Keval & Sasse, 2008).

Research has demonstrated that when presented with familiar faces the human visual system is able to tolerate low resolution and significant levels of degradation to face images without affecting performance (Hancock, et al., 2000; Sinha, Balas, Ostrovsky, & Russell, 2006b). Sinha, et al., (2006a) assessed FR performance as a function of image resolution and the impact of presenting internal facial features alone compared to a full face. Images of celebrities were blurred to varying degrees (from 5x7 pixels to 150x210 pixels) and presented to participants. Three different groups of participants were shown images which included either the full faces; only internal features (i.e., eyes, mouth and nose) placed side by side; or, with the same internal features in their original facial configuration. Performance was very poor for the two conditions in which only internal facial features were presented. However, for the full-face condition, FR performance was very robust to image resolution degradation. Participants were able to recognise more than half of the celebrity faces with image resolution of 7x10 pixels, and almost all faces when resolution was 19x27 pixels. This finding demonstrates that when it comes to recognition of familiar faces, the human visual system is able to deal with high levels of degradation.

Burton et al., (1999) assessed face matching and recognition performance on poor quality CCTV images. Results from the first experiment revealed significantly better performance for participants who were familiar with targets. In a subsequent experiment in which only the performance of familiar stimuli was assessed, it was found that even when body and gait information of the target were concealed, 73% of the targets were still successfully recognised. These findings may serve to demonstrate that only familiarity with target individuals would result in good recognition accuracy when images are of poor quality. This highlights some of the issues associated with security surveillance if images of poor quality are to be used for security and legal purposes (Burton, et al., 1999). It has recently been argued that the quality of CCTV and other surveillance imagery may not make it suitable for identification of people. Poor resolution of this imagery may result in many important details needed for identification purposes not being present (Porter, 2006; Kovesi, 2009). As a result, the suitability of such imagery for face identification purposes and as evidence in courts has been debated (Edmond, 2010; Edmond et al., 2010; Porter, 2009).

#### *2.2.1.2 Viewpoint and Illumination*

Viewpoint/pose and illumination have been found to be similar in the way that they affect human face matching and recognition performance. To match a face from a new viewpoint/pose or under different illumination conditions, we must be able to not only access and extract something unique to that face, but we must also be able to access this information from a novel view or under different lighting conditions. Empirical evidence has shown that humans can accurately recognise faces from different views or under different illumination conditions when those faces are familiar (O'Toole, Jiang, Roark, & Abdi, 2006).

However, when faces are not familiar, that is not the case. Viewpoint and illumination have been shown to impact even the most simple one-to-one face matching tasks with no memory load or time constraints on inspection of images. For example, Hill and Bruce (1996) found near-ceiling performance when participants were presented with two images of the same person under the same condition (e.g., profile image lit from top). However, when two presented images of the same person showed individuals in different poses, or were taken under different lighting conditions, performance suffered considerably (Hill & Bruce, 1996). Hill and Bruce (1996) further showed that human ability to process facial identity when faces are lit from below is impaired. Bottom lit images were associated with a disadvantage relating to accuracy compared to top lit images. This is said to happen because in natural settings faces are almost never lit from below (Hill & Bruce, 1996; Sinha, et al., 2006b).

Illumination leads to a change of the overall light intensity reflected back from a face, as well as the distribution of visible shadows. Illumination variation can also produce dramatic differences in the appearance of a face, larger than those associated with changes in viewpoint (Tarr, Kersten, & Bulthoff, 1998).

#### *2.2.1.3 Motion*

In naturalistic settings, faces are almost always in motion. O'Toole, et al., (2006) proposed that facial motion could improve performance as it exposes the viewer to a unique set of facial movements which may be characteristic to a particular individual. Movement also provides structural information about a face which can be used to enhance the quality of the perceptual representation of that face (O'Toole, et al., 2006; O'Toole & Tistarelli, 2010; Thornton & Kourtzi, 2002).

Although, the usefulness of motion seems plausible FR research findings have not been consistent, especially when face familiarity is considered (O'Toole, Roark, & Abdi, 2002). Benefits of motion were shown when recognising familiar faces (Knight & Johnston, 1997; Lander & Bruce, 2003; Pike, Kemp, Towell, & Phillips, 1997; Zhao, Chellappa, Phillips, & Rosenfeld, 2003). More specifically, Knight and Johnson (1997) found that it was easier to recognise famous faces when in motion than from still images. This was also shown for recognition from poor quality security surveillance videos (Bruce, Hancock, & Burton, 1998; Burton, et al., 1999). However, no benefit of motion has been found for the recognition of unfamiliar faces (Christie & Bruce, 1998). It has been suggested that the variability in results relating to unfamiliar face recognition and motion could be attributed to the types of motion tested (i.e., rigid or forced motion as opposed to naturalistic motion)[5] and how the stimuli are presented to participants (O'Toole, et al., 2002; Roark, Barrett, Spence, Abdi, & O'Toole, 2003).

[5] Rigid/forced motion may involve instructed movement. Naturalistic motion happens naturally without any instruction for particular type of movement.

In relation to face matching, research on the effect of motion is limited. Thornton and Kourtzi (2002) evaluated the impact of motion in a face matching task. The face matching task involved first presenting one stimulus (either still image or a moving video) for 540ms, which would be removed and immediacy after the second stimulus (always a still face image) would be displayed. They found significant performance differences between when still and moving imagery was presented first. However, no further work using face matching has been identified.

### *2.2.2 Individual Specific Factors*

#### *2.2.2.1 Ethnicity/Race*

The effect of ethnicity/race of a face on human recognition and matching performance has received a substantial amount of empirical attention. In general, it has been confirmed that people are better at matching and recognising faces of their own race. This phenomenon is referred to as the ***other-race effect*** (Caldara & Abdi, 2006; Chiroro & Valentine, 1995; Furl, Phillips, & O'Toole, 2002; Goldstone, 2003; Johnson & Fredrickson, 2005; Meissner & Brigham, 2001; Slone, Brigham, & Meissner, 2000; Valentine, 1991; Valentine, Chiroro, & Dixon, 1995; Walker & Hewstone, 2006, 2008; Walker & Tanaka, 2003; Young, Bernstein, & Hugenberg, 2010).

The most prominent explanation for the other-race effect is referred to as the ***contact hypothesis***, which asserts that superior recognition ability for one's own race is associated with the amount of exposure to one's own race as opposed to other races. After being exposed to one particular race, humans acquire more detailed perceptual information and, in turn become more expert at distinguishing unique features that characterise faces of their own race (Walker & Hewstone, 2006). It therefore seems plausible to suggest that increased interaction with other-race faces would improve face matching and recognition ability for faces of other races, and could possibly reduce the other-race effect (Caldara & Abdi, 2006; Chiroro & Valentine, 1995; Goldstone, 2003; Walker & Hewstone, 2006). However, results have been mixed. While a number of studies support the claims behind the contact hypothesis (Chiroro & Valentine, 1995; Goldstone, 2003; Meissner & Brigham, 2001; Walker & Hewstone, 2006), there have also been studies that have not found support (Levin, 1996, 2000; Malpass & Kravitz, 1969). Levin (2000) has argued that contact hypothesis ignores the complexities of human social cognition and feature coding differences that are associated with classifying other and same-race faces. However, Furl, et al., (2002) have argued that the lack of consistency among these studies may be due to different testing and analysis methods and the inconsistency in the definition of "contact".

Other explanations, such as ***perceptual learning*** and the ***configural-featural hypothesis*** for the other-race effect have also been suggested. Perceptual learning refers to an increased ability to extract information from an environment as a result of being exposed to stimuli and information from that environment (Gibson, 1969). This is often aligned with increased differentiation which assists with distinguishing between relevant and redundant facial information. It has also been suggested that discrimination of own-race faces is more accurate compared to other-race faces because individuals are able to more accurately discriminate between relevant and redundant facial information (Meissner & Brigham, 2001). The configural-featural hypothesis stems from work relating to face inversion which demonstrates that when highly familiar stimuli are inverted, their processing is affected. This was said to occur because individuals who are familiar with the stimuli rely on configural information, as opposed to individuals to whom the stimuli are not familiar (Meissner & Brigham, 2001). Therefore, Rhodes, Brake, Taylor and Tan (1989) suggested that increased experience with own-race faces would result in a greater inversion effect, while that would not be the case for other-race faces. However, mixed results have been reported. Some confirm this assertion (Rhodes, et al., 1989), while others found a greater inversion effect for other-race faces (Buckhout & Regan, 1988).

#### *2.2.2.2 Age*

The impact of age on face matching and recognition ability can be considered in terms of the age of the individual who is being observed and the individual who is doing the observing (i.e., the Human Operator). Of interest here is the extent to which age can affect an individual's face matching ability. Research has consistently shown that increasing age is associated with a decline in a number of cognitive abilities, including FR accuracy (Lamont, Stewart-Williams, & Podd, 2005; Salthouse, 2004).

In the same way that it has been reported that people are better at recognising and remembering members of their own racial and gender groups, studies of recognition memory have often found interactions between the age of the observer and age of the face stimuli being observed (Anastasi & Rhodes, 2006; Lamont, et al., 2005; Wright & Stroud, 2002). However, the extent of this effect is often debated. Some studies report that people are generally better at recognising individuals of their own age (Anastasi & Rhodes, 2006), other work has reported this effect for only younger (Wiese, Schweinberger, & Hansen, 2008), or only older adults (Kuefner, Cassia, Picozzi, & Bricolo, 2008; Lamont, et al., 2005).

Different findings have also been reported about the age at which face processing ability fully matures. Some research has suggested that face processing reaches adult levels at the age of 16 (Grill-Spector, Golarai, & Gabrieli, 2008; Itier & Taylor, 2004). However, a review by McKone, Crookes and Kanwisher (2009) has suggested that, although face processing can improve throughout childhood, face processing abilities in adults are similar to those of children as young as four years of age. In line with that, some researchers argue that face processing improvements beyond childhood can be explained by differences in attention, concentration and/or general memory rather than any explicit changes and improvements relating to face perception mechanisms specifically (Crookes & McKone, 2009; Itier & Taylor, 2004). However, contrary to previous findings, a recent study has reported that the ability to learn and recognise unfamiliar faces improves until the early 30s (Germine, Duchaine, & Nakayama, 2011).

#### *2.2.2.3 Gender*

The impact of gender on face matching and recognition performance can be considered from the perspective of the Human Operator's gender, as well as the gender of the stimuli or the person who is being viewed.

In relation to the impact of Human Operator gender on matching and recognition performance, the findings have not been consistent. There has been some support for the existence of an own-sex bias (equivalent to an own-race bias) (Rehnman, 2007; Rehnman & Herlitz, 2007). Some research has consistently demonstrated a female advantage for recognising female faces independent of ethnicity and age of those faces (Lewin & Herlitz, 2002; Rehnman & Herlitz, 2007; Shepherd, 1981; Slone, et al., 2000; Wright & Sladden, 2003). It has also been found that females are better at remembering faces and outperform males even on remembering and recognising male faces (Frias, Nilsson, & Herlitz, 2006; McKelvie, Standing, Jean, & Law, 1993; Rehnman, 2007; Rehnman & Herlitz, 2007). Similarly, some studies report that males are better at recognising their own gender, thus also reporting own gender bias (Cellerino, Borghetti, & Sartucci, 2004; Wright & Sladden, 2003). However, other studies have even found that males recognised more female than male faces, or that they remembered and recognised male and female faces equally well (Godard & Fiori, 2010; Lewin & Herlitz, 2002; McKelvie, et al., 1993; Rehnman, 2007).

Work reported above focuses on assessing gender differences using the recognition paradigm. Megreya, Bindemann and Havard (2011) assessed face matching performance in two experiments and found that female participants were more accurate compared to male participants. More specifically, female participants also demonstrated an own-sex bias, however, only when assessing true stimuli pairs. It was further found that female own-sex bias on true stimuli persisted while matching the internal and external facial features. These findings are consistent with recognition research which reports female FR advantage.

In terms of which faces are better recognised, male or female, yet again, mixed findings have been reported. Some results have demonstrated that female faces are better remembered and recognised because they are found to be more distinct (Godard & Fiori, 2010; Rehnman & Herlitz, 2007; Slone, et al., 2000). However, other findings have indicated that male faces are better recognised (Cellerino, et al., 2004; Hill & Bruce, 1996).

### *2.2.3 Human Operator Specific Factors*

#### *2.2.3.1 Fatigue and Stress*

It has long been reported that an individual's ability to maintain attention and vigilance gradually decreases when required to perform demanding cognitive tasks for long periods of time (Parasuraman, 1986; Parasuraman et al., 2009). Face matching and recognition tasks in applied settings are often performed continuously for long periods of time. Human Operators performing those tasks have to maintain high levels of attention and vigilance, especially as the probability of impostors may be low. Early work by Parasuraman (1986) has demonstrated that decrements in attention can lead to decreases in performance in monitoring and search tasks.

Fatigue is often reported after inadequate sleep, or long periods of physical and/or mental exhaustion and can affect an individual's emotional, behavioural, and cognitive functioning. It is said to be associated with a decrease in motivation, an aversion to effort and a decrease in overall quality of work and performance (Beurskens et al., 2000; Robert & Hockey, 1986). Empirical evidence has demonstrated that fatigue, stress and distractions (e.g., additional tasks) can affect human performance across different settings (see Staal (2004) for an overview). Additional tasks such as checking biographical data or even consulting certain decision aids can redirect attention away from the primary task, or can even reduce attentional resources available to sufficiently attend to either task.

Research on multitasking has suggested that the introduction of subsequent and additional tasks distracts from the primary task and thus, can negatively affect performance (Haga, Shinoda, & Kokubun, 2002; Staal, 2004). It has also been reported that performance decreases under time pressure compared to performance in non-time pressured conditions (Ariely & Zakay, 2001; Staal, 2004). Face matching tasks in applied settings are often performed for long periods of time and under time pressures, often in conjunction with additional tasks (which may be distracting to the primary face matching task).

Lee, Vast and Butavicius (2006) investigated the impact of attention, time pressure, and task demands on face matching performance accuracy. Separately, high time pressure and an additional task caused a small and non-significant decrease in performance. Only when combined, time pressure and an additional task resulted in a significant performance decrease. In a related study, Fletcher, Butavicius and Lee (2008) assessed the impact of increased attention on internal facial features on face matching performance. In this study eye-tracking was used and stimuli were presented for 2 or 6 seconds. Attention to internal facial features was greater in the 2-second condition. However, performance was higher in the 6- compared to 2-second condition. Also, in the 2-second condition faster responses were associated with lower performance, indicating a speed-accuracy trade-off. It appears that face matching may be improved if more time is allowed to process additional aspects of the face. However, Fletcher, et al., (2008) found that when 6 seconds were given, accuracy improved by only 3%. They therefore suggested that visual information essential for face matching can be collected in under 6 seconds, as the mean response time in the 6-second condition was approximately 3.5 seconds.

#### *2.2.3.2 Individual Differences*

Although, general research findings suggest that human performance with unfamiliar faces is poor, a considerable amount of variation in performance has been reported. Performance on different matching tasks with unfamiliar faces has been reported to range from 50% to 100% (Lee, et al., 2006; Megreya & Burton, 2006). This may serve to indicate that some individuals are better at performing this task compared to others. It would therefore be useful to consider these differences more closely to understand why some individuals may be better.

In line with this, recent research assessed the possibility of super-recognisers – individuals who perform statistically better compared to the general public on recognition tasks (Russell, Duchaine, & Nakayama, 2009). In a series of two experiments, Russell, et al., (2009) assessed four such individuals and found that their performance was significantly higher when recognising upright and inverted faces compared to that of control groups, who performed better compared to developmental prosopagnosics (Russell, et al., 2009).[6] Interestingly, Russell, at al., (2009, p. 256) stated that "super-recognisers are about as good as many prosopagnosics are bad". This finding shows the existence of an "extraordinary FR ability" and serves to demonstrate that the range of human recognition and face processing ability is wider than previously thought.

Another related FR study by Li et al., (2010) found that extroverts correctly recognised more faces compared to introverts. More specifically, it was the gregariousness facet which reflects the degree of inter-personal interaction that positively correlated with FR ability. Other extraversion facets – warmth, excitement-seeking, assertiveness, activity, and positive emotion – were not found to correlate with FR performance (Li, et al., 2010). Although this research focuses on recognition tasks, it provides a justification for further work to examine if personality traits are able to predict face matching performance. Such knowledge may have the potential to assist with selection of personnel most suitable for face matching tasks.

In terms of face matching tasks, Schretlen, Pearlson, Anthony and Yates (2001) found a positive correlation between performance on the Benton Facial Recognition Test with perceptual speed and total cerebral volume. Megreya and

[6] Prosopagnosia, also referred to as face blindness, is an impairment in the recognition of faces (Prosapognosia Research Centres).

Burton (2006) conducted a more comprehensive assessment and found modest, but significant correlations between performance on 1-to-10 face matching task and tests of perceptual speed, visual short-term-memory, and figure matching. They further found large individual differences on unfamiliar face matching, and that unfamiliar face matching performance did not correlate with recognition of familiar faces but that it did correlate with matching inverted familiar and unfamiliar faces (Megreya & Burton, 2006).

#### *2.2.3.3 Expertise*

The impact of Human Operator expertise on face matching performance has not received much empirical attention. This may be attributed to a common assumption that because face recognition and matching are performed as part of daily social interactions and have long been conducted within various security settings, that people are very good at these tasks and no assessment of their performance is required. Face matching training practices and procedures have therefore not been widely developed and applied. Exceptions are predominantly within the realm of forensics, such as facial reconstruction from eyewitness memory and the use of facial mapping to match faces from crime scenes (Spaun, 2009). However, scientific rigour and reliability of these procedures and operational practices have been questioned (Edmond, 2010; Edmond, et al., 2010). Aside from this, for the purposes of many other applied applications (e.g., Customs, Immigration, policing, Army, access control security) face matching training is most commonly acquired on-the-job. However, recent interests in biometrics, and FR in particular, have highlighted the importance of face matching expertise.

One of the earliest studies to assess the impact of training on face processing was conducted by Woodhead, Baddeley and Simmonds (1979). In a series of three experiments evaluating performance on a recognition and (many-to-many) face matching tasks, training was not found to improve performance. The authors suggested that the reason why training did not improve performance may have been because of the inaccurate assumptions about how face matching and recognition occur. The focus of the training course was on the assessment and comparisons of individual facial features rather than being focused on the face as a whole (Woodhead, et al., 1979). Furthermore, Woodhead, et al., (1979) stated that as face processing is integral to our lives and daily social interactions, it is perhaps, over-learned and may not be improved even by structured training.

Burton, et al., (1999) compared face memory performance of 20 police officers with an average of 13.5 years of service to students who were familiar and unfamiliar with presented faces. All participants were first shown 10 poor quality CCTV type videos and told that they would need to later identify shown individuals. After 1 minute rest period, participants were shown 20 high quality still images and asked to rate each image from 1, indicating that the presented face definitely did not appear in the videos, to 7, indicating that the face definitely appeared in the videos. Participants who were familiar with presented faces performed significantly better compared to both lay participants who were unfamiliar with presented faces and the police officers. Also, no statistical difference in the performance of participants who were unfamiliar with faces and the police was found.

Lee, Wilkinson, Memon and Houston (2009) compared performance of trained and untrained Human Operators. The trained group consisted of individuals enrolled in human identification course at the University of Dundee. Participants compared 12 greyscale CCTV clips (15 seconds in duration) and greyscale still photographs. In line with the previous two studies, this study also did not find a difference between the performance of trained and untrained participants. Even when further dividing the trained group of participants into those who had 1 year of experience compared to those who had three or more years, no difference in performance were found (Lee, et al., 2009).

Unlike the previous three experiments Wilkinson and Evans (2009) found that two experts performed better compared to an untrained group, and concluded that training and experience in facial analysis results in more reliable and accurate performance compared to that of the general public. However, this study has been criticised for being “misleading” (Edmond, et al., 2010). Edmond et al., (2010) are particularly concerned with claims made by this study in terms of providing support for the use of expert evidence in court. They claim that the real problem here was not whether experts performed better compared to lay people, but whether their opinions are sufficiently scientifically reliable to be admitted into courts (Edmond, 2010).

More recently, Semmler, Ma-Wyatt, Heyer, and MacLeod (2012) evaluated the effectiveness of a simple face matching training protocol on performance of a one-to-one face matching task. Two groups of novice Human Operators matched facial stimuli while having their gaze tracked through an eye tracker. While the control group (i.e., the free viewing group) was allowed to make their face matching decisions by focusing anywhere on the image, the training group was required to fixate their gaze on four different internal facial features. During the second stage, both groups – the free viewing and the "trained" – matched the same stimuli. Findings showed no difference between the two groups. The authors argued that this may be due to the short duration of training, in terms of viewing time and the number of stimuli that were presented (Semmler, et al., 2012).

For certain biometric and forensic applications, such as signature and fingerprint analyses conducted by Human Operators, training procedures and applied practices have long been in place (Dewhurst, Found, & Rogers, 2008; Dror & Cole, 2010; Dror & Mnookin, 2010). Consequently, an extensive body of work within this field has assessed differences between signature/document examination experts and lay people. This work has demonstrated that forensic document examiners (FDEs) performed more accurately but also differently compared to lay individuals (Bird, Found, & Rogers, 2010; Dewhurst, et al., 2008; Dyer, Found, & Rogers, 2006).

Empirical work on fingerprint analysis has also been extensive. This work has served to demonstrate that decisions made by fingerprint experts can be significantly affected by extraneous context and information (Dror & Charlton, 2006; Dror, Charlton, & Pe´ron, 2006). In one of the most controversial studies, five highly experienced fingerprint examiners were unknowingly presented with latent prints that they had previously examined and declared to be a match. During this study, the prints were presented in a context which would suggest that they were not a match (i.e., with most of the supporting evidence pointing that way). Of the five fingerprint experts, only one did not change their original opinion; three changed their decisions and with that contradicted their own previous identification decisions; and one was not able to make a definite conclusion (Dror & Charlton, 2006). This indicated that even highly trained and experienced individuals are able to be influenced by contextual information. Further, Dror and Cole (2010) provide a concise overview of the many other factors that affect performance of forensic pattern examiners (e.g., emotional context, expectation and motivation). They call for a thorough empirical exploration of the many factors and their impact on perception and decision making within forensic science as well as other related areas of expertise (Dror & Cole, 2010).

### *2.2.4 The Impact of Impostors*

Impostors, foils or mismatched presentations can be thought of as distractor stimuli which are presented alongside true or matched stimuli in face matching and recognition experiments to aid performance assessments. The impact of impostors on applied face matching performance can be considered in two distinct ways. The first relates to impostor types or categories and the extent to which experimentally generated impostors reflect what happens in the real world. The second relates to the frequency or prevalence of impostors presented in experimental settings and actually present in real world settings. Although empirical research has paid much attention to stimulus creation and presentation, it has not explicitly focused on the impact of different impostor types or different frequencies of impostor stimuli and their impact on face matching performance. This is especially of relevance as the occurrence of impostors in applied settings is likely to be very low (Bindemann, Avetisyan, & Blackwell, 2010; Hillstrom, Sauer, & Hope, 2011). The following paragraphs discuss the most common approaches that research has taken to deal with impostor types and frequencies.

The majority of research relies on the generation of a single type of impostor stimulus that depicts a face that is of a different, but similar looking individual to the target face (Fletcher, et al., 2008; Henderson, Bruce, & Burton, 2001; Megreya & Burton, 2006, 2007, 2008; Vast & Butavicius, 2005). Other experimental work has relied on the generation of “easy” versus “difficult” impostor categories (Kemp, et al., 1997; O’Toole et al., 2007). An “easy” stimulus set would involve face images that are very different to that of the target so that the target is generally easy to detect. A “difficult” stimulus set would involve presentations of stimuli that is very similar to the target but is in fact different, making the target difficult to identify. Also, those impostor types are generated in a number of different ways. In some experiments those selections are made by the experimenter (Butavicius, et al., 2008; Fletcher, et al., 2008; Kemp, et al., 1997; Vast, 2004). Other experiments use computer algorithms to produce similarity ratings (O’Toole, Phillips, et al., 2007); or a separate group of participants who produce similarity ratings on which stimulus pairings/presentations are prepared (Bruce, et al., 1999; Henderson, et al., 2001; Megreya & Burton, 2006, 2007, 2008). Consequently, it would be useful to consider different ways in which impostors can be generated in the real world (Section 3.3) and if different impostor types have a different impact on face matching performance.

For impostor frequencies, the majority of experiments use the 50% impostor rate (Bruce, et al., 1999; Hillstrom, et al., 2011; Kemp, et al., 1997; Megreya & Burton, 2006, 2007, 2008). However, there are examples where 20% (Butavicius, et al., 2008) or even 10% rates were used (Vast & Butavicius, 2005). Since it is believed that occurrences of impostors in the real world are not very common, the extent to which current empirical performance results reflect the real world Human Operator performance is not known.

Perhaps, a starting point would be to assess if Human Operator performance is affected by the presentation of different frequencies of impostors (Chapter 4). A study by Bindemann, et al., (2010) compared performance on impostor detection when participants were presented with 2% and 50% of impostors and did not find a difference in impostor detection for the conditions. They suggested that these results imply that the rare occurrence of impostors in applied settings would not impair a Human Operator's ability to detect them (Bindemann, et al., 2010). It could equally be suggested that their finding provides preliminary evidence that the current practice to predominantly utilise a 50% impostor rate in research does not underestimate the difficulty of impostor detection. As part of current research (Chapter 4), Bindemann, et al.'s work is extended by assessing the impact of four impostor frequencies and four impostor types on Human Operator one-to-one face matching performance.

### *2.2.5 Experimental versus Live Evaluations*

Applications such as passport control, identity verification by police officers, security guards, and similar, are probably the most common examples of face matching in the real world. In such environments a combination of factors can, positively or negatively, affect Human Operator performance. Face matching and recognition research is commonly conducted in experimental/laboratory environments where it is possible to control for certain variables in order to appropriately assess the impact of others. It is therefore legitimate to ask if performance rates obtained in controlled experimental settings will translate into applied settings, especially in relation to photo to live person matching. In applied settings Human Operators are usually approached by an individual who presents a token containing an image. Conducting experiments where participants acting as Human Operators are presented with *live* individuals who present a photograph for matching are logistically difficult. Perhaps for that reason, there have, to date, been only a small number of *live* evaluations.

Kemp, et al., (1997) assessed human performance on a live person-to-photo face matching task. In an attempt to simulate the real world, this experiment was conducted in a supermarket, and is often described in the following way. Six female cashiers were asked to verify the identities of 44 live shoppers by matching them to a photo on the credit card that they had presented. Correct identifications were made on only 67% of occasions, with cashiers falsely accepting more than 50% of the fraudulent presentations. These results are often cited as baseline of Human Operator applied face matching performance. While to some extent these results confirm the difficulty that humans have with matching unfamiliar faces, they have to be considered with caution. When reporting these results, it is often neglected that in addition to assessing the photographs, the cashiers also verified shoppers' signatures. In all situations the signatures were the true signatures of the shoppers. This may have affected the obtained results in two ways. First, the demands of an additional signature verification task may have distracted the cashiers from the primary face matching task. Second, as signatures always matched, there is a possibility that positive matching decisions made by cashiers were based on the signature alone.

The second live-to-photo matching experiment was conducted by Butavicius, et al., (2008). During this experiment 10 (nine male) Defence Force personnel verified identities of 50 *live* individuals by matching them to a photo on an ID card that they had presented. Human Operators were seated in separate rooms where they were provided with a laptop which was used to record their decisions. Overall 95% of correct decisions were made compared to 67.4% reported by Kemp, et al., (1997). Butavicius, et al., (2008) attribute this difference to methodological differences. Perhaps, it would also be useful to consider that 10 judges who took part in Butavicius et al., (2008) experiment were Defence Force personnel who, compared to cashiers or lay persons, may have a heightened sense of security and more motivation to perform better in such a task.

In the third study, Megreya and Burton (2008) conducted three experiments with the aim to assess the impact of *liveness* on a set of face memory and matching tasks. The third experiment was the most similar to the two previous ones conducted by Kemp, et al., (1997) and Butavicius et al., (2008) in that the face matching task was a one-to-one task. In the static condition participants viewed a static video image and a high quality digital photograph on a projector screen. In the *live* condition participants viewed a live target and a high quality digital photograph on a projector screen. In both conditions stimuli were presented simultaneously and participants were asked to make a decision about if they were of the same individual. Overall accuracy for the static video image presentation was 84% and 83% for the *live* condition. The authors found a statistical difference between the static and the *live* condition when impostor and real presentations were considered separately. It was found that in the *live* condition, participants were more likely to claim that two images were a match. Consequently, this resulted in more correct responses and false positives compared to the static video image condition. This indicates a response bias in the *live* condition for participants to claim that two items are a match.

In considering the reported research findings it seems rather striking that such notable performance differences can be found in what appears to be a simple matching task. Overall accuracy in the *live* one-to-one face matching task varied significantly with Kemp, et al., (1997) reporting 67%, Butavicius, et al., (2008) reporting 95%, and Megreya and Burton (2008) reporting 83% accuracy. In considering these results it is important to take into account the details of the matching tasks. Although the tasks in each experiment seem very similar, there are methodological differences that may have contributed to performance variations and therefore, need to be taken into consideration. More broadly, though, these findings are consistent with previous research which has shown that unfamiliar face matching is challenging to the human perceiver.

Previously, Chapter 1 introduced the research problem, focusing on applied one-to-one face matching. This chapter has considered Human Operator face processing and the many factors that can affect face matching and recognition performance, described next is Image Preparation, outlining the preparation of stimuli which were used during performance experiments.

# *Chapter 3*

## *Image Preparation: Acquisition, Sourcing, Look-Alike Definition and Impostor Generation*

This chapter outlines the way that stimuli (i.e., still and video imagery) were collected and prepared for use in the subsequent performance experiments. In addition to providing a methodological overview, this chapter also provides applied reasoning for the methodological decisions.

As shown in Figure 4, the Image Preparation phase consisted of five distinct stages, discussed below.[7]

[7] Preliminary sections of the methodology were presented during the early stages of this research (Calic, 2007, 2008; Calic, McLindin, & MacLeod, 2009).

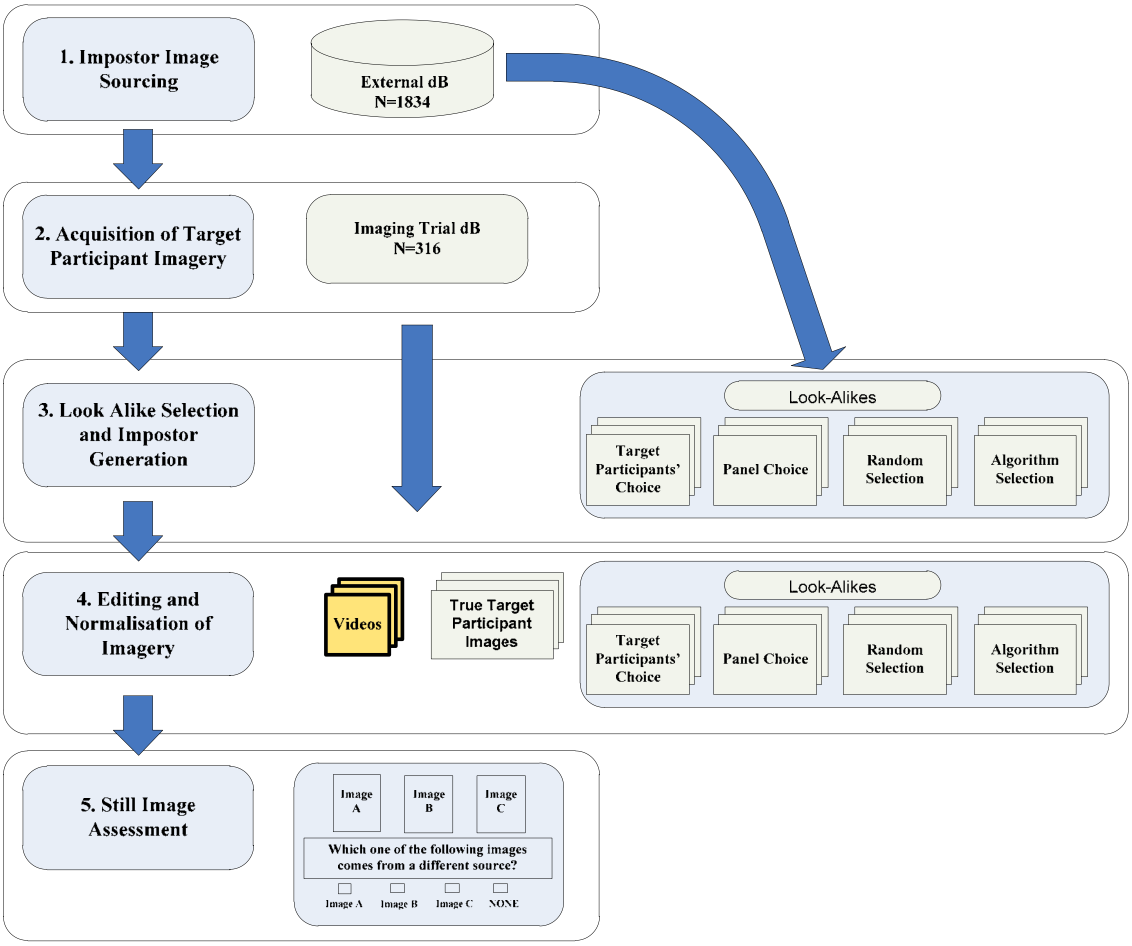

1. Impostor Image Sourcing
External dB
N=1834
2. Acquisition of Target Participant Imagery
Imaging Trial dB
N=316
3. Look Alike Selection and Impostor Generation
Look-Alikes
Target Participants' Choice
Panel Choice
Random Selection
Algorithm Selection
4. Editing and Normalisation of Imagery
Videos
True Target Participant Images
Look-Alikes
Target Participants' Choice
Panel Choice
Random Selection
Algorithm Selection
5. Still Image Assessment
Image A
Image B
Image C
Which one of the following images comes from a different source?
Image A
Image B
Image C
NONE

*Figure 4: Five components of Image Preparation*

1. ***Impostor Image Sourcing (Section 3.1).*** Still images, some of which input into the generation of impostor stimuli, were sourced from publicly available databases.
2. ***Acquisition of Target Participant Imagery (Section 3.2).*** Still and video imagery was acquired during the Imaging Trial. This imagery mainly input into generation of true stimuli.
3. ***Look-Alike Selection and Impostor Generation (Section 3.3).*** Still images

that input into impostor stimuli were selected and impostors were generated.

4. ***Editing and Normalisation of Imagery (Section 3.4).*** All imagery was edited to comply with DFAT Passport Photograph Guidelines (Department of Foreign Affairs and Trade (DFAT), 2005).
5. ***Assessment of Still Imagery (Section 3.5).*** An evaluation was conducted with all edited imagery to assess its suitability for use in the experiments.

The following sections detail each of the five stages of Image Preparation.

## *3.1 Impostor Image Sourcing*

Impostor image sourcing involved gathering images which were used to select look-alikes and generate impostor stimuli (Section 3.3). Still images were sourced from universities and various research institutions that have their own, or have access to database/s of face images, presented in Table 2. This set of images is referred to as the ***external database***.

*Table 2: External image databases used in this research*

| ***Database Name*** | ***Source*** |
|---|---|
| University of Essex Face Recognition Database | (Spacek, last updated 2008) |
| Indian Face Database | (Jain & Mukherjee, 2002) |
| PAL Face Database | (Minear & Park, 2004) |
| MIT-CBCL Face Recognition Database | (Weyrauch, Huang, Heisele, & Blanz, 2004) |
| Caltech Faces | (Weber, 1999) |
| The Psychological Image Collection at Stirling (PICS) | (School of Natural Sciences (Psychology), accessed 2008) |
| Georgia Tech Face Database | (Georgia Institute of Technology, 1999) |
| Face Database VIS_DB | (Nowosielski, 2006) |
| The Colour Facial Recognition Technology Database | (Phillips, Moon, Rizvi, & Rauss, 2000; Phillips, Wechsler, Huang, & Rauss, 1998) |
| GTAV Face Database | (Tarres & Rama, 2005) |

External databases contained coloured face photographs with neutral facial expression. It was ensured that only one image per individual was selected for use, resulting in 1,824 individual images (724 female and 1,100 male faces).

## *3.2 Acquisition of Target Participant (TP) Imagery*

Still and video imagery was acquired (i.e., the Imaging Trial) and will be referred to as the ***internal database***.[8]

### *3.2.1 Target Participant Recruitment and Participation*

TPs were recruited from the Defence Science and Technology Organisation (DSTO) through a series of site wide emails, the Defence Magazine, and by word of mouth. Three Hundred and sixteen (79 females and 237 males) TPs were recruited. Their ages ranged from 19 to 64 (*M* = 40.26, *SD* = 10.94). TPs were required for the acquisition of still and video images to generate the internal database, and for the creation of TPs' Choice impostor category (Section 3.2.2.2).

### *3.2.2 Experimental Conduct*

Upon arrival, TPs were informed about their participation requirements and the treatment of collected data. They read the Information Sheet (Appendix A), Guidelines for Volunteers (Appendix B) and signed the Consent Form (Appendix C) if they wished to participate.

[8] The majority of the TPs were available to attend both, the Imaging Trial and the subsequent *live* Experiment 2 (Chapter 5).

Each TP was randomly allocated a unique number which was used to track their participation throughout the experiment and subsequently link their imagery with appropriate demographic information. Each TP was also given a Participant Record Sheet (Appendix D) which provided a paper trail of their movement through the experiment.

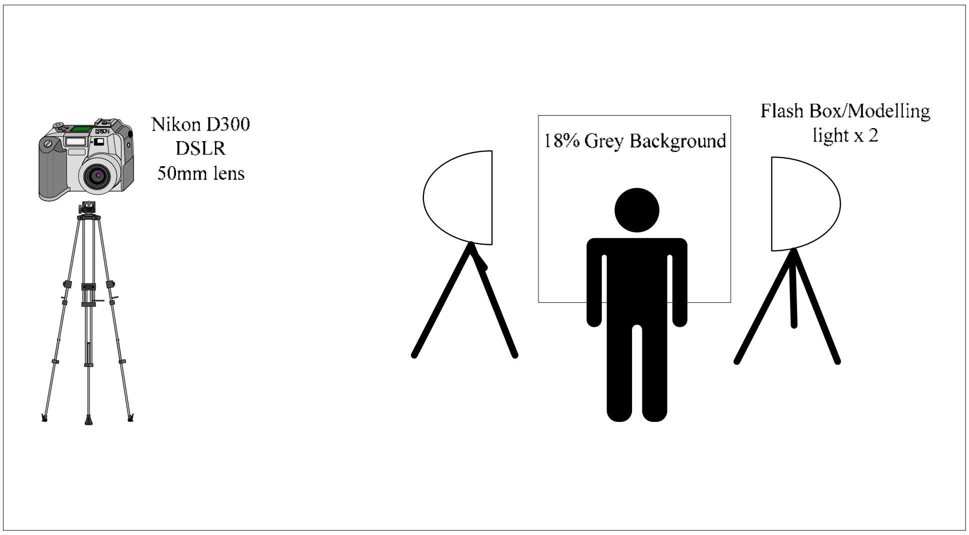
#### *3.2.2.1 Imagery Acquisition*

To facilitate applied relevance of the results, the acquisition of still and video imagery was in compliance with DFAT Passport Photograph Guidelines (Department of Foreign Affairs and Trade (DFAT), 2005). It should be noted that DFAT does not have guidelines for video acquisition. Therefore, general principles associated with still image acquisition (e.g., controlled/uncluttered background, neutral facial expression, etc.,) were also applied to the acquisition of video imagery.

##### *3.2.2.1.1 Still Imagery: Equipment Set Up and Acquisition*

Photographic equipment, a Digital Single Lens Reflex (DSLR) Nikon D300 camera was assembled in a room free of natural light. A 50 mm f 1.4 lens was used and after balancing the lights an f-stop of f11 was set at shutter speed of 1/60th second.

In compliance with DFAT Passport Photograph Guidelines a tripod with the camera was positioned 1.2 m away from where TPs were required to stand (Department of Foreign Affairs and Trade (DFAT), 2005). Two Bowen Esprit GM500 mono block lights with diffusion umbrellas were located 3 m from TPs and positioned one on each side at a 45 degree angle to the TP. The background used 18% grey. TPs were asked to stand at a designated spot in the imaging room, straighten their posture, look directly at the camera, and maintain a neutral facial expression. Two images of each participant were acquired. The imaging set up and TP acquisition process are demonstrated in Figure 5.



NOTE:
These figures/tables/images have been removed to comply with copyright regulations.
They are included in the print copy of the thesis held by the University of Adelaide Library.

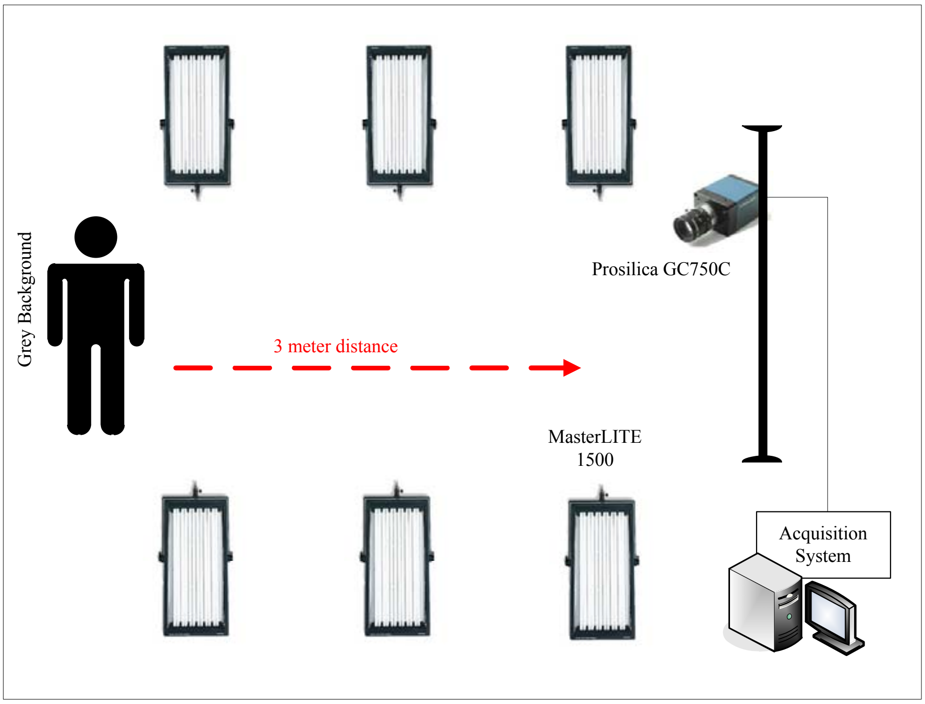
*Figure 5: Image Preparation: Still imagery acquisition equipment set up, acquisition process and an example of a still image*

#### *3.2.2.1.2 Video Imagery: Equipment Set Up and Acquisition*

Video imaging equipment was assembled in a corridor with no windows and minimal natural light. Six Masterlite1500 lights, three on each side, were positioned to ensure sufficient lighting. Light meters were used to adjust the positioning and intensity of lights to minimise shadows and ensure even lighting.

A Prosilica GC750C visible wavelength Gigabit Ethernet video camera was used for the acquisition of video imagery (752 x 480, 60 frames per second (fps)). The camera was stationed at one end of the corridor, allowing approximately 3 meters[9] walking distance to TPs. The starting position from which TPs commenced their approach to the camera was marked with red tape. Grey material was hung from the ceiling behind the TPs. This created a neutral, uncluttered background, in line with DFAT Passport Photograph Guidelines (Department of Foreign Affairs and Trade (DFAT), 2005). TPs were asked to traverse the corridor at a steady walking pace looking straight at the camera, positioned in front of them at head height. Participants were required to remain quiet and maintain a neutral facial expression for the duration of the recording. As they approached the camera they stepped on a mat, positioned approximately 1.2 m from the camera, which had a buzzer underneath to indicate to the TP that they should stop. The equipment set up and acquisition process are exemplified in Figure 6.

[9] This distance is based on operational estimations of how long it would take an individual to walk from the top of the queue line to a Customs Primary Line Officer at an airport (Graves, 2008).



NOTE:
This figure/table/image has been removed
to comply with copyright regulations.
It is included in the print copy of the thesis
held by the University of Adelaide Library.

*Figure 6: Image Preparation: Video imagery acquisition equipment set up and acquisition process*

#### *3.2.2.2 Target Participant Look-Alike Selection and Exit*

Once both still and video imaging were completed, TPs were asked to provide some basic demographic details (e.g., age, gender, and ethnicity). They were then asked to look through the booklet of external database images and select an image that they believed looked the most similar to them. After making their selection TPs were asked to, on a five point Likert scale (ranging from 1 (very confident) to 5 (not confident at all)) rate how confident they were that they would be able to use that image as a form of ID. This completed participant involvement in the Imaging Trial and took approximately 20-30 minutes.

## *3.3 Look-Alike Selection and Impostor Generation*

The previous two sections – impostor image sourcing (Section 3.1) and acquisition of TP imagery (Section 3.2) – addressed the creation of external and internal databases which were used for the generation of image stimuli that input into face matching experiments. This section describes the process by which each type of impostor stimuli was generated from the external database, and provides applied reasoning for the selection of the particular impostor categories.

For the purposes of the current work, look-alike selection facilitated the creation of different types of impostors. This was necessary for the execution of impostor attacks which may have the potential to expose FR systems' vulnerabilities.[10] The creation of different types of impostors, used throughout current experiments, was based on work by Graves, et al., (2003) who identified four classes of impostor attacks, based on the amount of effort required by the attacker. These impostor categories were adopted and modified based on applied considerations. As a result, the following four impostor categories were explored (Figure 7, p.92):[11]

1. ***TP Choice***. An image selected from the external database by each TP as being someone who they believed looked the most similar to them (Section 3.2.2.2). This simulated the situation in which an individual selected another person's identification documentation (e.g., passport) which contains an image that they believe will pass for themselves.
2. ***Panel Choice.*** An image from the external database which was most frequently selected by a panel of six judges as most closely resembling a TP. For each TP, each of the six judges independently looked through the external database of images and chose the one that they believed looked the most similar to that TP. The judges also indicated a five point Likert scale rating about how confident they were that a TP would be able to use the selected image as a form of ID. Imagery that was most frequently selected by all judges, and for which the highest confidence ratings were

[10] In this context *system* does not only refer to an automated algorithm. It refers to the larger applied setting, which may incorporate both, automated and human FR capabilities.

[11] Preliminary work on impostor types was presented at the Cognitive Science Society Conference (Calic & McLindin, 2009).

provided, was used as part of this impostor type. Operationally, this may be equivalent to a person obtaining a fraudulent identification (e.g., passport, visa application or equivalent) generated by group selection, such as a counterfeiting organisation.

3. ***Algorithm*[12] *Selection.*** This was an external database image selected by an automated FR algorithm as closely resembling the TP's image. Each TP image was matched by the FR algorithm generating a similarity score, with the highest being selected. This simulated the situation in which an individual uses fraudulent identification documentation (e.g., passport) that has been selected by an FR algorithm.
4. ***Random (based on Gender and Ethnicity).*** This was an external database image selected using the random function in Microsoft Excel based on the TP's gender and ethnicity. The internal and external databases were divided generically based on gender and then into two ethnicity groups (i.e., White Caucasian and Other[13]). On the basis of these categories a random image was chosen for each TP image (i.e., the internal database) from the corresponding gender-ethnicity category in the external database. This is similar to the zero-effort impostor category which assumes that an unauthorised user is not concerned about the likelihood of their attack

[12] This algorithm is described in Section 6.2.3.2.

[13] This crude division of ethnicities was based only on visible facial information. Had more information about the photographed individual's ethnicities from external databases been available, an attempt would have been made to make a more precise division of ethnicities.

succeeding, and is therefore prepared to use any found token to attempt to obtain access (Graves, et al., 2003). Of the four assessed impostor categories, this would be the one that is least likely to occur.

## *3.4 Editing and Normalisation of Imagery*

The following sections detail the process used to convert, edit, normalise, and prepare internal and external still and video imagery for input into the performance experiments.

### *3.4.1 Still Imagery*

For still imagery (i.e., from internal and external databases) to be used within same experiments, alongside one another, it was important to ensure that images appeared generically similar, and appeared to have been acquired under the same or similar imaging conditions. This was important because, with imagery looking visibly different, there was a concern that during performance evaluations Human Operators would be able to detect impostors by identifying image quality differences between impostor (i.e., external database) images and internal images rather than based on the assessment of a person's face. Consequently, still imagery editing focused on modifying external imagery backgrounds to increase consistency with internal imagery and to make it compliant with the DFAT Passport Photograph Guidelines (Department of Foreign Affairs and Trade

(DFAT), 2005). In addition to the backgrounds, changes were also made to size, contrast, brightness, and ratio aspects of external imagery.

It was first ensured that all images that were used were of Joint Photographic Experts Group (JPEG) image format and they were edited using Adobe Photoshop CS4. Changes included adjusting all images to a ratio of 4:5 (width:height). External imagery backgrounds were standardised to a set Red Green Blue (RGB) value of 160, 146, 142. The mean size of external images was 44.09 KB, with average pixel dimensions of 328x410. For the internal images the mean size was 59.30 KB with average pixel dimensions 331x414. Modifications to still images are exemplified in Figure 7 where each TP image is presented alongside the original impostor image and also after that image had been modified to look similar to TP imagery.

A general note regarding the appearance of modified images needs to be made. In some instances the lack of contrast tended to make images look flat (e.g., facial features were slightly less discernible) and poor image quality resulted in blurring (e.g., eyes not appearing sharp). To manage these limitations an assessment of modified external images alongside internal images was conducted. This assessment helped exclude external imagery which even after modification, still looked visibly different to the internal images and would have the potential to affect Human Operator face matching decisions (Section 3.5).

| ***Impostor Type*** | ***Target Participant Image*** | ***Modified Impostor Image*** | ***Original Impostor Image*** |
|---|---|---|---|
| ***TP Choice*** | | | |
| ***Panel Choice*** | | | |
| ***Algorithm Selection*** | | | |
| ***Random Selection*** (based on Gender and Ethnicity) | | | |

NOTE:
These figures/tables/images have been removed
to comply with copyright regulations.
They are included in the print copy of the thesis
held by the University of Adelaide Library.

*Figure 7: Image Preparation: Exemplifying impostor types and how external images were modified to appear similar to internal imagery*

### *3.4.2 Video Imagery*

The preparation of video imagery acquired during the Imaging Trial (Section 3.2) involved converting and cropping the imagery to ensure that it was in a commonly used format and therefore manageable for experimental work. Video images were converted from the native sequence file (i.e., .seq) into a QuickTime H.264 Movie file (i.e., .mov), the best quality available from the software. Conversion of the files was done using StreamPix Version 4, the same software used to acquire the video imagery (NorPix: Digital Video Recording Software, 2009).

Once converted, each video stream was viewed and only sections that involved TPs approaching the camera were selected. Based on applied estimations of how long it would take an individual to walk from the top of the queue line to the Customs Primary Line Officer, it was decided that each participants' video would be selected for an average of eight seconds (Graves, 2008). This process also ensured that all video imagery was of the same length and that there was no unnecessary footage (e.g., participant standing to receive instructions). MPEG Streamclip was used to crop imagery (Cinque, 2009).

## *3.5 Assessment of Still Imagery*

Having edited all still imagery, an assessment of modified external imagery alongside internal imagery was conducted to ensure that it was, or appeared to be, of the same quality. Based on this assessment some of the external database images were excluded from use in performance experiments. This was done by considering participants' assessments of the images, in terms of external images which participants most frequently believed looked similar and those which they most frequently believed looked different to the internal imagery.

One hundred and ten impostor images were selected and modified (approximately 30 images for each of the four impostor types). An additional 117 internal database images were randomly selected for use in this experiment. This imagery was presented to 19 participants (7 males and 12 females with a mean age of 30.84 years (*SD* = 9.63)) in the form of 76 combinations of three images. Presented imagery consisted of combinations of modified impostor images and internal images (e.g., Internal-Internal-Internal; Internal-Internal-External; Internal-External-Internal; Internal-External-External, etc.). Figure 8 shows how the imagery was presented to participants.

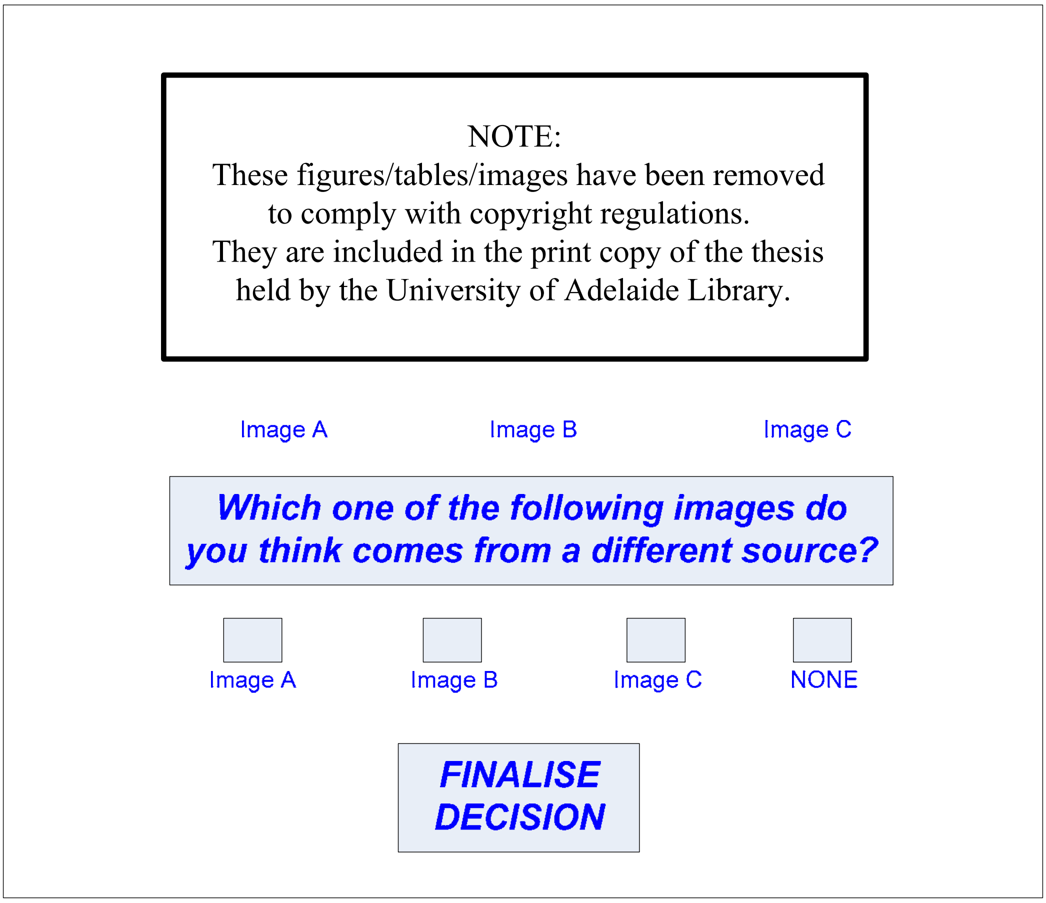


Image A
Image B
Image C
Which one of the following images do you think comes from a different source?
□ Image A
□ Image B
□ Image C
□ NONE
FINALISE DECISION

*Figure 8: Image Preparation: Assessment of still imagery stimuli presentation*

At the commencement of the experiment, participants were informed the following:

> *"You are going to be presented with three photographs that* <u>*may*</u> *come from different sources. Please have a careful look at the presented photographs and decide if there is ONE or NO photographs that come from a different source.*
>
> *Your decision should not be based on the similarity of the physical/facial*

> *characteristics of the people presented in the three photographs.*
>
> *We are interested in the general appearance of the three images presented."*

Once an image was selected, participants were asked to justify (in a free text format) why they selected that image. It was hoped that this additional information would assist with further understanding the most notably distinguishing aspects of internal and external images.

The results of this assessment directly input into the selection of external imagery that would be used throughout the performance experiments and presented alongside internal images (exemplified in Figure 9, on page 104). Image selection was based on participants' responses in terms of which external images they most frequently selected to look similar or dissimilar to internal images. Out of 110 external images that were assessed, 37 images were not selected for use in performance experiments. These images were not used for two reasons. Firstly, on more than 50% of occasions participants deemed them as looking different to the internal images, and secondly, participants never chose them as looking similar to the internal images. The remaining 73 images were determined suitable for use in performance experiments. It was found that while participants deemed those images as looking different to the internal images, this occurred much less frequently (on average on 15% of occasions), and they also believed that those images looked similar to the internal images on 24% of occasions.

An examination of participants' reasons for selecting particular images indicated that differences in lighting and colour were the most commonly cited reasons. Additionally, the importance of image clarity (both poor and good) and image quality, in general, were also emphasised.

The still and video imagery required for use in the face matching experiments was now ready for use (Chapter 4 to Chapter 6).

# Chapter 4

## Experiment 1: Scoping Laboratory One-to-One Face Matching

Experiment 1 sought to assess the impact of impostor frequencies and types on one-to-one face matching performance of untrained Human Operators. It was the first in the series of three face matching experiments and was conducted in a laboratory setting. As previously discussed the aims of this research are ecologically motivated, focusing on one-to-one face matching performance within applied access control settings. In considering applied performance, there are several factors that can affect face matching performance and this experiment focused on the impact of impostors, both in terms of impostor frequency and type.

Previous research has not explicitly focused on operationally relevant generation of impostor stimuli, and only very minimally on the presentation of different frequencies of impostors being operationally based. The most commonly used form of impostor type depicts facial images that are different, but similar looking to the target face (Fletcher, et al., 2008; Henderson, et al., 2001; Megreya & Burton, 2006, 2007, 2008; Vast & Butavicius, 2005). The use of *easy* and *difficult* impostor categories is also very common (Kemp, et al., 1997; O'Toole, Phillips, et al., 2007). These impostor categories have been reported to be generated by either experimenters themselves (Butavicius, et al., 2008; Fletcher, et al., 2008; Kemp, et al., 1997; Vast, 2004); computer algorithms (O'Toole, Phillips, et al., 2007), or a separate group of participants who produce similarity ratings (Bruce, et al., 1999; Henderson, et al., 2001; Megreya & Burton, 2006, 2007, 2008).

Similarly, in relation to impostor frequencies that participants are presented with, the majority of experiments use 50% impostor rate (Bruce, et al., 1999; Hillstrom, et al., 2011; Kemp, et al., 1997; Megreya & Burton, 2006, 2007, 2008). One experiment used 20% (Butavicius, et al., 2008) and another 10% (Vast & Butavicius, 2005). It is commonly believed that the occurrence of impostors in applied settings is very low (Bindemann, et al., 2010; Hillstrom, et al., 2011). It is therefore logical to ask about the extent to which the results from most empirical experiments which use the 50% impostor rate reflect what actually happens in applied settings. Bindemann, et al., (2010) evaluated impostor detection when participants were presented with 2% and 50% of impostors and found no difference in performance. The authors argued that these results imply that the infrequent occurrence of impostors in applied settings would not affect Human Operators' ability to detect impostors (Bindemann, et al., 2010). Similarly, these findings also provide preliminary evidence that the common research practice to use a 50% impostor rate would not affect face matching performance differently than in applied settings where the occurrence of impostors is thought to be significantly lower. Experiment 1 further extends the work conducted by Bindemann, et al., (2010) by evaluating the impact of four impostor frequencies and four impostor types on one-to-one face matching performance. As a scoping experiment, its findings directly input into the methodologies of Experiments 2 and 3.

## *4.1 Method*

### *4.1.1 Participants*

One hundred and fifteen (73 female and 42 male) participants were recruited from The University of Adelaide, School of Psychology student pool and received credit for their participation. They were aged between 17 and 46 (*M* = 21.18, *SD* = 5.19). The participants are referred to as Human Operators.

### *4.1.2 Materials*

#### *4.1.2.1 Target Participant Stimuli*

Of 316 TPs who participated in the Imaging Trial (Section 3.2), imagery from 200[14] (65 females and 135 males) was used in this experiment. Their ages ranged from 19 to 62 (*M* = 39.98, *SD* =11.40).

#### *4.1.2.2 Stimuli Presentation*

Matlab R2009a and the imaging toolbox were used for the display of video and still imagery. Video imagery, acquired during the Imaging Trial, was approximately 7 to 8 seconds in duration and showed a TP walking towards the camera. Still photographs consisted of images which were either, acquired during the Imaging Trial (i.e., used for true stimuli) or selected from the external database (i.e., used for impostor stimuli). Once the experiment was prepared, it was displayed on 17 inch monitors (1152x870 resolution).

[14] Due to time restrictions, no all TPs' images were used in Experiment 1. During the preparation of the experiment it was realised that more than 200 stimuli pairs had the potential to make the experiment over 2 hours in duration. As this had the potential to affect performance by significantly fatiguing participants, it was decided to reduce the number of stimuli pairs to 200.

### *4.1.3 Design and Procedure*

#### *4.1.3.1 Human Operator Performance Testing*

Human Operators were seated in front of a computer monitor and asked to read the Information Sheet (Appendix E). Their consent was indicated by clicking the "NEXT" button on the experimental application. Further details about the experiment were then presented and they were asked to complete a set of demographic questions (e.g., age, gender, ethnicity etc.). Once that was completed, Human Operators were presented with two test examples of experimental stimuli, after which the experiment commenced.

To assess the impact of impostor frequency on face matching performance, three impostor frequency conditions (i.e., 10, 20 and 30%) were considered. Additionally, a control condition with 0% impostor stimuli was also included. Human Operators were randomly assigned to participate in one of the four groups.

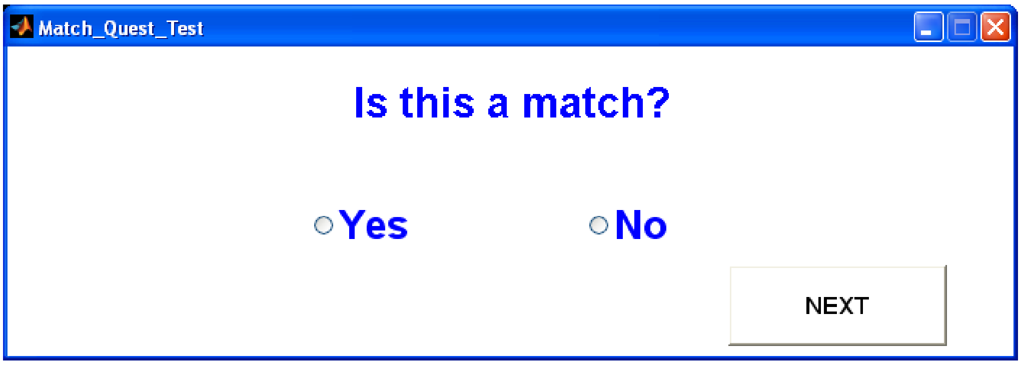
Human Operators viewed 200 randomised stimulus pairs which included a video and a still image. As shown in Table 3 the number of real and impostor images varied according to the impostor frequency condition that Human Operators were assigned to. Within each condition, all four impostor types were presented. Human Operators were presented with 0, 20, 40 or 60 impostor stimuli depending on whether they were in 0, 10, 20 or 30% impostor condition.

*Table 3: Impostor frequency conditions and the number of images presented as part of each condition*

| | ***Impostor Frequency Condition (%)*** | | | |
|---|---|---|---|---|
| ***Image Type*** | **0** | **10** | **20** | **30** |
| | (*n*=25) | (*n*=28) | (*n*=31) | (*n*=31) |
| **TP Choice Impostor** | N/A | 5 | 10 | 15 |
| **Panel Choice Impostor** | N/A | 5 | 10 | 15 |
| **Algorithm Selected Impostor** | N/A | 5 | 10 | 15 |
| **Random Selection Impostor** | N/A | 5 | 10 | 15 |
| **Real Imagery** | 200 | 180 | 160 | 140 |
| **Total** | 200 | 200 | 200 | 200 |

Figure 9 illustrates the way that stimuli were displayed during the experiment. Human Operators were first presented with a TP video, shown in the left corner of the screen. Once the video finished, the last frame of the video (a close-up of the TP) remained on the screen. A still photograph was then displayed in the right corner of the screen. This photograph was either a true photograph of the TP or an impostor photograph, from the four impostor types. With both video and still image displayed, Human Operators were then asked: “Is this a match?” and provided with “Yes” and “No” options.

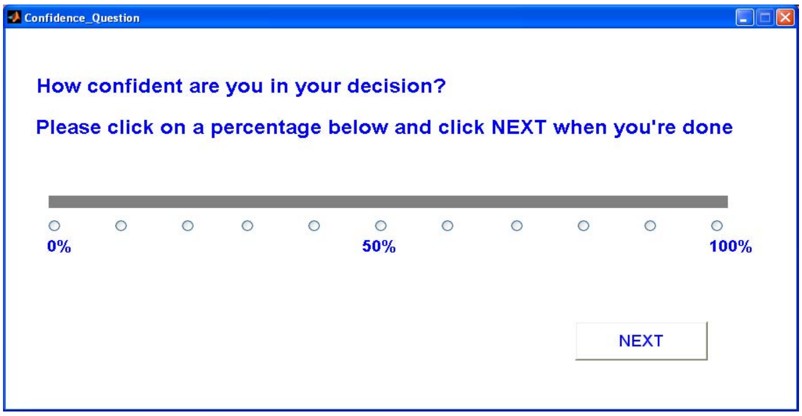
NOTE:
This figure/table/image has been removed
to comply with copyright regulations.
It is included in the print copy of the thesis
held by the University of Adelaide Library.



Is this a match?

*Figure 9: Experiment 1: Stimuli presentation*

After indicating their decision and clicking "NEXT", Human Operators were not able to return to this screen in the event they changed their mind (They were informed about this prior to the commencement of the experiment and were encouraged to carefully consider their decisions). After clicking "NEXT", Human Operators were then asked to indicate how confident they were in their decision on a scale from 0 to 100 percent (Figure 10). This process was repeated for all 200 stimuli. This experiment took on average 1 hour and 26 minutes to complete (ranging from 1:06 to 2:10 hours).



How confident are you in your decision?

Please click on a percentage below and click NEXT when you're done

*Figure 10: Experiment 1: Decision confidence rating scale*

## *4.2 Analysis*

This section provides an overview of how performance data was analysed. The same analysis method was used in all three face matching experiments.

In addition to performance rates/percentages, performance was also analysed using Signal Detection Theory (SDT) statistics (Section 1.2.2.1). Human Operator ***sensitivity*** or ***discrimination*** was represented using *d'* (dee-prime) which incorporates hits and false alarms to measure the distance between the signal and noise means in standard deviation units. ***Criterion*** or ***bias*** was measured using $\beta$ (beta) which estimates Human Operator's tendency to respond "yes" or "no"

(Stanislaw & Todorov, 1999). In addition to these measures, Receiver Operating Curves (ROC) were used to visualise performance. ROCs, shown in Figure 11, are commonly used to demonstrate the relationship between false alarm rates presented on the *x*-axis and hit rates presented on the *y*-axis. Figure 11 shows examples of different curves which correspond to different levels of sensitivity. The higher the value of *d'* the better the ability to distinguish between signal and noise (Macmillan & Creelman, 2005; McNicol, 2005; Stanislaw & Todorov, 1999; WISE (Web Interface for Statistics Education)).

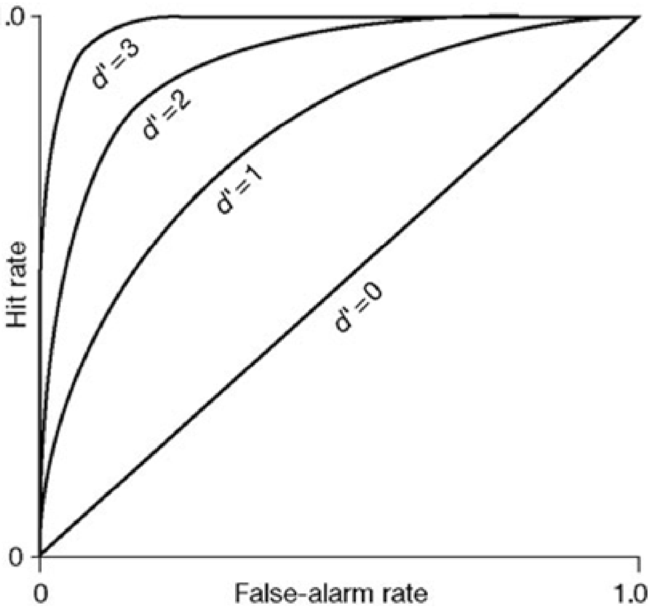


*Figure 11: The ROC curve for different values of d'*

SDT statistics were obtained by using Human Operator confidence ratings. Although, Human Operators were asked to make a yes-no binary decision, these decisions were followed by a confidence scale rating, which were used in line with SDT rating experiments.[15]

Furthermore, within the current context, where the interest is to examine Human Operator ability to detect impostors, it is appropriate to define indices of performance with the focus on impostor-related decisions. Therefore, terminology used to explain performance is defined differently to the conventional way applied by the SDT where the focus is on ***true*** stimuli and when participants accurately respond "yes" (Macmillan & Creelman, 2005; McNicol, 2005; Stanislaw & Todorov, 1999). Here, the focus is on ***impostor*** stimuli and when Human Operators accurately respond "no" or reject impostors. The terminology used throughout this thesis is presented in Table 4 and is defined in the following way:

1. ***Hit Rate*** (or Correct Reject Rate) is the proportion of impostor stimuli that was correctly rejected.
2. ***Miss Rate*** (or False Accept Rate) is the proportion of impostor stimuli that was not detected.
3. ***False Alarm Rate*** (or False Reject Rate) is the proportion of true stimuli that was incorrectly rejected.
4. ***Correct Response Rate*** (or Correct Accept Rate) is the proportion of true stimuli that was accepted.

[15] The way that measures are calculated and ROCs obtained in rating experiments is covered by Macmillan and Creelman (2005), McNicol (2005), and Staislaw and Todorov (1999).

*Table 4: Response decision matrix*

| | ***Human Operator Responses*** | |
|---|---|---|
| | ***No*** (Different) | ***Yes*** (Same) |
| Impostor Stimuli<br>(i.e., mismatching stimuli or different individuals) | Hits | Misses |
| True Stimuli<br>(i.e., matching stimuli or same individuals) | False Alarms | Correct Responses |

To obtain an overall estimate of performance accuracy, all correct decisions were combined by considering Hits and Correct Responses.

## *4.3 Results*

Experiment 1 assessed the impact of four impostor frequencies and four impostor types on Human Operator one-to-one face matching performance. The following sections are divided accordingly to report the findings.

### *4.3.1 The Impact of Impostor Frequency*

As the focus of Experiment 1 was to assess Human Operator ability to detect impostors, the analyses predominantly considered the 10, 20, and 30% impostor frequency conditions, with only a minimal consideration of the 0% impostor condition. Overall accuracies across all four conditions ranged from 62 to 100% (*M* = 94.45, *SD* = 7.90), presented in Table 5.

*Table 5: Experiment 1: Overall accuracies and SDT measures by impostor frequency*

| ***Impostor***<br>***Frequency (%)*** | ***Overall Accuracy***<br>*M (SD)* | ***Mean d'***<br>*M (SD)* | ***Mean β***<br>*M (SD)* |
|---|---|---|---|
| **0** | .90 (.14) | NA | NA |
| **10** | .96 (.05) | 3.59 (.66) | 1.22 (1.67) |
| **20** | .96 (.04) | 3.80 (.52) | .89 (1.56) |
| **30** | .95 (.05) | 3.65 (.68) | .80 (1.44) |

Comparing overall accuracies for the four impostor frequency conditions revealed that the 0% condition differed to the 10, 20 and 30% conditions. However, a one-way Analysis of Variance (ANOVA) found no statistically significant differences between the four conditions, *F*(3, 111) = 2.57, *p* = .06. Furthermore, overall accuracies for the 10, 20 and 30% conditions were not found to be statistically different, *F*(2, 90) = 0.67, *p* = .52. This was further confirmed when mean *d’* measures were considered, *F*(2, 90) = 0.86, *p* = .43.

Performance rates are shown in Table 6 and follow a similar pattern to overall accuracies. A one-way ANOVA revealed no statistical differences in the rates between the 10, 20, and 30% impostor conditions. Specific results were: Hit Rate (*F*(2, 87) = 1.08, *p* = .35); Miss Rate (*F*(2,87) = 0.89, *p* = .41); False Alarm Rate (*F*(2, 87) = 0.11, *p* = .90); and Correct Response Rate (*F*(2, 87) = 0.13, *p* = .88).

*Table 6: Experiment 1: Performance rates by impostor frequency*

| | ***Impostor Frequency (%)*** | | |
|---|---|---|---|
| | **10**<br>*M (SD)* | **20**<br>*M (SD)* | **30**<br>*M (SD)* |
| **Hit Rate** | .91 (.12) | .94 (.08) | .91 (.10) |
| **Miss Rate** | .09 (.12) | .06 (.08) | .09 (.10) |
| **False Alarm Rate** | .04 (.06) | .03 (.06) | .04 (.07) |
| **Correct Response Rate** | .96 (.06) | .97 (.06) | .96 (.06) |

However, when the 0% impostor condition was considered, it was found that False Alarm Rate (*M* = .09, *SD* = .14) and Correct Response Rate (*M* =.91, *SD* = .14) of the 0% impostor condition were significantly different to the other three conditions' False Alarm Rates (*F*(3, 111) = 2.96, *p* = .04) and Correct Response Rates (*F*(3, 111) = 3.18, *p* = .03).

Having considered performance rates by impostor frequency, Human Operator performance is presented graphically using ROC curves (Figure 12). The figure visually confirms that the performance of the three impostor conditions was very similar.

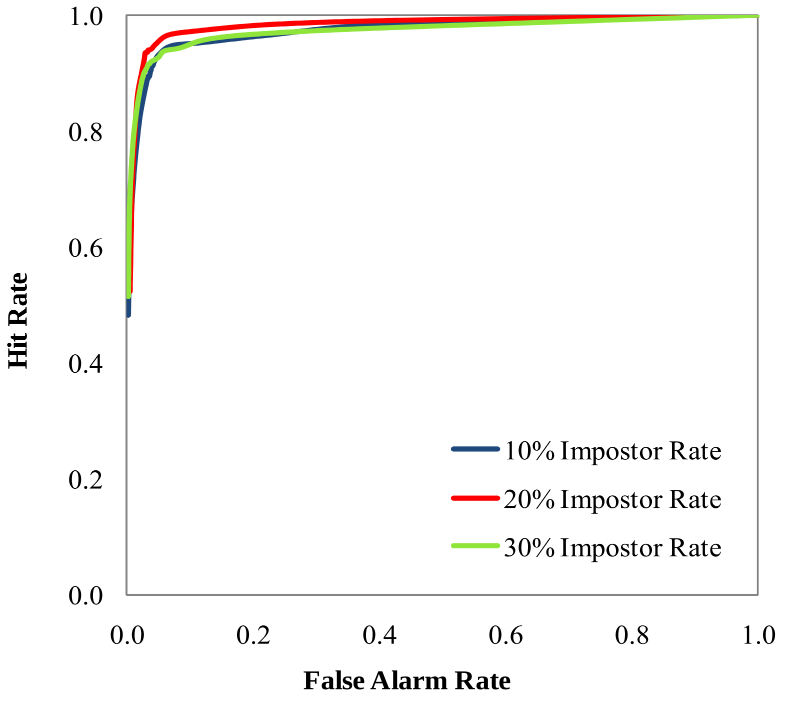


*Figure 12: Experiment 1: ROCs by impostor frequency conditions*

### *4.3.2 The Impact of Impostor Types*

In terms of impostor and true stimuli presentations, it was found that performance was better on true (*M* = .96, *SD* = .06) compared to impostor (*M* = .92, *SD* = .09) stimuli. This difference was statistically significant, $t(89) = 151.17$, $p < .001$.

These results were further considered in terms of different impostor types by each impostor frequency condition, shown in Table 7.

*Table 7: Experiment 1: Hit Rates for impostor types by impostor frequency*

| | ***Impostor Frequency*** | | |
|---|---|---|---|
| | **10** | **20** | **30** |
| ***Impostor Type*** | *M (SD)* | *M (SD)* | *M (SD)* |
| **TP Choice** | .84 (.25) | .89 (.15) | .89 (.12) |
| **Panel Choice** | .87 (.19) | .88 (.20) | .83 (.16) |
| **Algorithm Selection** | .96 (.09) | .99 (.03) | .97 (.06) |
| **Random Selection** | .96 (.08) | .99 (.02) | .98 (.06) |

Analysis of variance revealed a statistically significant main effect for impostor type ($F(2.13, 183.45) = 27.43$, $p < 001$, Partial $\eta^2 = .24$) but not for interaction of impostor type and frequency ($F(4.26, 183.45) = 1.05$, $p = .39$, Partial $\eta^2 = .02$).[16]

[16] Partial (eta) $\eta^2$ is a measure of effect size and indicates what proportion of the variance in the dependent variable is attributable to the factor in question (Dancey & Reidy, 2002).

A repeated measures ANOVA further revealed a significant effect of impostor type within the 10% impostor condition ($F(3, 28) = 5.70$, $p = .007$, Partial $\eta^2 = .18$), the 20% impostor condition ($F(3, 31) = 9.13$, $p = .001$, Partial $\eta^2 = .23$), and the 30% impostor condition ($F(3, 31) = 22.43$, $p = .001$, Partial $\eta^2 = .43$).

However, when different impostor types were considered across the three impostor conditions, one-way ANOVA revealed no statistically significant differences between the rates obtained for each impostor type. Specific scores for the individual impostor type categories were: TP choice ($F(2, 87) = 0.89$, $p = .41$); Panel choice ($F(2, 87) = 0.78$, $p = .46$); Algorithm selection ($F(2, 87) = 1.06$, $p = .35$); and Random selection ($F(2, 87) = 2.64$, $p = .08$).

Given that no differences in performance were found between the impostor frequency conditions, both for overall performance and the impostor types (shown in Table 7), further analyses considering the difference between impostor types involved combining scores across all impostor frequency conditions. Overall rates, with all impostor frequency conditions combined are presented in Table 8.

*Table 8: Experiment 1: Hit Rates by impostor type (frequency combined)*

| | ***Combined Frequency Condition*** |
|---|---|
| ***Impostor Type*** | *M (SD)* |
| **TP Choice** | .87 (.18) |
| **Panel Choice** | .86 (.19) |
| **Algorithm Selection** | .97 (.07) |
| **Random Selection** | .98 (.06) |

Table 8 shows the consistent disparity in performance rates between TP and Panel chosen impostors compared to Algorithm and Randomly selected impostors. These differences were explored by conducting a paired samples t-test which revealed significant differences between TP and Algorithm ($t(88) = -5.72$, $p < .001$), TP and Random ($t(89) = -5.45$, $p < .001$), Panel and Algorithm ($t(88) = -6.10$, $p < .001$), and Panel and Randomly selected ($t(89) = -6.34$, $p < .001$) impostors. Furthermore, there were no statistical differences between TP and Panel ($t(89) = .76$, $p = .45$), and Algorithm and Randomly selected ($t(88) = .38$, $p = .70$) impostor imagery.

### *4.3.3 Confidence Ratings and Decision Latency*

Providing that there was no difference in performance across the three impostor frequency conditions, confidence and decision latency were combined across all impostor frequency conditions (Table 9). This enabled adequate focus on differences between impostor types and the way that they impacted on confidence and decision latency.

*Table 9: Experiment 1: Confidence ratings overall and by impostor type*

| | ***Confidence Rating (%)***<br>*M (SD)* |
|---|---|
| **Overall** | 90.68 (9.23) |
| **TP Choice** | 84.86 (13.28) |
| **Panel Choice** | 81.75 (13.49) |
| **Algorithm Selection** | 91.57 (10.63) |
| **Random Selection** | 94.18 (8.48) |

Table 9 shows that when Human Operators were presented with TP and Panel chosen impostors their confidence was generally lower compared to when presented with Algorithm and Randomly selected impostors, perhaps indicating they that they had more difficulty with these types of impostors. These results seem to follow a very similar pattern to face matching performance results (Section 4.3.2). However, when explored further by conducting paired samples t-tests, it was found that there were statistically significant differences between all impostor conditions. Specific results included TP and Panel (*t*(89) = 4.27, *p* < .001); TP and Algorithm (*t*(88) = -7.68, *p* < .001); TP and Random (*t*(89) = -8.76, *p* < .001); Panel and Algorithm (*t*(88) = -9.50, *p* < .001); Panel and Random (*t*(89) = -12.78, *p* < .001); and Algorithm and Random (*t*(88) = -2.86, *p* = .005).

Table 10 shows overall decision latency and decision latency by impostor type.

*Table 10: Experiment 1: Decision latency overall and by impostor type*

| | ***Decision Latency (s)***<br>*M (SD)* |
|---|---|
| **Overall** | 3.51 (0.54) |
| **TP Choice** | 4.14 (1.69) |
| **Panel Choice** | 4.03 (1.42) |
| **Algorithm Selection** | 3.79 (1.86) |
| **Random Selection** | 3.48 (1.23) |

It can be seen that Human Operators were making their face matching decisions in approximately 4 seconds. Paired samples t-tests were conducted to assess if there were any differences in the time Human Operators took to consider different impostor types. Statistically significant differences were found for TP and Random (*t*(88) = 3.71, *p* < .001), and Panel and Randomly (*t*(89) = 3.73, *p* < .001) selected stimuli. However, no differences were found for TP and Panel ($t$(89) = 0.68, $p$ = .50), TP and Algorithm ($t$(88) = 1.91, $p$ = .06), Panel and Algorithm ($t$(88) = 1.10, $p$ = .28), and Algorithm and Randomly ($t$(88) = 1.50, $p$ = .14) selected impostor stimuli.

In considering confidence and decision latency, there is a consistent pattern in terms of differences between TP and Panel chosen, and Algorithm and Randomly selected impostors, which is in line with performance rates obtained for these conditions (in-depth discussion in Section 7.1.1.2).

## *4.4 Discussion*

Experiment 1 was a scoping study which assessed the impact of impostor frequency and type on Human Operator one-to-one face matching performance. The main discussion of this study is contained within Chapter 7, in which the findings from Experiments 2 and 3 are also considered. However, before moving onto the next two experiments, the results relating specifically to the impact of impostor frequency on face matching performance are briefly considered. These results input directly into the methodologies of Experiments 2 and 3.

During this experiment, Human Operators were randomly assigned to either 0, 10, 20, or 30% impostor condition with the aim to assess if their one-to-one face matching performance would differ for the different conditions. It was found that face matching performance for the 0, 10, 20, and 30% impostor conditions was predominantly not affected. The current finding is consistent with Bindemann, et al., (2010) who compared impostor detection performance for impostor rates of 2 and 50% (Section 2.2.4).

Bindemann, et al., (2010) also reported that in the 2% condition, participants had a higher tendency to incorrectly reject true presentations compared to the 50% condition. This was also found in Experiment 1. In the 0% condition, Human Operators incorrectly rejected true presenters significantly more frequently than in the 10, 20 and 30% conditions. This may be attributed to Human Operators' preconceived ideas that experiments must have some incidence of impostor stimuli and therefore they set their criterions accordingly. Bindemann, et al., (2010) eliminated this effect in subsequent experiments by first exposing participants to 50% and then to the 2% impostor rates. They suggested that participants probably applied dissociable criteria for detecting impostors and matching true presenters, however they require an initial exposure to both to attain and stabilise that criteria (Bindemann, et al., 2010). This finding raises questions about if this also occurs in applied setting. If that is the case it may mean that unlike in experiments where it is believed that occurrence of impostors is high, in applied settings where low impostor prevalence is expected, Human Operator criterion will be different.

Finally, the current findings relating to the impact of impostor frequency on face matching performance may serve to confirm that previous research that predominantly utilised 50% impostor rates may not undermine the difficulty of an impostor detection task. Additionally, given that there were no statistical performance differences between different frequency conditions, it was decided that the highest – 30% – rate be used during Experiments 2 and 3. A higher rate was chosen to ensure that sufficient data was obtained for both true and impostor presentations to enable meaningful analyses.

The next chapter focuses on Experiment 2 which is a simulated *live* access control performance evaluation, followed by a replication of the *live* evaluation in a laboratory setting in Experiment 3.

# *Chapter 5*

## *Experiment 2: Live One-to-One Face Matching*

Experiment 2, reported in this chapter, is the central experiment of this body of work. The aim of Experiment 2 was to recreate an applied access control setting and assess Human Operator one-to-one face matching performance within this setting. This aim stems from the need to formally consider and assess the extent to which findings from controlled laboratory experiments reflect real world situations. Consequently, the work conducted as part of Experiment 2 is later replicated in Experiment 3 in the form of a controlled laboratory study.

As discussed previously (Chapter 1), the conduct of simulated *live* evaluations is associated with many logistical complexities. This is predominantly attributed to the many applied factors which can affect face matching performance that need to be considered when preparing and conducting such evaluations. Perhaps for that reason, only a handful of *live* evaluations has been conducted and reported in the literature (Section 2.2.5). It is interesting to note that *live* evaluations which assessed one-to-one face matching performance, reported significantly different findings. Overall accuracy reported by Kemp, et al., (1997) was 67%, Butavicius, et al., (2008) 95% and, Megreya and Burton (2008) reported 83% overall accuracy. This may be attributed to a number of methodological differences (e.g., the type and the way that stimuli were presented, differences in distracter tasks, etc.). However, only Megreya and Burton (2008) compared performance on the same stimuli presented *live* and in a laboratory setting. They found that participants were more likely to state that stimuli pairs were a match (i.e., they had a more lenient criterion) in the *live* compared to in the laboratory condition (further discussed in Section 7.1.2).

This work further extended the previous *live* studies by also considering a number of factors that can affect face matching performance within applied settings. As already mentioned, considered were the impact of impostors, expertise, and individual differences. What follows is a brief overview of previous empirical work which has considered these factors.

In terms of the impact of impostors, the findings of Experiment 1 influenced the methodologies of Experiments 2 and 3. As stated in the discussion of

Experiment 1 (Section 4.4), the impact of impostor frequency is not considered further. Instead, the highest impostor rate of 30% was used.[17] Also, the same four impostor types (i.e., TP Choice, Panel Choice, Algorithm Selection, and Random Selection) assessed in Experiment 1 were considered in Experiments 2 and 3.

Human Operator expertise within the area of face matching has not received extensive empirical attention. As outlined in Section 2.2.3.3, of the research that has been conducted, findings have been inconsistent. This provides a motivation to compare one-to-one face matching performance of trained/experienced Human Operators and lay individuals. Therefore, in addition to assessing *live* face matching performance, Experiment 2 incorporated this comparison.

Experiment 2 also considered the impact of individual differences (i.e., perceptual speed, personality, etc.) on face matching performance. This focus was motivated by the consistent amount of performance variation evident throughout much empirical work (Lee, et al., 2006; Megreya & Burton, 2006, 2007, 2008). However, reasons why this may be the case remain largely unexplained. Of the research that considered individual differences, Schretlen, et al, (2001) found that performance on the Benton Facial Recognition Test correlated positively with perceptual speed and total cerebral volume. Megreya and Burton (2006) found

[17] The actual impostor rate in Experiment 2 was 26.4% as it depended on the number of TPs who were able to attend on the day of the experiment.

performance on a 1-to-10 face matching task modestly correlated with perceptual speed, visual short-term-memory, and figure matching. Also, Li et al., (2010) found that extraverts correctly recognised more faces compared to introverts. However, much more work is needed to appropriately establish a consistent set of measures that could reliably predict face matching performance. Consequently, Experiment 2 aimed to further extend this work by specifically focusing on a one-to-one face matching task. It is anticipated that findings from this work would be of applied relevance (e.g., by informing selection of individuals more suited for face matching positions). Also theoretically, knowledge about what predicts face matching performance could assist with better understanding the fundamental processes that underlie unfamiliar face matching in general.

## *5.1 Method*

### *5.1.1 Participants*

#### *5.1.1.1 Target Participants*

Participants who took part in the Imaging Trial (Section 3.2) were approached to take part in Experiment 2. As discussed previously, having the same TPs (and therefore, the same stimuli) enabled comparisons between the current *live* evaluation and the laboratory evaluations (i.e., Experiments 1 and 3). Of the 316 TPs who participated in the Imaging Trial, 129 (33 females and 96 males) also volunteered to take part in Experiment 2. Their ages ranged from 21 to 64 ($M = 42.06$, $SD = 10.46$).

### *5.1.1.2 Human Operators*

Thirty two untrained/inexperienced and trained/experienced Human Operators (19 females and 13 males) participated in this experiment (Figure 13). Their ages ranged from 17 to 56 ($M = 34.32$, $SD = 10.21$).

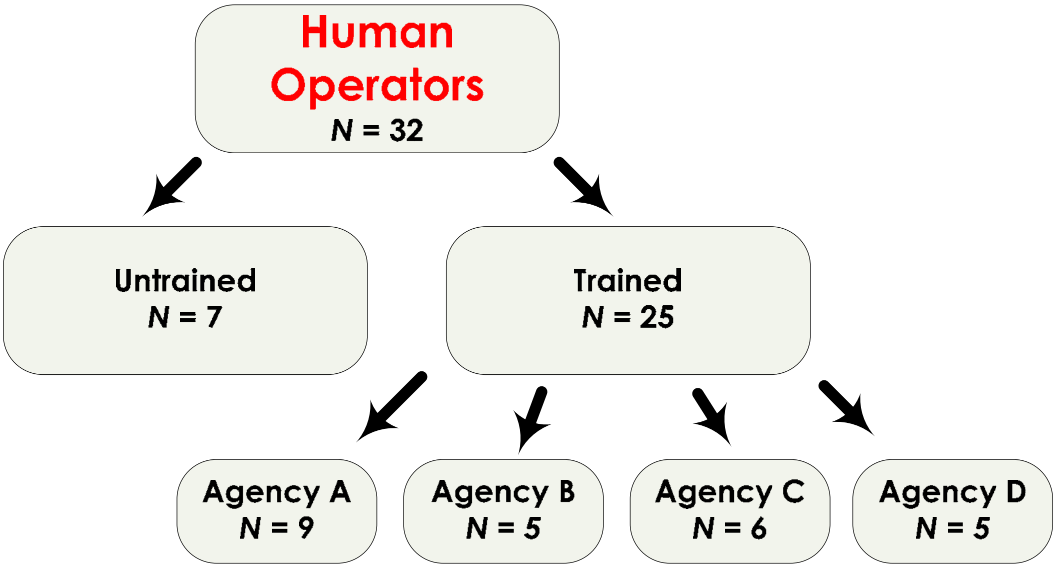


*Figure 13: Experiment 2: Untrained and trained/experienced Human Operators*

***Untrained/inexperienced Human Operators*** (referred to as untrained Human Operators) had no face matching expertise. The seven (5 females and 2 males) untrained Human Operators were recruited from The University of Adelaide, School of Psychology paid participants database. They were reimbursed for their participation. Their ages ranged from 17 to 42 ($M$ = 28.14, $SD$ = 9.29).

***Trained/experienced Human Operators*** (referred to as trained Human Operators) had face matching expertise and conducted such tasks as part of their employment. The twenty-five (14 females and 11 males) trained Human Operators were sourced from four government organisations which require their staff to perform various face matching tasks as part of their employment. Their ages ranged from 22 to 56 ($M$ = 36.13, $SD$ = 9.92). These individuals had on-the-job experience in conducting face matching tasks and/or had undergone some form of training. More specifically, four individuals had only on-the-job training, and 21 had received formalised face matching training. The formalised face matching training was conducted internally by each agency and usually ranged from a few hours to two days in duration. Years on the job varied from eight months to 24 years with a mean of 5.44 ($SD$ = 6.45) years across all trained Human Operators. The identity of the government organisations involved in this research is kept confidential for security reasons. Therefore, organisations are referred to as Agency A, B, C and D.

### 5.1.2 *Materials*

#### *5.1.2.1 ID Cards*

Figure 14 shows an example of the ID Cards which were prepared specifically for this experiment to be used by TPs. They served as a form of ID that was presented to each Human Operator. Each ID card featured a TP's true or impostor photograph which was positioned in the middle of the card with their unique ID number immediately under the photograph. The photograph was a colour passport style photograph consistent with the DFAT passport photographic guidelines and was prepared as part of the image preparation process (Chapter 3).

NOTE:
This figure/table/image has been removed
to comply with copyright regulations.
It is included in the print copy of the thesis
held by the University of Adelaide Library.

**TP4794**

*Figure 14: Experiment 2: An example of Target Participant ID card*

No additional information (e.g., demographics) was included on the ID card. To focus on only face matching performance evaluation it was decided that additional information, that could potentially affect Human Operator face matching decisions would not be included. Additionally, due to the involvement of different organisations with diverse applied settings it was not plausible to focus on only one application as part of this experiment. For that reason it was decided that the look of the ID card be simple and generic and not specifically attempting to simulate any one of the participating agencies applications or applied settings.

#### *5.1.2.2 Human Operator Record Booklet*

Human Operator Booklets (Appendix F) in were used by Human Operators. The first few pages of the booklet provided detailed instructions about the Human Operator role in the experiment. It also contained basic demographic questions (e.g., age, gender, training, etc.). The remainder of the booklet was dedicated to Human Operator face matching decision recording. Each page was allocated to one decision on which Human Operators were provided with space to record their decision and confidence rating relating to that decision. The last few pages of the booklet contained post-evaluation questions which asked Human Operators to describe how they perform face matching tasks in general (e.g., what aspects of the face they consider important).

#### *5.1.2.3 Individual Differences Tests*

Based on previous face matching work which considered individual differences (Section 2.2.3.2), the following tests were selected:

1. ***Perceptual Speed Test*** was taken from the *Kit of Factor-Referenced Cognitive Tests* (Ekstrom, French, Harman, & Dermen, 1976). The authors define perceptual speed as "the speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other very simple tasks involving visual perception" (Ekstrom, et al., 1976, p. 123). This test has three subscales:

    - ***Finding A's Test.*** Participants were presented with columns of words (five per page) and asked to locate as many *A*s within the presented words as possible, within 2 minutes.
    - ***Number Comparison Test.*** Participants were presented with pairs of multi-digit numbers and asked to classify these pairs as same or different. They were instructed to do this for as many pairs as possible within a 90 second time period.
    - ***Identical Pictures Test.*** Participants were presented with a target line-drawn figure which they matched to an array of five variants. They were required to match as many figures as possible within a 90 second time period.

2. ***Rational Experiential Inventory (REI).*** Participants were presented with a series of 40 statements and asked to, on a 5 point Likert scale, rate how applicable those statements were for them. The REI measures rational and experiential thinking or information processing styles, and includes self-

reported ability and engagement (Pacini & Epstein, 1999).

3. ***Glasgow Face Matching Test*** **(GFMT) (short version).** Participants were presented with a series of 40 grayscale face pairs (i.e., one pair per an A4 sheet) and asked to select whether the presented faces were of the same or different person (Burton, White, & McNeill, 2010). The GFMT measures human one-to-one face matching ability.

### *5.1.3 Design and Procedure*

This experiment was conducted as a single day activity. A simulated access control environment was assembled in a vacant building (shown in Figure 15).

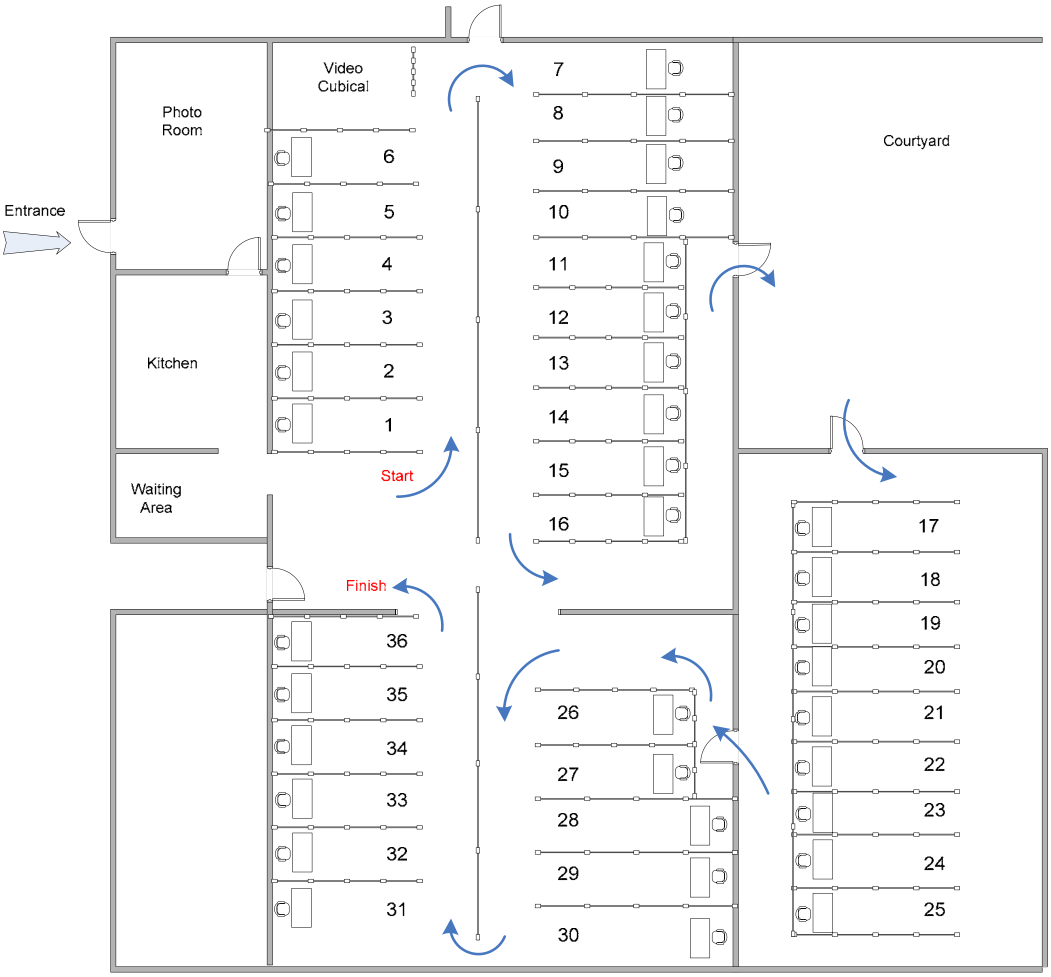


*Figure 15: Experiment 2: Testing environment showing cubicles in which Human Operators were seated and the arrows representing TP movement*

Cubicles containing a chair and a table were prepared for Human Operators. The cubicles were divided by light sheets of material which were hung from the ceiling. The purpose of having each Human Operator cubicle separated was to minimise any potential face familiarisation effects that may have impacted on face matching performance. This ensured that Human Operators were not able to see TPs prior to them entering their cubicle to commence their interaction.

#### *5.1.3.1 Target Participant Task*

After reading the Information Sheet (Appendix G), Guidelines for Volunteers Appendix B) and providing consent (Appendix C), TPs were given an envelope which contained an ID card. The ID card contained either their true image or an impostor image. As shown in Figure 16, out of 129 TPs, 95 were given ID cards which contained their true image and 34 were given ID cards which contained an impostor image. This meant that the evaluated impostor rate was 26.4%. Based on findings from Experiment 1 it was decided that 30% impostor rate would be used for all subsequent performance evaluations. However, since this was a simulated *live* experiment which was conducted over one day, it was not possible to control the number of participants who were available for participation on the day.

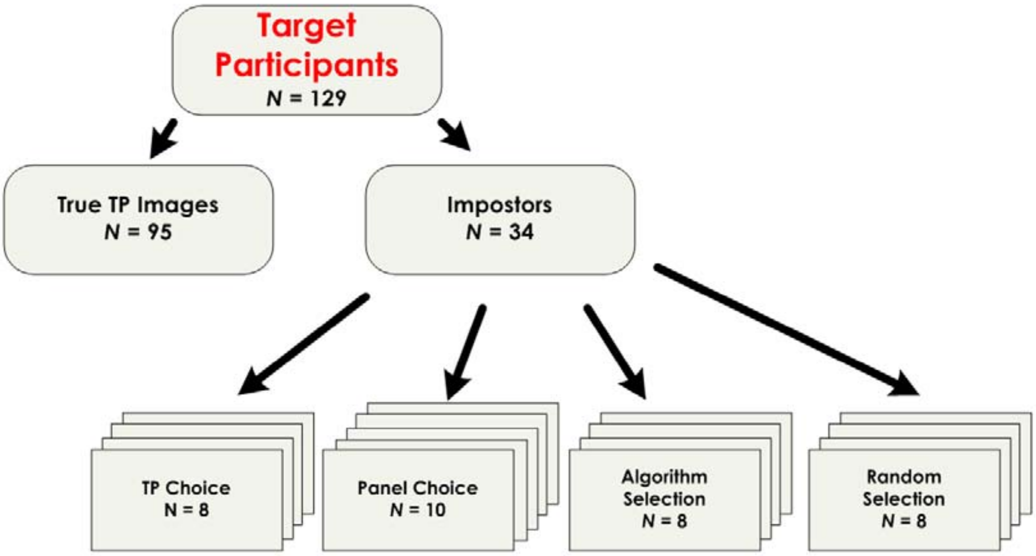


*Figure 16: Experiment 2: Target Participant imagery presented to Human Operators*

TPs were instructed not to open the envelope and inspect their ID card. They were also not informed about whether their ID card contained a true or impostor image. This was important as there was a possibility that TPs' behaviour could have been impacted if they knew that the image on their ID card was or was not their true image. The TPs task was to present their envelope to all Human Operators who took the ID card out of the envelope (without showing the card to TPs), inspected the photograph and made their face matching decision. As presented in Figure 15, TPs "looped" through the testing environment to ensure that they interacted with all Human Operators. All TPs were instructed on where they would start their interactions and how to navigate through the testing environment.

In addition to interacting with Human Operators, TPs were asked to participate in video imagery acquisition. This was done by setting up one Human Operator cubicle with video imaging equipment (the location of this cubicle can be seen in Figure 15). This imagery was acquired for subsequent use in Experiment 3 (Chapter 6).

The video imaging set up and procedure were based on the Imaging Trial methodology (Section 3.2.2.1.2). The same imaging equipment and lights were used for this set up. However, it should be noted that the aim here was not to replicate the "ideal" Imaging Trial environment but to image TPs under the same conditions in which they were seen by Human Operators. Therefore, a number of differences in the set up are noted. One such difference was associated with TPs' walking distance. During the Imaging Trial the walking distance was based on an operational estimation. However, the *live* Human Operator imaging set up did not allow for such a distance. This resulted in a shorter walking distance and therefore, a shorter video.[18] The size of the imaging location also dictated the number of lights that were required for the imaging set up. Instead of using six, this set up required four Masterlite1500 lights. It should also be noted that unlike during the Imaging Trial, there was a small amount of natural light in the location, but the Masterlites were configured so that they were the primary illumination source (Figure 17).

[18] Imaging Trial videos averaged 7 to 8 seconds, whereas videos acquired as part of this acquisition averaged 2 to 4 seconds in duration.

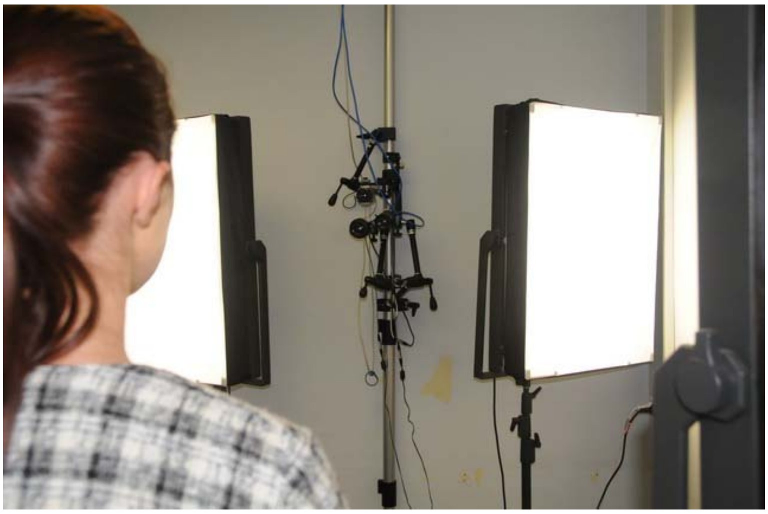


*Figure 17: Experiment 2: An example of Target Participant video image acquisition*

### *5.1.3.2 Human Operator Task*

Human Operators were briefed about their role in the experiment prior to their arrival to the testing location. They were emailed the Information Sheet

(Appendix H) and Guidelines for Volunteers (Appendix B) as well as verbally briefed on the bus on the way to the testing location. Upon arrival they provided their consent (Appendix C) and were given the Human Operator booklet. Each booklet was associated with a unique Human Operator number (e.g., HO0015 meaning Human Operator 15). Human Operators were then asked to provide basic demographic details in their booklets. This was followed by individual differences tests (Section 5.1.2.3) which all Human Operators completed at the same time prior to commencing the face matching part of the experiment. Once these tests were completed each Human Operator was asked to occupy a cubicle which corresponded to the Human Operator number on their booklet, and asked to remain in their cubicles for the duration of the experiment.

As shown in Figure 18, during the experiment Human Operators were approached by TPs who presented them with envelopes which contained their ID cards. Human Operators were instructed to assess the photograph on the ID card, compare that photograph to the presenting TP, and make a decision about whether the face in the photograph and the presenting individual were a match or not. Once a decision was made, Human Operators were further asked to indicate how confident they were in their decision. This was done on a ten point 0-100 Likert type scale. Human Operators then returned the ID card into the envelope and returned the envelope to the TP.

*Figure 18: Experiment 2: An example of Target Participant and Human Operator interaction*

Human Operators were asked to make their decisions as quickly and as accurately as possible. To minimise other effects that could assist Human Operators with making their face matching decisions, they were instructed not to talk to or ask TPs any questions, especially questions relating to their appearance.

Once all interactions with TPs were completed, all Human Operators completed a set of post-experiment questions relating to how they performed the face matching task and if there were facial characteristics that they focused on in particular. Additionally, a semi structured focus group discussion with all Human Operators was conducted after the experiment asking them some general questions about how they found the task and what they thought the impostor rate was.

#### *5.1.3.3 Logistics: Target Participant and Human Operator Interaction*

The experimental conduct was divided into three, one and a half hour long sessions during which Human Operators and TPs interacted. This allowed for approximately 30 to 40 TPs to participate in each session. At the commencement of each session, Human Operators were seated in their cubicles where they remained for the duration of the session. Upon their arrival, TPs were first briefed about their role in the experiment and provided with envelopes which contained their ID cards. They then entered the testing environment where Human Operators were seated. Each TP was initially positioned in front of a cubicle with their back facing the Human Operator. This was done to ensure similar interaction time between Human Operators and TPs. Once a TP completed their interaction with one Human Operator they were instructed to move onto the next cubicle. The progression of TPs was monitored so that they did not enter a cubicle until the previous TP had exited. This was organised in such a way so that all TPs "looped" through the rooms to ensure that they interacted with each Human Operator once.

## *5.2 Results*

### *5.2.1 Overall Performance*

Overall Human Operator performance accuracy ranged from 86 to 100% with an average of 93.77% (*SD* = 3.9). Overall accuracy is further explored by looking at performance rates, presented in Table 11. It can be seen that Human Operators did not detect 15% of impostors and that 3% of true presenters were rejected.

*Table 11: Experiment 2: Overall performance rates*

| | ***Overall Rates***<br>*M (SD)* |
|---|---|
| **Hit Rate** | .85 (.12) |
| **Miss Rate** | .15 (.12) |
| **False Alarm Rate** | .03 (.06) |
| **Correct Response Rate** | .97 (.03) |

### *5.2.2 The Impact of Human Operator Expertise*[19]

Human Operator performance was further explored by considering trained and untrained Human Operators separately, presented in Table 12. Trained Human Operators scores ranged from .86 to 1.00 and untrained Human Operators scores ranged from .87 to .96. An independent samples t-test revealed that this difference was statistically significant, $t(30) = -2.54$, $p = .03$. Furthermore, an independent samples t-test revealed that *d'* for values trained and untrained Human Operators were also statistically different, $t(30) = -2.80$, $p = .01$.

[19] Preliminary results of this work were presented at the Australasian Experimental Psychology Conferences (Calic, Macleod, McLindin, & Dunn, 2010).

*Table 12: Experiment 2: Overall accuracy, performance rates and SDT indices for trained and untrained Human Operators*

| | ***Trained Human Operators***<br>*M (SD)* | ***Untrained Human Operators***<br>*M (SD)* |
|---|---|---|
| **Overall Accuracy** | .95 (.04) | .91 (.04) |
| **Hit Rate** | .86 (.13) | .80 (.10) |
| **Miss Rate** | .14 (.13) | .20 (.10) |
| **False Alarm Rate** | .02 (.03) | .05 (.04) |
| **Correct Response Rate** | .98 (.03) | .95 (.04) |
| ***Mean d'*** | 3.41 (.66) | 2.63 (.59) |
| ***Mean β*** | 1.4 (1.3) | 1.11 (.98) |

However, when specific performance rates were considered, no statistical differences between trained and untrained Human Operators were found for Hit Rates (*t*(30) = -1.49 *p* = .16); Miss Rates (*t*(30) = -0.64, *p* = .16); False Alarm Rates (*t*(30) = -0.64, *p* = .53); and Correct Response Rates (*t*(30) = -1.99, *p* = .08).

Further to performance rates and SDT indices, Figure 19 graphically illustrates the difference between trained and untrained Human Operators performance.

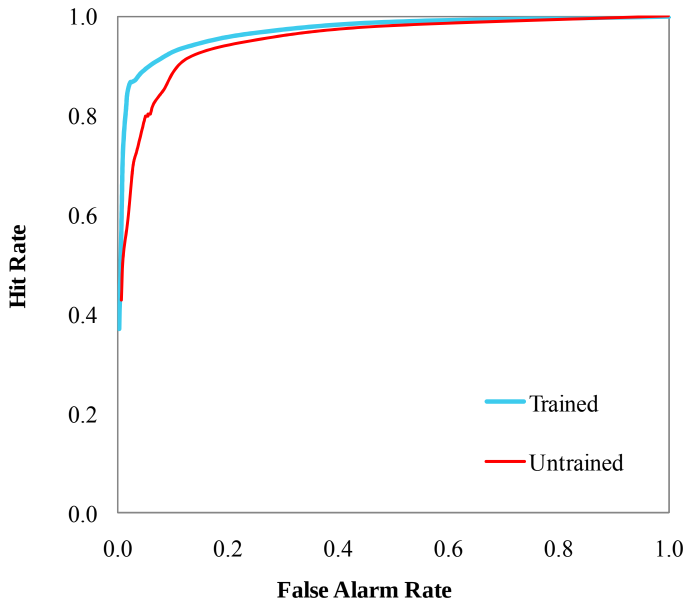


*Figure 19: Experiment 2: ROCs for trained and untrained Human Operator performance*

### *5.2.2.1 Trained Human Operator Performance by Agency*

To better understand the impact of Human Operator expertise, performance of trained Human Operators was also considered based on the agency. This was done because face matching tasks and the exact nature of the tasks differed between the agencies. It therefore made sense to assume that as a result of different experiences Human Operators may perform differently (Table 13).

*Table 13: Experiment 2: Overall accuracies and performance rates by agency*

| | ***Agency*** | | | |
|---|---|---|---|---|
| | ***A*** | ***B*** | ***C*** | ***D*** |
| | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* |
| **Overall Accuracy** | .96 (.02) | .95 (.04) | .91 (.04) | .97 (.001) |
| **Hit Rate** | .94 (.09) | .91 (.10) | .71 (.12) | .90 (.03) |
| **False Alarm Rate** | .02 (.02) | .04 (.04) | .02 (.03) | .01 (.01) |

Overall accuracies reveal that Human Operators from Agency C performed lower compared to the other three agencies. Statistical analyses partly confirmed this finding by showing that Agency C's overall accuracy was statistically lower than overall accuracies of Agency A (*t*(5) = 2.65, *p* < .05), and D (*t*(4) = -6.71, *p* < .01), but not of Agency B (*t*(4) = 2.04, *p* = .11). Furthermore, overall accuracies of Agencies A, B, D were not statistically different. Specific scores were: Agency A and B (*t*(4) = .31, *p* = .77); Agency A and D (*t*(4) = -1.07, *p* = .35); and, Agency B and D (*t*(4) = -1.30, *p* = .27).

The lower performance of Agency C is further confirmed when Hit Rates are considered. Agency C's Hit Rate was statistically lower than Hit Rates of Agency A (*t*(5) = 3.95, *p* = .01), B (*t*(4) = 4.70, *p* = .01), and D (*t*(4) = -14.43, *p* < .001). Furthermore, Hit Rates of Agencies A, B, D were not statistically different:

Agency A and B, ($t$(4) = 1.00, $p$ = .37); Agency A and D, ($t$(4) = 0.21, $p$ = .85); and, Agency B and D, ($t$(4) = 0.10, $p$ = .92). Also, no differences were found for agencies' False Alarm Rates: Agency A and B ($t$(4) = -2.51, $p$ = .07); Agency A and C ($t$(5) = -0.21, $p$ = .84); Agency A and D ($t$(4) = 0.30, $p$ = .78); Agency B and C ($t$(4) = 0.52, $p$ = .57); Agency B and D ($t$(4) = 1.76, $p$ = .15); and Agency C and D ($t$(4) = 0.88, $p$ = .43).

### *5.2.3 The Impact of Impostor Types*

Table 14 shows trained and untrained Human Operator Hit Rates by impostor type. Analysis of variance revealed a statistically significant main effect for impostor type ($F$(2.12, 63.50) = 16.89, $p$ < 001, Partial $\eta^2$ = .36) but not for impostor type and training interaction ($F$(2.12, 63.50) = 0.64, $p$= .54, Partial $\eta^2$ = .02).

*Table 14: Experiment 2: Human Operator Hit Rates by impostor type*

| | ***Trained Human Operators*** | ***Untrained Human Operators*** |
|---|---|---|
| ***Impostor Type*** | *M (SD)* | *M (SD)* |
| **TP Choice** | .76 (.23) | .63 (.22) |
| **Panel Choice** | .76 (.26) | .74 (.21) |
| **Algorithm Selection** | .95 (.10) | .91 (.06) |
| **Random Selection** | .99 (.03) | .95 (.07) |

A repeated measures ANOVA revealed a significant effect of the combined impact of impostor type on trained (*F*(3, 25) = 15.46, *p* < .001, Partial $\eta^2$ = .39), and untrained Human Operators (*F*(3, 7) = 7.20, *p* = .01, Partial $\eta^2$ = .55). These differences were explored further by separately considering different impostor types' scores for trained and untrained Human Operators.

Paired samples t-tests revealed that for ***trained*** Human Operators statistically significant differences were found for TP and Algorithm (*t*(24) = -3.95, *p* = .001), TP and Random (*t*(24) = -5.26, *p* < .001), Panel and Algorithm (*t*(24) = -3.90, *p* = .001), Panel and Random (*t*(24) = -4.48, *p* < .001), and Algorithm and Randomly (*t*(24) = 2.22, *p* = .040) selected impostors. TP and Panel chosen impostors (*t*(24) = 0.09, *p* = .93) were not found to be significantly different.

For ***untrained*** Human Operators, however, a completely different pattern of results was found. Two statistically significant differences were found between TP and Algorithm ($t(6) = -3.77$, $p = .01$), and TP and Randomly ($t(6) = -3.75$, $p = .01$) selected impostor types. No statistically significant differences were found between TP and Panel ($t(6) = 1.32$, $p = .23$); Panel and Algorithm ($t(6) = -2.09$, $p = .08$); Panel and Random ($t(6) = -2.09$, $p = .08$); and Algorithm and Randomly ($t(6) = 1.55$, $p = .17$) selected impostors.

### *5.2.4 Individual Differences*

This section examines the impact of individual differences on face matching performance. Table 15 presents overall results which combine trained and untrained Human Operator performance. The table shows Pearson's correlations between Human Operator performance and the individual differences tests.

*Table 15: Experiment 2: Pearson's correlations between Human Operator performance and individual differences tests*

| | ***Overall Accuracy*** | ***Hit Rate*** | ***False Alarm Rate*** |
|---|---|---|---|
| **Training/Experience** | **.43*** | .23 | .08 |
| **Confidence Ratings** | **.37*** | .17 | .30 |
| **Perceptual Speed** | | | |
| Finding A's | .29 | .12 | **.36*** |
| Identical Pictures | .34 | .23 | -.05 |
| Number Comparison | .33 | **.41*** | .08 |
| **Glasgow Face Matching Test** | **.48**** | .22 | .35 |
| **Rational-Experiential Inventory** | | | |
| Rational Ability | **-.36*** | **-.45**** | -.04 |
| Rational Engagement | **-**.29 | **-.42*** | .09 |
| Experiential Ability | -.08 | -.17 | -.19 |
| Experiential Engagement | -.19 | -.11 | -.24 |

* correlation is significant at *p* < .05
** correlation is significant at *p* < .01

Table 15 shows that overall accuracy was moderately[20] correlated with Human Operator expertise and confidence ratings. In terms of the Perceptual Speed tests, the Finding A's test correlated with False Alarm Rate indicating that better performance on this test was associated with incorrectly rejecting true presenters. No reliable associations were found with the Identical Pictures test. The Number Comparison test however, indicated a moderate correlation with Hit Rates, indicating that correct rejection of impostors was associated with good performance on this test. In terms of the face matching task, moderate correlations were found for the performance on the GFMT and the *live* one-to-one face matching task. As both tasks are one-to-one face matching tasks, this may not be surprising and will be discussed further in the overarching discussion (Chapter 7). Finally, in terms of thinking styles, negative correlations were found between Rational Ability/Engagement aspects of the scale while Experiential Ability/Engagement did not show any reliable associations with performance. These results are discussed in the overarching discussion (Chapter 7).

[20] The correlation coefficient is a measure of the strength of the relationship between two variables. Its values range between +1 and -1. Coefficient values can be interpreted in the following way:

- 0 denotes no relationship.
- +1 denotes a perfect positive relationship: as one variable's value increases, the other variable's value also increases.
- -1 denotes a perfect negative relationship: as one variable's value increases, the other variable's value decreases.
- Values between 0 and 0.3 (0 and -0.3) denote a weak relationship.
- Values between 0.3 and 0.7 (-0.3 and -0.7) denote a moderate relationship.
- Values between 0.7 and 1.0 (-0.7 and -1.0) denote a strong relationship (Dancey & Reidy, 2002).

The results presented in Table 15 above, combine trained and untrained Human Operators' performance. Consequently, further correlation analyses were conducted to consider any differences between trained and untrained Human Operators. Full results of these analyses are provided in Appendix I while Table 16 incorporates only columns where significant results were found.

*Table 16: Experiment 2: Pearson's correlations between trained and untrained Human Operator performance and individual differences tests*

| | ***Trained Human Operators*** | | ***Untrained Human Operators*** |
|---|---|---|---|
| | *Overall Accuracy* | *Hit Rate* | *False Alarm Rate* |
| **Confidence Ratings** | **.44*** | .19 | **.83*** |
| **Perceptual Speed** | | | |
| Identical Pictures | **.43*** | .22 | .31 |
| Number Comparison | .27 | **.42*** | .47 |
| **Glasgow Face Matching Test** | **.57**** | .29 | .49 |
| **Rational-Experiential Inventory** | | | |
| Rational Ability | **-.49*** | **-.45*** | .15 |
| Rational Engagement | -.29 | **-.41*** | .60 |

* correlation is significant at $p < .05$
** correlation is significant at $p < .01$

Examining the findings in Table 16, it can be observed that the majority of overall accuracy correlations (Table 15) were contributed by trained Human Operators' performance. The only exception is the confidence ratings, correlating highly with untrained Human Operators' False Alarm Rate. The only other surprising finding was the correlation between scores on Identical Pictures and overall accuracy for trained Operators. This finding is surprising considering that for overall performance no associations with performance on Identical Pictures were found.

Finally, independent samples t-test revealed no statistically significant differences between trained and untrained Human Operator scores on the individual differences tests. Specific results were: Findings As (*t*(30) = -0.33, *p* = .74); Identical Pictures (*t*(30) = 0.66, *p* = .55); Number Comparison (*t*(30) = -1.07, *p* = .29); GFMT (*t*(30) = -1.16, *p* = .26); Rational Ability (*t*(30) = 0.47, *p* = .64); Rational Engagement (*t*(30) = 1.75, *p* = .09); Experiential Ability (*t*(30) = -0.97, *p* = .34); Experiential Engagement (*t*(30) = 0.90, *p* = .38).

### *5.2.5 Confidence Ratings*

Table 17 shows confidence ratings for trained and untrained Human Operators. Overall confidence scores show that trained Operators were slightly less confident compared to untrained Human Operators. However, this difference was not statistically significant, *t*(30) = 0.38, *p* = .71. In considering confidence by impostor type, the only significant difference between trained and untrained Operators was found for Randomly selected impostors, *t*(30) = 2.08, *p* = .05.

Confidence ratings for TP choice (*t*(30) = -0.05, *p* = .96), Panel choice (*t*(30) = 1.04, *p* = .32) and Algorithm selection (*t*(30) = 1.29, *p* = .21) impostor types were not found to be statistically different.

*Table 17: Experiment 2: Confidence ratings overall and by impostor type*

| | ***Confidence Rating (%)*** *M (SD)* | |
|---|---|---|
| | ***Trained Human Operators*** | ***Untrained Human Operators*** |
| **Overall** | 87.88 (8.91) | 89.00 (6.19) |
| **TP Choice** | 74.65 (13.70) | 74.46 (6.53) |
| **Panel Choice** | 75.64 (10.75) | 80.29 (10.36) |
| **Algorithm Selection** | 87.45 (10.01) | 91.25 (5.73) |
| **Random Selection** | 90.65 (13.93) | 97.14 (3.73) |

Examining between group variance for each impostor type (i.e., looking at the rates horizontally), a repeated measures ANOVA revealed statistically significant differences for confidence ratings obtained by trained (*F*(3, 25) = 30.66, *p* < .001, Partial $\eta^2$ = .56) and untrained (*F*(3, 7) = 25.32, *p* < .001, Partial $\eta^2$ = .81) Human Operators. Paired samples t-tests were performed to separately explore trained and untrained Human Operators' confidence ratings.

For ***trained*** Human Operators' confidence ratings, statistically significant differences were found for TP and Algorithm (*t*(24) = -6.45, *p* < .001), TP and Random (*t*(24) = -6.53, *p* < .001), Panel and Algorithm (*t*(24) = -6.45, *p* < .001) and, Panel and Randomly (*t*(24) = - 6.27, *p* < .001) selected impostors. However, confidence ratings for TP and Panel (*t*(24) = .53, *p* = .60) and, Algorithm and Randomly (*t*(24) = 1.76, *p* = .09) selected impostors were not significantly different.

A similar pattern was found for ***untrained*** Human Operators. Confidence ratings for TP and Algorithm (*t*(6) = -7.76, *p* < .001), TP and Random (*t*(6) = -8.79, *p* < .001), Panel and Algorithm (*t*(6) = -3.04, *p* = .02), Panel and Random (*t*(6) = -4.32, *p* = .01) and, Algorithm and Randomly (*t*(6) = 3.47, *p* = .01) selected impostors were statistically significant. Confidence ratings for TP and Panel (*t*(6) = 2.11, *p* = .08) chosen impostors were not found to be statistically different. These results will be discussed in light of Experiments 1 and 3 findings in the overarching discussion in Chapter 7.

### *5.2.6 The Post-Evaluation Survey*

The post-evaluation survey included a series of questions relating to the face matching task, its difficulty, and the way that Human Operators performed the task.

In terms of how the participants found the task, the most common response provided by trained and untrained Human Operators was that the task was: “relatively easy” “Overall, quite easy”, “very easy”. Nevertheless, in relation to ***untrained*** Human Operators (*n* = 7) it should be noted that two reported the task being difficult and one stated that it was at the “correct level of difficulty”. Sometimes Human Operators further justified their answers by also stating that the majority of the decisions were straightforward with only some that were more challenging. Examples of this include the following explanations:

> “The tasks were complex in some but majority were easy to determine.” (Human Operator 2, trained)
>
> “90% of the subjects posed no problem and were easier to ID than in real life.” (Human Operator 8, trained)
>
> “Fairly easy, kept thinking they were trying to make it tricky so would hesitate on the percentage of certainty.” (Human Operator 11, untrained)
>
> “I found the majority of the faces to be recognisable. There were a small number of faces which proved difficult to distinguish from the photos.” (Human Operator 15, trained)

When asked about how they performed the face matching task and if there were any particular features that they focused on, Human Operators reported different methods and aspects of face which they predominantly used to assist their decision making. However, regardless of the method that was employed, both trained and untrained Human Operators reported concentrating on individual facial features (e.g., eyes, nose, ears, eyebrows, jaw line, facial marks, mouth, facial creases, etc.) and to a lesser extent age, skin type, facial expression, etc. Examples of Human Operator descriptions of the process and the features which they used the most can be exemplified by considering the following responses:

> "Sectioned the face, concentrating on details such as eye width and shape, nose, philtrums and lips, jaw line and ears. Significant marks such as moles, scars etc." (Human Operator 1, trained)
>
> "I began with a quick overview of both the faces and the photos. In order to make my decision I dissected the different facial features to determine which were the same." (Human Operator 15, trained)
>
> "I looked at the eyebrows, nose, mouth and jaw line. Some faces had a very distinguishable nose so that is all I looked at. Others had distinguishable mouths or cheekbones" (Human Operator 16, untrained)
>
> "By focusing on 6 segments of the face: eyes, ears, nose, mouth, shape of face and facial marks (moles etc) and individually addressing and comparing each." (Human Operator 36, trained)

A notable difference in the responses provided by trained and untrained Human Operators is evident in the detail provided. Trained Human Operators provided much more details about how they conducted the task and which facial features they focused on. The above examples demonstrate this. Some additional examples are provided here for ***trained*** Human Operators:

> "Interacted with client. Sighted any facial features and marks, Look at ears (shape and size), look at picture, another look at face. Split in to 6 sections." (Human Operator 2, trained)
>
> "Look at the subjects as they approached and searched for distinguishing features. Did the same for the photo and matched distinguishing features." (Human Operator 23, trained)

Some examples of ***untrained*** Human Operators' responses can also be considered:

> "I looked at the face then looked at the picture and decided if the face matched the picture." (Human Operator 6, untrained)
>
> "I looked at the person, then at the picture then the person again and made a decision." (Human Operator 11, untrained)

In terms of any other comments made by Human Operators it should be noted that one Human Operator thought that they were being presented by the same TPs on multiple occasions. When asked what they thought could be done differently, one

Human Operator suggested ensuring that Human Operators and TPs are at eye level. This may be an important consideration as previous research has demonstrated that face angle can substantially impact human face matching abilities (Section 2.2.1.2). Finally, as stated by one Human Operator, it is important to be cognisant of the extent to which such experimental efforts are assessing and measuring actual performance:

> “This was more difficult than the "real thing" in that I felt conscious of the fact there were impostors rather than could be! I spent more time on each person and felt I kept double checking my first impression, I was doubting myself and my judgement far more than I would normally. This could adversely affect accuracy.” (Human Operator 3, trained)

## *5.3 Discussion*

Experiment 2 simulated an access control environment in an attempt to assess one-to-one face matching performance within an applied setting. One hundred and twenty-nine TPs presented IDs for inspection by 32 Human Operators who decided if the TP face and that presented in the photo on the ID were of the same individual. In addition to assessing one-to-one face matching performance, considered was the impact of four impostor types, Human Operator expertise, and individual differences on face matching performance. Provided here is a brief discussion of *only overall accuracy* which is directly relevant to Experiment 3. The main discussion is contained within Chapter 7 where all findings are discussed in the context of all experiments and the overarching research aims.

Overall Human Operator performance on the *live* one-to-one face matching task in Experiment 2 was high, with an average of just under 94% accuracy. Compared to previous *live* evaluations, the current finding is most similar to 95% overall accuracy reported by Butavicius, et al., (2008). This is followed by 83% reported by Megreya and Burton (2008), and finally Kemp, et al., (1997) who reported 67% overall accuracies. It may seem surprising to find such a notable performance disparity on this seemingly simple task. However, as discussed in Chapter 1, assessing applied face matching performance is associated with many complexities. For example, the exact nature of the task and the way in which stimuli are presented, as well as various applied factors that can differently affect performance (e.g., the quality and presentation of stimuli). Therefore, in considering the specific details of the *live* evaluations conducted so far, it may almost be reasonable to expect performance differences.

Kemp, et al., (1997) had cashiers verify shoppers' identities by assessing photographs and signatures displayed on credit cards. Megreya and Burton (2008) presented Human Operators with *live* individuals and displayed photographs on a projector screen. Butavicius, et al., (2008) procedure most closely resembled the procedure of Experiment 2, with one group of participants acting as targets who approached Human Operators and presented their IDs for Human Operators to assess. Consequently, in considering only the specific face matching tasks, it may be understandable that notable performance differences were found (discussed further in Chapter 7, Section 7.1).

Therefore, despite all the effort invested in attempting to simulate applied settings, the results need to be considered as carefully as those from laboratory experiments. Applied claims need to be appropriately qualified by explaining the exact nature of the face matching task as well as any other factors that may have affected performance. It is therefore logical to ask if all the effort that is necessary to simulate an applied setting and conduct a *live* performance evaluation is justified. This leads to one of the main questions addressed within this thesis, about the extent to which performance results from laboratory experiments are equivalent to those obtained by *live* performance evaluations. If that is the case this may reduce, if not eliminate, the need to conduct *live* performance evaluations, allowing for research within this field to focus on tailoring laboratory experiments to specific applied settings as well as hopefully lead to a form of methodological standardisation when assessing face matching performance (Chapter 7). To address this aim the next step is to replicate the *live* performance evaluation conducted as part of Experiment 2 by conducting it within a laboratory setting.

# *Chapter 6*

# *Experiment 3: Laboratory One-to-One Face Matching*

Experiment 3 was the final Human Operator one-to-one face matching performance experiment and intended to replicate the *live* Experiment 2 in the form of a controlled laboratory experiment while using the same stimuli. As such, it was methodologically very similar to Experiment 1. Human Operators were presented with a video of TPs, acquired during Experiment 2 (Section 5.1.3.1) and the same still image presented on TP ID cards during Experiment 2 (acquired during the Imaging Trial, Section 3.2). The results from this experiment help address the main aim of this research regarding the extent to which the findings from controlled laboratory one-to-one face matching performance experiments are translatable to real world access control settings.

Experiment 3 also evaluated the impact of the four impostor types and individual differences on face matching performance, however it only considered performance of Human Operators with face matching expertise. The reason for this was twofold. First, to reliably evaluate the current level of applied face matching performance, this study sought to assess performance of individuals who conduct face matching as part of their employment. Second, this study also sought to demonstrate the usability of the current methodology beyond only evaluating human performance, by assessing the performance of an automated FR system using the same stimuli. Presented next is a brief overview of previously conducted human and algorithm performance comparisons relevant to face matching tasks.

## *6.1 Human-Algorithm Performance Comparisons*

Performance assessments comparing human and automated FR performance have predominantly been conducted with the primary focus on improvement and development of automated systems (Adler & Maclean, 2004; Adler & Schuckers, 2007; Ding, Shu, Fang, & Ding, 2010; Hancock, Bruce, & Burton, 1998; O'Toole, Abdi, Jiang, & Phillips, 2007; O'Toole, et al., 2000; O'Toole, Phillips, et al., 2007; O'Toole, Phillips, & Narvekar, 2008; Phillips et al., 2007). As previously discussed (Chapter 1), this focus was motivated by increased security concerns which have emphasised the importance of reliable, accurate, and quick identification and verification of individuals. It therefore makes sense to know how the newly developed automated systems compare to the system that has traditionally been relied on, the Human Operator. A detailed discussion of all comparative evaluations is beyond the current scope. However, the following summary provides a brief overview relevant to current research.

Burton, et al., (2001) compared performance of a principal components analysis (PCA) based algorithm with humans on a one-to-ten face matching task, by presenting the same stimuli to both. They found that the automated system performed as well as or better than humans. Further, it was found that when the presenting images were rotated to 30 degrees, both human and automated performance declined, however, automated performance was more severely affected. Burton, et al., (2001) suggested that it is likely for automated FR performance to have been affected by changes in camera and even slight illumination variation.

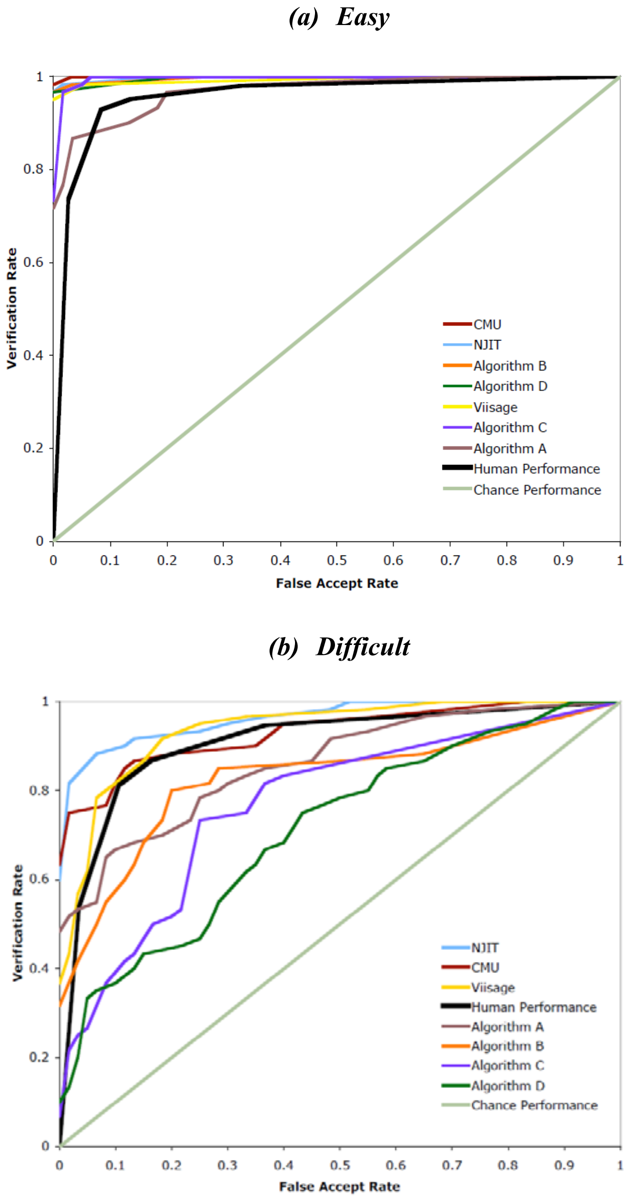
In another series of studies, Adler and Maclean (2004) and Adler and Schuckers (2007) aimed to develop a technique for the comparison of human-algorithm performance. To that end, Adler and Maclean (2004) first compared one-to-one face matching performance of humans with a range of best performing automated systems which were developed and available in 1999, 2001 and 2003. This evaluation was further extended by Adler and Schuckers (2007) to also include algorithms from 2005 and 2006. The results of the first evaluation revealed that even the best performing algorithms available in 2003 were outperformed by humans (Adler & Maclean, 2004). Adler and Schuckers (2007) however, reported a substantial improvement in algorithm performance. The best performing algorithm in 2006 revealed that while 29.2% of human participants performed better compared to automated systems, 37.5% performed worse than the algorithm (Adler & Schuckers, 2007).

Another comparative assessment was conducted as part of the Face Recognition Grand Challenge (FRGC) (O'Toole, Phillips, et al., 2007). Seven state-of-the-art algorithms matched all possible pairs of 16,028 target[21] (controlled illumination) and 8,014 probe images[22] (uncontrolled illumination). This resulted in almost 128 million face pairs. Algorithms produced a similarity score for each face pair, indicating their "decision" about whether the presented stimuli were the same or different. As it would not be plausible to present millions of stimuli to human participants, a set of 120 easy and 120 difficult face pairs was generated using a sampling procedure defined by a baseline PCA algorithm (details in O'Toole, Phillips, et al., 2007). Forty nine untrained participants viewed two still facial images for 2 seconds and rated their level of similarity on a 5-point Likert scale, ranging from 1 (sure they are the same person) to 5 (sure they are not the same

[21] Target image is referred to as a raw biometric image stored in the database. It is also often referred to as a gallery image, database image or a stored image (Jain, Flynn, & Ross, 2008).

[22] Probe image is an image that is acquired during authentication. Probe images are also often referred to as query or input images (Jain, et al., 2008).

person). Figure 20 shows ROCs which compare human and algorithm performance.



*(a) Easy*

*(b) Difficult*

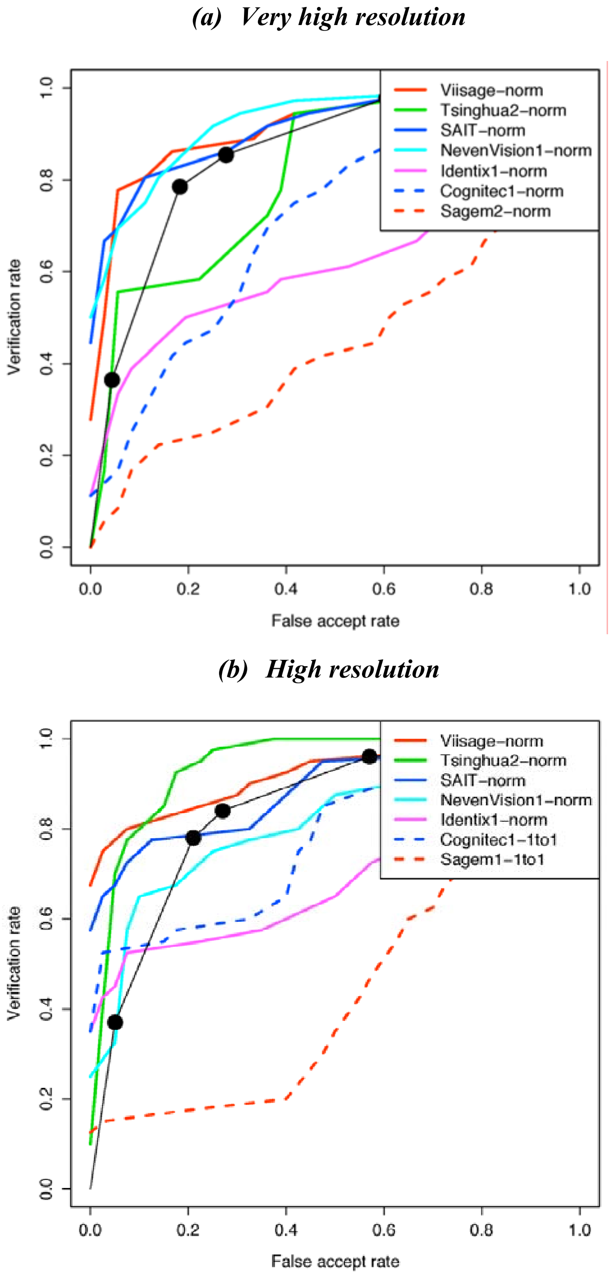
*Figure 20: Human and algorithm performance on easy and difficult face pairs, adapted from O'Toole, Phillips, et al., (2007)*

Figure 20 (a) shows performance on easy face pairs where six out of the seven algorithms performed better compared to humans. Figure 20 (b) shows that when presented with difficult face pairs, three algorithms were more accurate and four algorithms were less accurate than humans.

This work was further extended during the Face Recognition Vendor Test (FRVT) 2006 by assessing the same seven algorithms on a set of 5,402 ***very high resolution*** and 7,192 ***high resolution*** images (O'Toole, et al., 2008; Phillips, et al., 2007).[23] Unlike in FRGC, stimuli presented to human participants were not divided into easy and difficult stimuli. Instead, ***moderately difficult pairs*** were created based on the performance of the seven assessed algorithms. Selected were image pairs from the middle range of algorithm performance which were incorrectly judged by three to five out of the seven assessed algorithms. In the very high resolution condition, 25 untrained participants were presented with 36 true and 36 impostor face pairs. For the high resolution condition, 28 untrained participants were presented with 40 true and 40 impostor face pairs. Both conditions' stimuli were shown for 2 seconds and participants rated their similarity on the same scale used in the FRGC evaluation. Figure 21 (a) shows

[23] The *very high resolution* imagery was acquired using a 6 Megapixel Nikon D70 camera. The average face size for the controlled images was 400 pixels between the centres of the eyes and 190 for the uncontrolled images. The *high resolution imagery* was acquired using a 4 Megapixel Canon PowerShot G2. The average face size for the controlled images was 350 pixels between the centres of the eyes and 110 for the uncontrolled images (O'Toole, et al., 2008).

findings in the ***very high resolution*** condition, and Figure 21 (b) in the ***high resolution*** condition.



*(a) Very high resolution*

*(b) High resolution*

*Figure 21: Human (black) and algorithm performance on very high and high resolution imagery, adapted from O'Toole, et al., (2008)*

In considering the results from the FRGC and FRVT, a similar pattern emerges with algorithms consistently outperforming humans. Overall, these findings may suggest that, although algorithm performance can still be improved, it could enhance security within applied settings where humans currently conduct face matching tasks. Nevertheless, Human Operators assessed as part of FRGC and FRVT were undergraduate students who had no training or experience in conducting face matching tasks. It could therefore be argued that their performance is not representative of what actually occurs within applied settings within which Human Operators may have training and do have experience in conducting these tasks. Experiment 2 found that trained/experienced Human Operators performed significantly better than the untrained/inexperienced group.

Interestingly, a recent study compared face matching performance of 4,504 trained/experienced Human Operators with an automated FR system (Ding, et al., 2010). Human Operators and the algorithm were presented with image pairs which consisted of scanned licence photographs and still imagery acquired by a video camera positioned in a hallway. This imagery was divided into "easy", "middle" and "hard" categories as determined by similarity scores obtained by a different algorithm to the one which performance was assessed. They found that on the whole, the FR algorithm surpassed Human Operators. More specifically however, algorithm performance was superior for "easy" and "middle" image categories while on the "hard" category the performance of Human Operators was superior (Ding, et al., 2010). Consequently, Ding, et al., (2010) proposed a human-algorithm fusion which would involve Human Operators assessing imagery after it had been matched by the algorithm. They hope that this would reduce Human Operator workload and potentially increase accuracy and efficiency.

Having briefly considered empirical work focusing on comparing human and algorithm performance, it should be noted that the primary aim of the current human-algorithm assessment stems from wanting to more broadly apply the methodology that has been developed for the assessment of Human Operator applied performance. Therefore, the current evaluation differs from the ones previously conducted in that its primary focus has been the assessment of Human Operator abilities. Nevertheless, it is anticipated that the results will provide an indication of trained/experienced human and algorithm performance on the same one-to-one face matching task. Finally, it is important not to depart from the main aim of this research which focuses on assessing the feasibility of extrapolating laboratory findings to applied settings. This experiment is the final assessment, the laboratory assessment, and its results will be compared to those of the *live* Experiment 2 to answer this question.

## *6.2 Method*

### *6.2.1 Participants*

#### *6.2.1.1 Human Operators*

Ninety (54 females and 36 males) trained and/or experienced Human Operators participated in this experiment. Their ages ranged from 22 to 60 (*M* = 38.88, *SD* = 11.18). As shown in Figure 22, they were sourced from five different government agencies and conduct various face matching tasks as part of their employment. As in Experiment 2, the identity of the agencies is kept confidential for security reasons. They are referred to as Agency A, B, C, D, and E.[24]

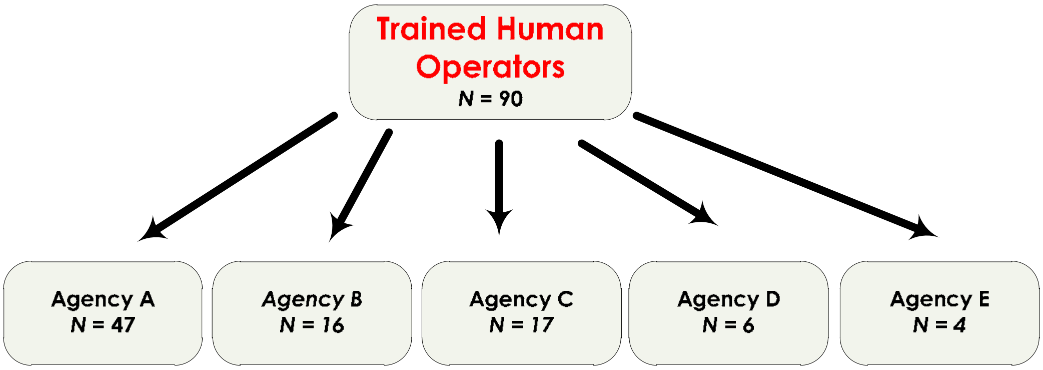


*Figure 22: Experiment 3: Trained/experienced Human Operators*

[24] Participating organisations were similar to those that took part in Experiment 2, however a different group of participants from these organisations took part in the experiment.

### *6.2.2 Materials*

#### *6.2.2.1 Target Participant Stimuli*

One hundred TPs' (31 females and 69 males) still and video imagery was used. Their ages ranged from 21 to 64 (*M* = 42.23, *SD* = 10.76). These stimuli were selected from a pool of 129[25] TPs who took part in both the Imaging Trial (Section 3.2.2.1) and also in the *live* performance evaluation as part of Experiment 2 (Chapter 4).

#### *6.2.2.2 Stimuli Preparation and Presentation*

Video imagery was acquired during the imaging part of Experiment 2 (Section 5.1.3.1). Videos were approximately 2 to 4 seconds in duration and displayed a TP walking towards the camera. Still imagery was the same as that used during both, Experiments 1 and 2, and consisted of images acquired during the Imaging Trial as well as impostor images from the external database. Matlab R2009a and the imaging toolbox were used for the display of TP video and still imagery. Once the experiment was prepared, it was displayed on 17 inch monitors (1152 x 870 resolution) and presented to Human Operators.

[25] The entire set of 129 TPs was not able to be used due to time restrictions. It should however be noted that this enabled the impostor rate to be set at exactly 30%.

#### *6.2.2.3 Individual Differences Tests*

To enable a meaningful comparison with findings from Experiment 2, the same battery of individual differences tests was used (Section 5.1.2.3). It should be noted that due to time constraints[26] associated with the conduct of this experiment, out of 90 Human Operators, 70 completed the individual differences tests.

#### *6.2.2.4 Automated Face Recognition System*

The automated FR software that was used in this experiment was provided for experimental use in August, 2008 by SAFRAN Morpho (former Sagem Sécurité). Detailed specifications of the software and how it works were not made available due to commercial reasons.

It should be noted that although improved versions of this software may be available as the vendor makes modifications and improvements, the results reported in this research are reflective of only the version of software that was made available for this research in August 2008.

[26] Different government organisations were able to allocate different amounts of time for the participation in the experiment.

### *6.2.3 Design and Procedure*

#### *6.2.3.1 Human Operator Performance Testing*

The same evaluation methodology as for Experiment 1 (Section 4.1.3.1) was adopted. Human Operators were seated in front of the monitor and provided with instructions about the experimental conduct. They first completed a set of demographic questions (e.g., age, gender, ethnicity etc.,) and were shown two examples of how stimuli would be presented throughout the experiment. Once that was completed, the main part of the experiment commenced.

During the experiment, Human Operators were first presented with a video of a TP approaching the camera. Once the video was shown, the last (close-up) image of the TP remained displayed on the screen. Alongside the video a still photograph was displayed. With the imagery displayed on the screen, Human Operators were asked: "Is this a match?" and provided with "Yes" and "No" options. After indicating their decisions, Human Operators were further asked to indicate how confident they were in their decision by clicking from 0 to 100 percent (divided into increments of 10). As shown in Figure 23, each Human Operator was presented with 100 video and still image pairs in a random order. Exactly 30 impostor images, equating to 30% impostor rate, were presented.

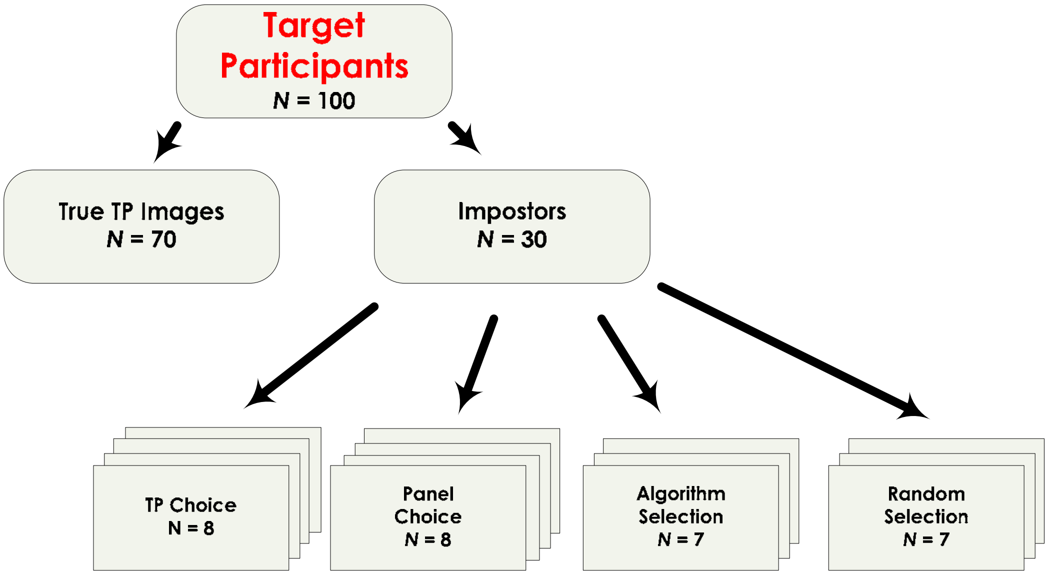


*Figure 23: Experiment 3: Target Participant imagery presented to Human Operators*

#### 6.2.3.2 *Algorithm Performance Testing*

Automated FR system performance was assessed offline. The same 100 TPs' video and still image pairs that were presented to Human Operators were also presented to the automated FR system. For each presentation a match score was obtained. The obtained match scores were used to present the results in the form of an ROC curve to enable comparison with Human Operator performance.

## *6.3 Results*

Presented first are results relevant to only Human Operator performance (Section 6.3.1), followed by human and automated system results (Section 6.3.2).

### *6.3.1 Human Operator Performance*

Overall Human Operator accuracy ranged from 64 to 100% with an average of 93.18% (*SD* = 5.7). Table 18 presents performance rates and SDT statistics.

*Table 18: Experiment 3: Performance rates and SDT indices*

| | ***Trained Human Operators***<br>*M (SD)* |
|---|---|
| **Hit Rate** | .92 (.08) |
| **Miss Rate** | .08 (.08) |
| **False Alarm Rate** | .07 (.09) |
| **Correct Response Rate** | .93 (.09) |
| ***Mean d'*** | 3.31 (.58) |
| ***Mean β*** | .30 (1.33) |

Further to performance rates and SDT indices, Figure 24 provides a visual illustration of Human Operator performance.

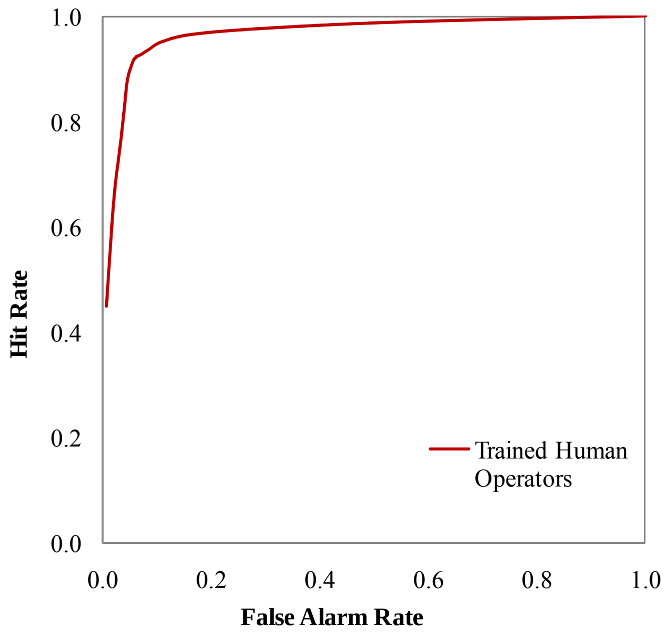


*Figure 24: Experiment 3: ROC of trained Human Operator performance*

#### *6.3.1.1 Human Operator Performance by Agency*

Performance of Human Operators was further divided by agency, presented in Table 19.

*Table 19: Experiment 3: Overall accuracies and performance rates by agency*

| | ***Agency*** | | | | |
|---|---|---|---|---|---|
| | ***A*** | ***B*** | ***C*** | ***D*** | ***E*** |
| | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* | *M(SD)* |
| **Overall Accuracy** | .92 (.04) | .94 (.06) | .89 (.07) | .95 (.03) | .93 (.04) |
| **Hit Rate** | .84 (.10) | .94 (.07) | .90 (.05) | .91 (.10) | .95 (.05) |
| **False Alarm Rate** | .04 (.07) | .06 (.08) | .11 (.12) | .03 (.03) | .10 (.13) |

Overall accuracies show that performance of all agencies was very similar, with only slightly lower performance achieved by Agency C. However, paired samples t-test only found a statistically significant difference for overall accuracies of Agency C and D (*t*(6) = -2.81, *p* < .05). All other comparisons were not significant.

Specific performance rates produced mixed results. Only Hit Rates for Agency A and B (*t*(5) = -2.86, *p* < .05); Agency A and E (*t*(6) = -3.13, *p* < .05); and Agency D and B (*t*(16) = 2.14, *p* < .05) were found to be statistically significantly different. All other Hit Rates and all False Alarm Rates were not found to be statistically different.

#### *6.3.1.2 The Impact of Impostor Types*

Table 20 shows Human Operator performance rates by impostor type. A difference in the rates for TP and Panel chosen compared to Algorithm and Randomly selected impostors can be observed.

*Table 20: Experiment 3: Hit Rates by impostor type*

| ***Impostor Type*** | ***Human Operators***<br>*M (SD)* |
|---|---|
| **TP Choice** | .85 (.18) |
| **Panel Choice** | .89 (.13) |
| **Algorithm Selection** | .98 (.05) |
| **Random Selection** | .98 (.05) |

A repeated measures ANOVA revealed statistically significant differences for rates obtained for different impostor types, $F(3, 90) = 36.12$, $p < .001$, Partial $\eta^2 = .29$. Paired samples t-tests further revealed statistically significant differences for TP and Panel ($t(89) = 2.53$, $p = .01$), TP and Algorithm ($t(89) = -6.95$, $p < .001$), TP and Random ($t(89) = -6.99$, $p < .001$), Panel and Algorithm ($t(89) = -6.18$, $p < .001$) and, Panel and Randomly ($t(89) = -6.10$, $p < .001$) selected impostors. However, no statistically significant difference was found between Algorithm and Randomly ($t(89) = -0.47$, $p = .64$) selected impostors. These results are similar to Experiments 1 and 2 findings, and are further considered in the overarching discussion (Section 7.1.1.2).

#### *6.3.1.3 Human Operator Confidence Ratings and Decision Latency*

Table 21 shows overall confidence rating and confidence ratings by impostor type.

*Table 21: Experiment 3: Confidence ratings overall and by impostor type*

| | ***Confidence Rating (%)***<br>*M (SD)* |
|---|---|
| **Overall** | 88.50 (8.78) |
| **TP Choice** | 82.50 (12.39) |
| **Panel Choice** | 84.60 (11.56) |
| **Algorithm Selection** | 90.30 (9.40) |
| **Random Selection** | 92.26 (7.96) |

Considering the raw confidence, it can be seen that confidence was lower when Human Operators viewed TP and Panel compared to when they viewed Algorithm and Randomly selected impostors. However, paired samples t-tests revealed statistically significant differences between all impostor conditions. Specific results for all combinations included TP and Panel (*t*(89) = -3.39, *p* = .001); TP and Algorithm (*t*(89) = -9.32, *p* < .001); TP and Random (*t*(89) = -10.76, *p* < .001); Panel and Algorithm (*t*(89) = -7.12, *p* < .001); Panel and Random (*t*(89) = -8.89, *p* < .001); and Algorithm and Random (*t*(89) = -3.08, *p* = .003).

Table 22 shows overall decision latency and decision latency by impostor type.

*Table 22: Experiment 3: Decision latency overall and by impostor type*

| | ***Decision Latency (s)***<br>*M (SD)* |
|---|---|
| **Overall** | 2.91 (0.76) |
| **TP Choice** | 3.24 (1.79) |
| **Panel Choice** | 3.01 (1.15) |
| **Algorithm Selection** | 3.03 (1.40) |
| **Random Selection** | 2.80 (0.99) |

Looking at decision latency by impostor type, it can be seen that Human Operators were making their face matching decisions in less than 4 seconds. Paired samples t-tests found statistically significant differences between the time taken to make decisions for TP and Random (*t*(89) = 3.27, *p* = .002) and Panel and Randomly selected impostor images (*t*(89) = 2.11, *p* = .04). No statistically significant differences were found between decision latencies for TP and Panel (*t*(89) = 1.66, *p* = .10), TP and Algorithm (*t*(89) = 1.17, *p* = .24), Panel and

Algorithm ($t(89) = -0.11$, $p = .91$) and, Algorithm and Randomly ($t(89) = 1.67$, $p = .10$) selected impostors.

#### *6.3.1.4 Human Operator Individual Differences*

Table 23 presents Pearson's correlations for Human Operator performance and confidence ratings, and individual differences tests. It can first be noted that confidence ratings were positively and moderately correlated with overall accuracy, however, negatively and also moderately correlated with False Alarm Rate. In terms of specific perceptual speed test's subscales, two correlations were found. Finding A's negatively correlated with False Alarm Rate, and Number Comparison subscale was associated with overall accuracy. A measure of face matching ability (i.e., GFMT) correlated moderately with overall performance accuracy, and also moderately, but negatively with False Alarm Rate. Finally, of the four Rational-Experiential Ability subscales, a negative weak correlation was found for Experiential Engagement and Hit Rate. These results are discussed further in the overarching discussion (Chapter 7).

*Table 23: Experiment 3: Pearson's correlations between trained Human Operator performance and individual differences tests*

| | ***Overall Accuracy*** | ***Hit Rate*** | ***False Alarm Rate*** |
|---|---|---|---|
| **Confidence Ratings** | **.38**** | .08 | **-.33**** |
| **Perceptual Speed** | | | |
| Finding A's | .22 | -.07 | **-.24*** |
| Identical Pictures | .12 | .07 | -.08 |
| Number Comparison | **.26*** | .09 | -.21 |
| **Glasgow Face Matching Test** | **.38**** | -.08 | **-.39**** |
| **Rational-Experiential Inventory** | | | |
| Rational Ability | -.20 | -.11 | .13 |
| Rational Engagement | -.15 | -.22 | .03 |
| Experiential Ability | .05 | -.13 | -.12 |
| Experiential Engagement | .05 | **-.25*** | -.18 |

* correlation is significant at *p* < .05

** correlation is significant at *p* < .01

### *6.3.2 Human Operator and Automated System Performance*[27]

Automated system performance was calculated based on separately considering match scores obtained for true and impostor stimuli presentations. True presentation match scores input into the generation of correct responses (and false alarms, in the form of 1 – correct response rate = false alarm rate). Impostor match scores input into miss rate (and Hit Rates, as 1 – Miss Rate = Hit Rate). This enabled the construction of appropriate ROC curves to enable comparison with Human Operator performance. This approach has been used in previous comparative assessments (e.g., O'Toole, Phillips, et al., (2007); O'Toole, et al., (2008)).

Automated FR performance was evaluated using video imagery acquired as part of Experiment 2 and still imagery used for the ID cards prepared during Image Preparation (Chapter 3). This ensured that Human Operators and the automated FR system both viewed TPs under similar conditions. Human Operator and FR algorithm performance are presented in the form of an ROC in Figure 25.

[27] Preliminary results of this work were presented at the International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society (Calic, McLindin, & Macleod, 2010).

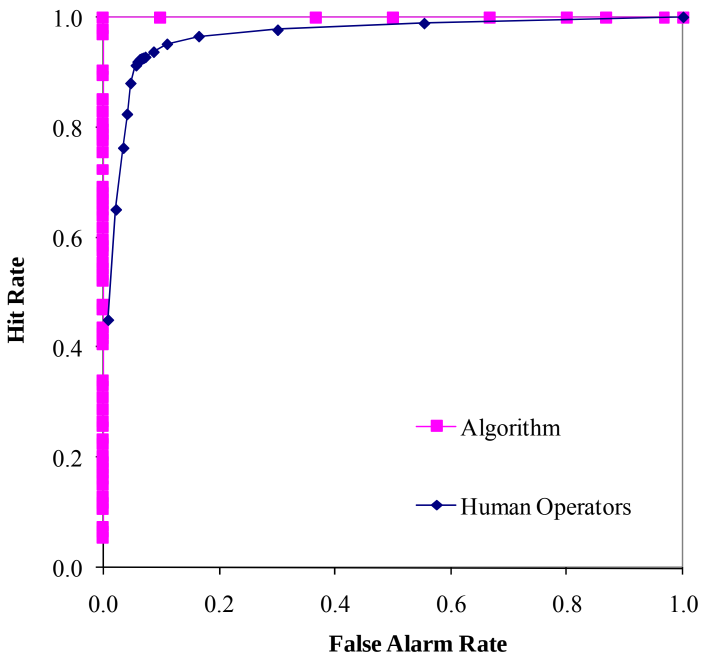


*Figure 25: Experiment 3: ROCs of Human Operator and algorithm performance*

A visual inspection of Figure 25 shows that algorithm performance is superior to that of trained and/or experienced Human Operators. In addition to visually comparing performance, it is also useful to consider algorithm match scores to assess if there were any differences in the way that the algorithm matched different impostor types. Table 24 shows mean match scores for true and impostor imagery. It can be observed that overall, impostor and true imagery scores were substantially different. This difference was found to be statistically significant ($t(94) = -9.52$, $p < .001$). However, as can be seen by the obtained match scores, there were no statistically significant differences between the match scores obtained for impostor stimuli.

*Table 24: Experiment 3: Automated FR system match scores by image type*

| ***Image Type*** | ***Match Score***<br>*M (SD)* |
|---|---|
| **True Imagery** | 5895.20 (2955.17) |
| **TP Choice** | 1240.88 (194.42) |
| **Panel Choice** | 1189.13 (176.11) |
| **Algorithm Selection** | 1183.57 (215.71) |
| **Random Selection** | 1060.71 (236.71) |

## *6.4 Summary of Experiment 3*

Experiment 3 was the final performance experiment. It was important for a number of reasons, outlined below:

- First, Experiment 3 replicated the *live* one-to-one face matching evaluation in the form of a controlled laboratory performance experiment. The findings from this evaluation input into answering the main aim of this research about the extent to which the findings from controlled laboratory one-to-one face matching experiments can be extrapolated to real world access control settings. This will be discussed in light of all findings in the main discussion (Chapter 7).

- Second, Experiment 3 continued to consider the impact of impostor types and individual differences on face matching performance. Also, in order to provide an assessment of the applied level of Human Operator performance, Experiment 3 only assessed the performance of trained and/or experienced Human Operators.
- Third, this study also evaluated performance of an automated FR system, using the same stimuli as that presented to Human Operators. As such, this study was instrumental in demonstrating the usability of the current methodology beyond only evaluating human performance.

As stipulated in brief discussions of Experiments 1 and 2 (Sections 4.4 and Section 5.3), the results of all three experiments are best considered collectively. Consequently, to best address the main aims of the current research and to avoid repetition, Chapter 7 jointly outlines the results of Experiments 1, 2 and 3, and discusses the key findings in light of previous empirical work.

# *Chapter 7*

## *Discussion and Conclusion*

The principal aim of this research was to evaluate the feasibility of extrapolating one-to-one face matching performance findings from a laboratory setting to the real world access control environment, and, in the process, to support the development of an ecologically motivated performance evaluation methodology that could be used for future performance assessments. The approach taken to address this aim stemmed from the focus on identity verification or the one-to-one face matching task, that is predominantly performed in access control settings. This focus ensured the applicability and relevance of current findings to appropriate real world settings and enabled the evaluation of factors that may affect face matching performance in these settings. Consequently, this research evaluated the impact of different rates of impostor frequency, different types of impostors, Human Operator expertise, and individual differences on one-to-one face matching performance.

A preliminary evaluation (Experiment 1) examined the effects of impostor frequency and type, and helped to establish parameters for Experiments 2 and 3. Experiment 2 compared trained and untrained Human Operator face matching performance in a simulated *live* access control setting. Experiment 3 replicated Experiment 2 within a laboratory setting and assessed trained Operator performance. Findings from Experiments 2 and 3 assist with answering the following questions:

1. What is the performance on a one-to-one face matching task within a *live* access control environment?
2. What is the performance on a one-to-one face matching task within a laboratory environment when the same stimuli used in the *live* setting are presented?
3. How do findings from the *live* and laboratory settings compare, and to what extent can the findings from the laboratory experiment be extrapolated to the *live* access control evaluation?

Results of all experimental work were presented throughout Chapters 4 to 6. However, discussions were brief, focusing only on the results relevant to the following experiments. This was done to avoid repetition and to allow for an overarching discussion (this chapter) which would jointly consider the results from all three experiments. Section 7.1 summarises and discusses one-to-one face matching performance across the three experiments.

## *7.1 Discussion of the Key Findings*

The findings of the three performance experiments conducted are discussed in light of previous similar experimental studies. It should be noted that because the frequencies of impostors presented in previously conducted experiments (i.e., 50%) and current evaluations (i.e., 30%) differ, overall accuracies' calculations are not based on the same proportions of true and impostor stimuli. Therefore, overall accuracies may not provide an accurate means for comparison between the experiments. Consequently, to appropriately compare current findings with previous work, the focus is on Hit Rates and False Alarm Rates.[28] First, laboratory Experiments 1 and 3 are considered in light of previous similar studies. Second, simulated *live* Experiment 2 is compared with previous *live* evaluations.

***Laboratory Experiments***

Table 25 provides a summary of laboratory Experiments 1 and 3 results and compares them to previous similar one-to-one face matching experiments.

[28] Note the definition of Hits and False Alarms (Section 4.2). To enable comparisons, previous works' findings are considered in terms of the current definitions of Hits and False Alarms.

*Table 25: Comparison of laboratory Experiments 1 and 3 with previous laboratory one-to-one face matching performance evaluations*

| | | ***Overall Accuracy (%)*** | ***Hits (%)*** | ***False Alarms (%)*** |
|---|---|---|---|---|
| **Lab Experiment 1** | | 96 | 92 | 4 |
| **Lab Experiment 3** | | 93 | 92 | 7 |
| **Megreya & Burton (2007)** | | 80 | 84 | 22 |
| **Megreya & Burton (2008)** | | 84 | 85 | 15 |
| **Burton, et al., (2010)** | **Short version** | 81 | 82 | 21 |
| | **Long version** | 89 | 88 | 8 |

Looking at Table 25 it can be seen that Experiments 1 and 3 had similar Hit Rates and False Alarms. Previous similar evaluations reported slightly lower Hit Rates and higher False Alarms. For example, on a one-to-one face matching task where Human Operators viewed two still images, Megreya and Burton (2007) found a Hit Rate of 84% and False Alarm Rate of 22%. In another study, Human Operators inspected a photograph and a static video image and achieved a Hit Rate of 85% and a False Alarm Rate of 15% (Megreya & Burton, 2008). Also, when assessed on the long and short versions of the Glasgow Face Matching Test (GFMT) which require participants to compare two greyscale still images shown simultaneously, Hit Rates and False Alarms were 82% and 21%, and 88% and 8%

for the short and long versions respectively (Burton, et al., 2010). These results demonstrate notable differences in the performance among similar empirical evaluations that focus on one-to-one face matching.

This performance variability may be attributed to differences among the face matching tasks, selected stimuli, and the conditions under which these stimuli were presented. While Megreya and Burton (2007, 2008) and Burton, et al., (2010) used greyscale still imagery, during Experiments 1 and 3, Human Operators viewed high quality colour imagery, both video and still. The stimuli and presentation were guided by applied access control settings where a live individual approaches a Human Operator who inspects their ID, which contains a high quality still taken under optimal conditions (e.g., a passport). Therefore, while the task may seem the same as it involves one-to-one face matching, the characteristics of the stimuli and the way in which they were presented may have differently affected Human Operator performance.

It should also be noted that of the research that focuses on one-to-one face matching performance, surprisingly, no previous work has been found that evaluated performance by presenting a high quality video and a still photograph to replicate access control settings. One similar experiment was conducted by presenting poor quality CCTV videos and still photographs (Lee, et al., 2009).

They reported 67% overall performance for trained and untrained participants.[29]

***Simulated Live Experiments***

Table 26 presents the results of simulated *live* Experiment 2 and three previous *live* one-to-one face matching performance evaluations.

*Table 26: Comparison of Experiment 2 with previous live one-to-one face matching performance evaluations*

| | | ***Overall Accuracy (%)*** | ***Hits (%)*** | ***False Alarms (%)*** |
|---|---|---|---|---|
| **Live Experiment 2** | **Untrained** | 91 | 80 | 5 |
| | **Trained** | 95 | 86 | 2 |
| **Kemp, et al., (1997)** | | 67 | 51 | 11 |
| **Megreya & Burton (2008)** | | 83 | 77 | 11 |
| **Butavicius, et al., (2008)** | | 95 | 91 | 4 |

Kemp, et al., (1997) were the first to evaluate *live* person-to-photo face matching performance, however, not within an access control setting. They assessed performance in a supermarket setting where six cashiers verified identities of 44

[29] The results of overall performance are compared because Lee at al, (2009) report only overall (identification) performance rate.

participating shoppers' by matching them to a photograph on their credit card. Four different types of image stimuli were presented to shoppers. Impostor images were divided into hard (i.e., matched foil) and easy (i.e., unmatched foil) impostor categories. True presentations were also divided into changed (i.e., involving small paraphernalia changes such as adding/removing prescription glasses, jewellery, etc.,) and unchanged appearance photographs. Impostor selections were made by experimenters who sourced target participants (i.e., the shoppers) and impostor imagery from a database of one hundred and fifty undergraduate students. Kemp, et al., (1997) reported a Hit Rate of 51% and False Alarm Rate of 11%. Although this result is most commonly cited as a baseline of applied human one-to-one face matching performance, a number of potential methodological issues associated with this study should be considered (Section 2.2.5).

Most notably, it is worth reiterating that in addition to the face matching task, the cashiers inspected shoppers' signatures, which were also presented on the credit card. The signatures were consistently valid, true signatures of the shopper carrying a particular card, even when the card was depicting an impostor image. Consequently, there is a possibility that the match decisions were based on, or influenced by, the signature. A number of additional potential problems with the way that cashiers were instructed to perform the task need to be noted. The cashiers were instructed to "process this card normally but that they should also check the photograph" (Kemp, et al., 1997, p. 216). It should be noted that “normal” processing of the card only involved verifying signatures and did not involve the checking of the photographs. Therefore, cashiers may have been biased to predominantly focus on the signatures, which were always true signatures of the shoppers. Kemp, et al., (1997) stated that while debriefing the cashiers the cashiers explained that they were very reluctant to request a second sample of a shoppers signature because, from their experience, this often provoked an aggressive response. Therefore, they believed that challenging a shopper on the basis of their appearance was even more likely to provoke an aggressive response, and that they would only challenge a shopper if they were absolutely certain that the photograph was not of the person presenting it. Therefore, performance rates reported by Kemp, et al., (1997) need to be considered with caution.

The two other *live* studies were conducted by Megreya and Burton (2008), and Butavicius, et al., (2008). Both reported higher Hit Rates compared to Kemp, et al., (1997). However, there is still a substantial difference between the two findings. Megreya and Burton (2008) reported Hit Rates of 77% and False Alarm Rates of 11% when a live individual and a greyscale still photograph projected on a screen were compared simultaneously. This performance is notably different to the current *live* evaluation findings. As explained for the laboratory studies, this performance difference is most probably attributable to the type and the presentation of image stimuli.

The design of Butavicius, et al., (2008) most closely resembles that of Experiment 2, as they aimed to mimic an applied access control setting. Fifty target participants presented with an ID to 10 Human Operators who were army personnel. The impostor frequency was 20% and impostors were selected by two experimenters as most closely resembling the target participants. Experiment 2 extended the work of Butavicius, et al., (2008). In addition to evaluating performance within a simulated *live* access control setting, Experiment 2 expanded the work on impostors by incorporating four different types of impostors which were generated in ways conceivable in the real world. Experiment 2 evaluated performance of both trained and untrained Human Operators as well as a number of individual differences which were subsequently correlated with performance (Section 7.1.1.4). Consequently, Experiment 2 focused on more than just the evaluation of *live* face matching performance. Furthermore, performance of the *live* Experiment 2 is compared to performance in laboratory evaluation using the same stimuli (Experiment 3) to address the extent to which laboratory findings can be extrapolated to the *live* setting.

Butavicius, et al., (2008) found the same overall accuracy that was achieved by trained Human Operators in Experiment 2.[30] However, when Hit Rates are considered notable differences in performance are noted. In Butavicius, et al.,

[30] Please note that overall results can be compared as an equivalent percentage of impostor and true stimuli were presented.

(2008) impostors were generated by two experimenters who selected imagery which they believed most closely resembled target participants. This is equivalent to the current Panel Choice impostor category. Therefore, when Hit Rates for Panel Chosen impostors are considered for trained and untrained Human Operators as part of Experiment 2 a notable difference in performance can be seen. While Butavicius, et al., (2008) report a Hit Rate of 91%, untrained Human Operators as part of Experiment 2 achieved a Hit Rate of 74% and trained achieved a Hit Rate of 76% on Panel Choice impostor category (Table 27). It is therefore surprising to see that even with very similar experimental designs the performance can differ substantially. This only serves to confirm the complexities associated with evaluating face matching performance and the extent to which performance depends on numerous factors.

Comparing the performance of laboratory (Experiments 1 and 3) and *live* (Experiment 2) studies with similar relevant experiments has revealed a notable difference in performance. As discussed above, this difference may be attributed to the differences in the characteristics and the presentation of the image stimuli.

Performance rates also reveal that the seemingly simple one-to-one face matching task is error prone. For example, as part of Experiment 2 Human Operators failed to detect 8 to 20% of impostors and on 2 to 7% of occasions, incorrectly declared that true presenters were impostors. From an applied perspective, these results may be concerning. Consequently, in line with previous research, current results presented confirm the difficulty that people have with processing facial information of unfamiliar individuals (Bruce, et al., 2001; Burton, et al., 1999; Hancock, et al., 2000; Megreya & Burton, 2006, 2007, 2008; Pike, et al., 2000). It has been suggested that this occurs because unfamiliar face matching relies on situation specific representations (e.g., pose, lighting, expression, etc.,) which limit generalisation beyond specific representations. This is different to familiar face recognition for which generalisation beyond any specific conditions is easily achieved (Bruce & Young, 1986). Hancock, et al., (2000) have also suggested that processing of unfamiliar faces relies on image matching, rather than on a more sophisticated face matching strategies used for familiar faces. However, much work is still needed to improve our understanding of unfamiliar face processing.

### *7.1.1 Effects of Experimental Factors and Human-Algorithm Performance Comparison*

The following sections discuss the impact of impostor frequency (Section 7.1.1.1) and impostor type (Section 7.1.1.2); Human Operator expertise (Section 7.1.1.3); and individual differences (Section 7.1.1.4) on one-to-one face matching performance. Comparison of human and automated system performance is discussed in Section 7.1.1.5.

#### *7.1.1.1 Impostor Frequency*

The impact of different frequencies of impostors on one-to-one face matching performance was considered because in the majority of applied settings the occurrence of impostors is not known, yet researchers typically present 50% of impostor stimuli (Section 2.2.4). Consequently, the impact of impostor frequency was assessed in Experiment 1 during which Human Operators were presented with either, 0, 10, 20 or 30% of impostor stimuli.

As these findings were relevant to the design of Experiments 2 and 3 they were considered in the discussion section of Experiment 1, Section 4.4. In summary, it was found that face matching performance was not affected by different impostor frequency. No difference in overall performance between the 10, 20, and 30% impostor conditions was found. A similar result was reported by Bindemann, et al., (2010) who assessed impostor rates of 2 and 50% and found that participants' impostor detection was not affected. Therefore, current findings may serve to confirm that previous research which predominantly utilises 50% impostor rates does not undermine the difficulty of the impostor detection task. Based on this finding, it was decided that the impact of impostor frequency would not be explored further beyond Experiment 1. Therefore, Experiments 2 and 3 used the higher 30% impostor rate, which ensured sufficient analysable data to support meaningful analyses.

A neat extension of this work would involve an evaluation of the extent to which informing Human Operators to expect more or less, or a certain frequency of impostors would affect their face matching performance. This scenario is also plausible from an applied perspective as there may be situations where Human Operators in applied settings would be pre-warned about the potential of incoming impostors, or through experience, would expect a certain rate of impostors.

#### *7.1.1.2 Impostor Types*

The focus on impostor types was motivated by the lack of empirical evidence considering real world impostor creation and its impact on face matching performance. Although previous research had paid much attention to stimulus creation and presentation, to author's knowledge, only Kemp at al., (1997) stated that their selection of impostor type (i.e., judged to look similar to target participant) was based on what might occur within the criminal community. No other studies have explicitly considered impostor generation based on the real world or ecologically plausible scenarios.

Consequently, current work considered four different ways in which impostors could be generated in the real world and assessed their impact on one-to-one face matching performance (Section 3.3). TP Choice simulated the situation in which an individual selected another person's identification documentation (e.g., passport) which contained an image that they believe looked most similar to them.

Panel Choice may be equivalent to a person obtaining a fraudulent identification (e.g., passport or visa application) generated by group selection. Algorithm Selection simulated the situation in which fraudulent identification documentation (e.g., passport) is selected by an FR algorithm. Random Selection would equate to using a found source of ID. Of the four assessed impostor categories, this would be the one that is least likely to occur. Table 27 provides a summary of results.

*Table 27: Human Operator Hit Rates by impostor type*

| | ***Experiment 1*** | ***Experiment 2*** | | ***Experiment 3*** |
|---|---|---|---|---|
| ***Impostor Type*** | ***Untrained***<br>*M (SD)* | ***Untrained***<br>*M (SD)* | ***Trained***<br>*M (SD)* | ***Trained***<br>*M (SD)* |
| **TP Choice** | .87 (.18) | .63 (.22) | .76 (.23) | .85 (.18) |
| **Panel Choice** | .86 (.19) | .74 (.21) | .76 (.26) | .89 (.13) |
| **Algorithm Selection** | .97 (.07) | .91 (.06) | .95 (.10) | .98 (.05) |
| **Random Selection** | .98 (.06) | .95 (.07) | .99 (.03) | .98 (.05) |
| **False Alarm Rate** | .04 (.06) | .05 (.04) | .02 (.03) | .07 (.09) |

A clear disparity between TP and Panel Choice compared to Algorithm and Random selected impostors can be observed.[31] It was found that TP and Panel chosen impostors were more difficult to detect compared to Algorithm and Random selected impostors for which the detection rate was always above 90%. While it may not be surprising to find that human selected impostors (i.e., TP and Panel Choice) were challenging for Human Operators, it may be surprising to see that Human Operators did not have much difficulty with detecting Algorithm selected impostors. This finding serves to show likely differences in human and algorithm processing of facial stimuli. It may seem that the algorithm does not focus on salient features (e.g., gender) that guide human face processing. Nonetheless, while the algorithm match score did not seem to be a good indicator of facial similarity as perceived by humans, when algorithm performance was assessed as part of Experiment 3 the algorithm was consistently able to accurately reject all impostor stimuli, even that selected by the algorithm (Section 7.1.1.5).

The disparity in detection of TP and Panel chosen impostors compared to Algorithm and Random selected impostors was further supported when confidence ratings and decision latencies were considered.[32] Human Operators consistently reported lower confidence when presented with TP and Panel chosen

[31] The only exception was a finding as part of Experiment 3 which revealed that there was a significant difference between performance on TP and Panel chosen impostors.

[32] Due to the logistical nature of the experiments, decision latency was not collected during the *live* access control simulation as part of Experiment 2.

impostors compared to when presented with Algorithm and Randomly selected impostors. However, further analyses did not reveal statistically significant differences. Decision latency findings indicated that decision latencies for TP and Random, and Panel and Random selected impostors were statistically different.[33]

Therefore, Hit Rates, confidence ratings and decision latency show a reliable disparity between impostor stimuli that were generated by participants (TP Chosen) and by a panel of judges (Panel Choice) compared to impostor stimuli generated by an algorithm (Algorithm Selection) and selected randomly (Random Selection). This finding is akin to hard and easy impostor conditions commonly used in FR research (Kemp, et al., 1997). It seems that the hard category is comparable to TP and Panel choice, while Algorithm and Random selections correspond to the easy category, presented in Table 28.

*Table 28: Human Operator Hit Rates by hard and easy impostor types*

| | ***Experiment 1*** | ***Experiment 2*** | | ***Experiment 3*** |
|---|---|---|---|---|
| ***Impostor Type*** | ***Untrained***<br>*M (SD)* | ***Untrained***<br>*M (SD)* | ***Trained***<br>*M (SD)* | ***Trained***<br>*M (SD)* |
| **Hard** | .86 (.16) | .68 (.18) | .76 (.22) | .88 (.14) |
| **Easy** | .97 (.05) | .93 (.06) | .97 (.05) | .98 (.04) |

[33] Specific results are in Sections 4.3.3, 5.2.5, and 6.3.1.3 for Experiments 1 to 3, consecutively.

Although this is an initial empirical exploration of the impact of ecologically generated impostors on face matching performance, the results are promising and the general concept could be considered as part of further empirical work. The methodology presented herein which was used to create the impostor categories can serve as an initial conceptual guide and motivate further studies to explore the impact of impostors.

#### *7.1.1.3 The Impact of Human Operator Expertise*

The primary reason for evaluating Human Operator expertise stems from the need to better understand if, and the extent to which experience with faces and facial imagery, beyond everyday social interactions, and specialised training improve face matching performance. The approach adopted here was to assess face matching performance of individuals who conduct face matching as part of their employment and/or may have received a form of face matching training (i.e., trained), and compare their performance to that of lay individuals (i.e., untrained).

Figure 26 provides a summary of face matching performance of untrained and trained Human Operators. Hit Rate is divided into easy and hard impostor types (Section 7.1.1.2).

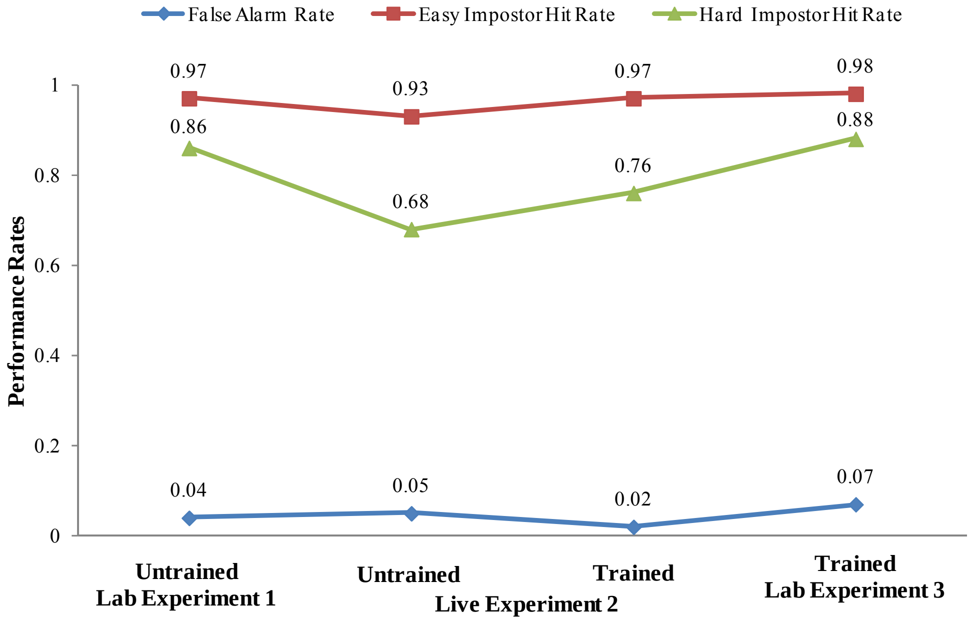


*Figure 26: A summary of untrained and trained Human Operator one-to-one face matching performance across all experiments*

A direct statistical performance comparison of untrained and trained Human Operators was conducted as part of Experiment 2.[34] Overall accuracies and *d'* measures for untrained (.91 and 2.63) and trained (.95 and 3.41) Human Operators were statistically significantly different, indicating superior one-to-one face matching performance for trained Human Operators. However, this difference was not found when specific performance rates were considered separately.

[34] A statistical comparison between untrained Human Operators from Experiment 1 and trained Human Operators from Experiment 3 cannot be conducted as two were distinct experiments and involved presentation of different stimuli.

Trained and untrained Human Operator performance can be explored further by considering the performance of participating agencies. Comparing performance across agencies as part of Experiment 2 (Section 5.2.2.1) revealed that Agency C had lower overall performance compared to other agencies.[35] Looking at Figure 27, it can be seen that Hit Rates of untrained Human Operators and Agency C are notably lower compared to other agencies, with Agency C's Hit Rates even lower than that of untrained Human Operators. Agency C, unlike other participating agencies, is substantially less required to conduct face matching tasks. Face matching tasks do not form the core of Agency C's business practices but are an additional task which is performed intermittently. Thus, their staff may be less trained and/or experienced and therefore their performance is more in line with that of untrained Human Operators.

[35] Equivalent results were found in Experiment 3 (Section 6.3.1.1).

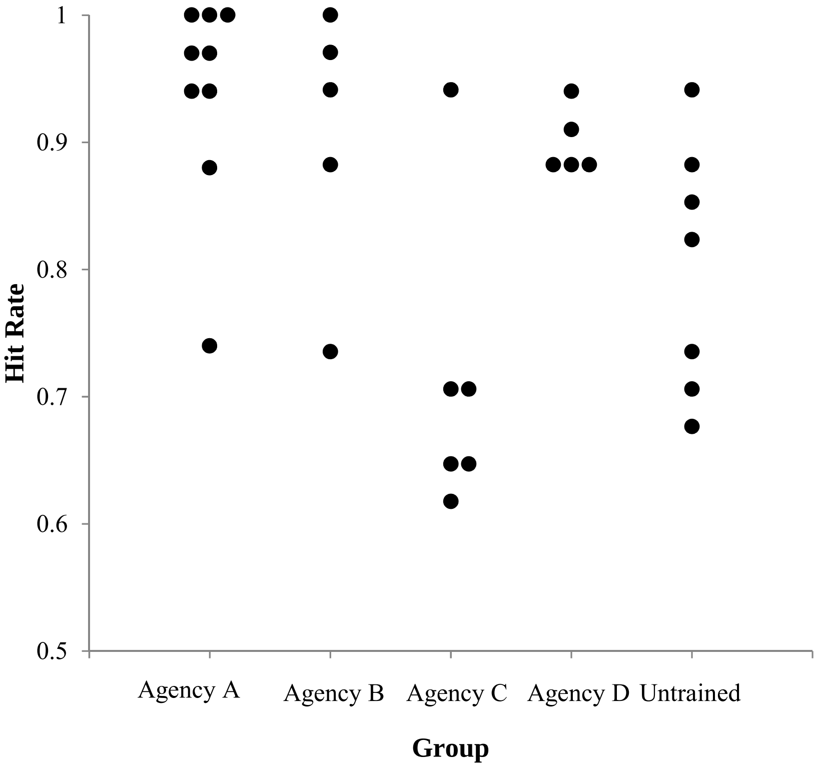


*Figure 27: Live Experiment 2: Hit Rates of participating agencies' and untrained Human Operators*

Previous empirical research has only minimally considered face matching expertise (Section 2.2.3.3). Burton, et al., (1999) compared face memory performance of police officers and students, and found no difference in the performance of the police and students. Furthermore, on a similar task where a group of trained and untrained Human Operators matched poor quality CCTV and still images Lee, et al., (2009) did not find a significant difference in performance. Contrary to these findings, Wilkinson and Evans (2009) found that two CCTV experts (the two authors of the paper) performed reliably better compared to untrained participants. However, this study has since received criticism for overstating their findings (Edmond, et al., 2010). Another related evaluation compared performance of nine state and federal agencies of trained Human Operators across Australia with untrained individuals. This evaluation found no performance difference between trained and untrained operators on *one-to-many* face matching task (Heyer, MacLeod, Hopley, Semmler, & Ma-Wyatt, 2011; Heyer et al., 2011). Although, one-to-many face matching task is different to *one-to-one* face matching evaluated as part of the current studies, it should be noted that both assessments recruited participants from similar state and federal Australian agencies and therefore the results are comparable to the current findings.

In light of previous works' findings, it may be difficult to make many claims about why trained Human Operators performed better as part of Experiment 2. This performance advantage could be attributed to a number of different aspects, such as selection, formal training or on-the-job experience. However, it is currently not possible to attribute their performance to any one or a combination of these aspects. Instead, it may seem that trained participants who conduct face matching tasks as part of their employment and who may have also received a form of face matching training have, through these activities, developed numerous strategies that assist them with the performance of this task. Because of general familiarity with the task, trained Human Operators may be more able to focus on task relevant features and successfully dismiss task irrelevant and potentially distracting extraneous factors. Essentially, trained/experienced individuals may be deploying their attentional resources and approaching the face matching task in a much more strategic and deliberate manner. Likewise, it may be that as a result of formalised training and/or their on-the-job experience they are more likely to focus on diagnostic facial features. This is important because it may mean that face matching is not associated with an inherent ability, but that it may be a skill that is able to be developed through formalised and on-the-job training.

This hypothesis needs to be evaluated through future research. A larger sample of trained and untrained participants would need to be evaluated. Also, in line with current experiments, a neat extension of this work could involve a replication of Experiment 3 by presenting the stimuli to an untrained group and statistically comparing the performance of the two. Further comparative evaluations could be conducted by presenting challenging and non-optimal facial imagery (e.g., non-portrait, poor quality, etc.,) as it is difficult to determine whether the current stimuli were not sufficiently difficult, and therefore, unable to appropriately tap into performance differences between trained and untrained participants. However more broadly, in conducing expertise relevant research, it is important to think about the concept of expertise and what face matching expertise constitutes.

Current findings may have practical implications, especially in light of recent debates about face matching and recognition expertise and admissibility of photographic evidence into court (Edmond, et al., 2010). The main problem associated with face matching expertise stems from the belief that people are generally good with faces which has historically resulted in ad-hoc recruitment for such tasks. As a result, individuals with various educational and employment backgrounds are recruited into a job that requires them to conduct face matching tasks. These individuals do not always have the opportunity to participate in formalised training regimes on how to conduct these face matching tasks prior to commencing employment. They may be required to attend a face matching course during their employment, however, that may not always occur as such courses/training may not be available.

Spaun (2009) from the US Federal Bureau of Investigation (FBI) proposed that a structured training process for facial image comparison experts needs to be developed and provides a thorough overview of attributes of such an expert post training. Spaun (2009) stated that such an expert needs to have extensive knowledge in "comparative science, image science and processing, bones of the head, muscles of the face, properties of the skin, aging and alteration, legal issues and case law, and the history of facial identifications and photographic comparisons." (p. 161). Furthermore, a distinction is made between two different levels of training and skill required, depending on whether controlled or uncontrolled facial imagery is being examined.[36] Consequently, a distinction was made between the type of training required for *Standardised Image Examiners* who would mainly deal with controlled imagery, and *Uncontrolled Image Examiners* who would be concerned with uncontrolled imagery. In an attempt to standardise, provide guidelines and reach a level of consensus from a practitioner perspective, the Facial Identification Scientific Working Group (FISWG) have worked on and released a series of relevant documents, namely, *Guidelines and Recommendations for Facial Comparison Training to Competency*, *Guidelines for Facial Comparison Methods*, and *Recommendations for Training Program in Facial Comparison* (Facial Identification Scientific Working Group (FISWG), 2011, 2012a, 2012b).

In line with this it is important for FR research to align its focus with practical requirements and consider Human Operator face matching and recognition expertise while it is still in embryonic stages. Appropriate research may focus on the development of training regimes and what they should entail to maximise applied benefits. At this stage, it is worth noting that training may be application specific as for instance, one-to-one face matching compared to one-to-many

[36] Controlled imagery would be of high quality, taken under controlled lighting, displaying neutral, uncluttered backgrounds. It would show individuals in a frontal pose with a neutral facial expression. Uncontrolled imagery would be commonly acquired under non-standardised environmental conditions, at various distances, displaying different facial angles (e.g., images acquired at crime scenes, surveillance images).

would differ substantially in the amount of time, effort and focus required to complete the task. In the first instance empirical work could focus on designing simple training, evaluating performance pre and post training, and assessing if improvement in performance occurred (similar to work by Semmler, et al., (2012)). In order for this research to be successful and applicable, research institutions and practitioner agencies need to work together. This would enable researchers to ask and address appropriate questions and ensure that rigorous and reliable research findings guide applied decision making.

Finally, in spite of the move towards a reliance on automated FR solutions, it has often been argued that applied face matching tasks are associated with complexities which may not allow for full automation (Butavicius, et al., 2008; Zhao, et al., 2003). Of course, this would depend on the specific application in question. Nonetheless, it should be acknowledged that the Human Operator will remain an important part of applied settings. It is therefore important to invest much time and resources into developing appropriate training solutions closely coupled with evaluations of environmental, imagery, as well as any other contextual factors which can impact on Human Operator performance in diverse applied settings.

#### *7.1.1.4 The Impact of Individual Differences*

Experiments 2 and 3 assessed the impact of individual differences on one-to-one face matching performance. As suggested by Megreya and Burton (2006), perhaps the large performance differences on unfamiliar face matching tasks may be explained and informed by how they perform on individual differences tests. A better understanding of individual differences relevant to face matching tasks may also have the potential to guide personnel selection and recruitment.

Current work used a set of standardised measures to evaluate Human Operator perceptual speed, face matching ability, and rational-experiential thinking styles (Section 5.1.2.3) and correlated the scores on these measures with face matching performance. In brief, analyses revealed that face matching performance correlated with measures of perceptual speed, face matching ability and rational-experiential thinking styles. Overall correlations found here were modest, which is in line with Megreya and Burton's (2006) findings.

More specifically, perceptual speed measure's subscales showed a similar pattern of results in both experiments. Face matching performance correlated with Finding A's[37] and Number Comparison subscales, but not with Identical Pictures subscale. A slightly different finding was reported by Megreya and Burton (2006) who used the same perceptual speed measure. They found that performance on a one-to-ten face matching task correlated with Finding A's and Identical Pictures

[37] However, correlations with Finding A's in Experiment 2 were positive and moderate compared to low and negative in Experiment 3.

subscales, but not with Number Comparison subscale. When details of the specific subscales are considered, the difference between these results seems sensible.

The Number Comparison subscale involves a comparison of two numbers presented side by side by assessing each individual digit of the two presented numbers. This task somewhat resembles the one-to-one face matching task as when comparing two facial stimuli side by side their features can be evaluated individually to determine similarity. It may therefore make sense to find that the Number Comparison subscale was correlated with one-to-one face matching. The Identical Pictures subscale, on the other hand, requires participants to find one (target) picture presented among four other pictures. This is similar to a one-to-many face matching task, and it seems appropriate that correlations with performance on Identical Pictures task and Megreya and Burton’s (2006) one-to-ten face matching task were found. While this seems intuitively plausible, further similar research using the same perceptual speed measures is required to explore this finding further. At this stage, however, it could be suggested that performance on one-to-one vs one-to-many face matching tasks may be better predicted by different types of individual differences measures.

Face matching ability was assessed by using the GFMT. Both experiments’ overall accuracies were found to be moderately correlated with the overall GFMT score. This may not be surprising as the current experiments and GFMT focus on one-to-one face matching. GFMT has been designed with the aim to provide a one-to-one face matching performance indicator. In light of the current findings it would be plausible to suggest that GFMT could be adopted within applied settings where one-to-one tasks are performed. The current findings are consistent with other work which focused on one-to-one and one-to-many face matching (Heyer, et al., 2010).

The Rational-Experiential Inventory was used to assess Human Operator thinking styles. Findings were not consistent across the two experiments. The results of Experiment 2 found moderate negative correlations for overall face matching accuracy and Rational Ability subscale, and for Hit Rate and Rational Ability and Engagement subscales. Experiment 3 findings showed only a weak negative correlation between Hit Rate and Experiential Engagement. According to Pacini and Epstein (1999) who developed the Rational-Experiential Inventory, rational ability is defined as conscious, analytical and relatively affect-free compared to experiential ability which is said to rely on preconscious, rapid, automatic processing which is associated with affect. It may therefore make more sense for face matching to be associated with rational ability. However, correlation analyses did not support this assumption. Instead, face matching performance was found to be negatively correlated with both rational ability and engagement and experiential engagement. Consequently, future work should further consider this instrument to establish its usability.

Finally, the types of tests used here need to be considered. It should be noted that the perceptual speed tests may not be appropriately tapping into any specific underlying abilities integral to face matching tasks. In considering this, it is important to be explicit about whether what is being evaluated really are individual differences rather than skills that can be developed after prolonged exposure to a complex perceptual tasks. The perceptual speed tests seem to focus more on attention, and skills and strategies that people may be able to adopt and develop when conducting such tasks. More broadly, however, further work on individual differences may focus on an entirely different set of measures. In the first instance, it would be valuable to consider the profiles of individuals who display above average face matching performance compared to poor performers.

#### *7.1.1.5 Human-Algorithm Performance Comparison*

Automated FR system performance was evaluated as part of Experiment 3. This allowed a comparison of human and automated performance on a controlled one-to-one face matching task. Additionally, this evaluation demonstrated the usability of the current methodology beyond only assessing human performance.

Experiment 3 found that one-to-one face matching performance of the automated FR system was superior compared to that achieved by trained Human Operators. This finding is in line with a number of previous human-automated comparisons although these studies did not assess trained Human Operators (O’Toole, Abdi, et al., 2007; O'Toole, Phillips, et al., 2007; O'Toole, et al., 2008). The only other study known to have compared trained Human Operator performance with that of an algorithm was conducted by Ding, et al., (2010) who assessed the performance of 4,504 trained/experienced individuals. They found that algorithm performance was superior for "easy" and "middle" stimuli categories while on the "hard" category Human Operator performance was superior compared to that of the algorithm. Consequently, Ding, et al., (2010) proposed a human-algorithm fusion which would involve Human Operators assessing imagery after it had been matched by an algorithm. They argued that this would reduce workload and ultimately have the potential to increase accuracy and efficiency in applied settings. Ding, et al., (2010) finding that Human Operators outperform algorithms on difficult face stimuli is further confirmed by a study conducted by Biswas, Bowyer, and Flynn (2011). Biswas, et al., (2011) compared performance of untrained Human Operators and two automated systems to distinguish between identical twins. They found that Human Operators outperformed two different automated FR systems.

Although algorithm performance improved significantly and continues to improve, it should be noted that performance of the current algorithm (Experiment 3) and those evaluated by Ding, et al., (2010) and Biswas, et al., (2011) cannot be generalised to all FR algorithms. A good example of the variability of different automated FR systems is presented in the FRVT evaluations where performance of numerous algorithms was compared (Blackburn, Bone, & Phillips, 2000; Phillips et al., 2003; Phillips, et al., 2007). Furthermore, when considering algorithm performance, it is important to be cognisant of differences based on the type of evaluation. Previous research has shown that algorithm performance differs significantly from technology to scenario, and especially to operational evaluations (Introna & Nissenbaum, 2009). While an algorithm may perform favourably as part of a technology evaluation where testing conditions are highly controlled, its performance may suffer substantially when assessed as part of a scenario or an operational evaluation where conditions are not controlled.

Furthermore, performance of the algorithm evaluated in Experiment 3 may be attributed to the ease of the face matching task for algorithm processing. This algorithm was developed based on earlier models which have been implemented into applied access control settings and may thus be robust to poor image quality and environmental variations. However, during Experiment 3, performance was assessed offline using high quality controlled video and still imagery. This image quality may not be available in all applied settings. Therefore, the high quality stimuli may have made the task easy for the algorithm. Consequently, similar to Human Operator findings, algorithm performance results may be affected by the task and stimuli that are presented. Experiment 3 and previous similar work have used good quality imagery in controlled settings. Perhaps these results would be different if algorithm performance was assessed on poor quality imagery in less controlled settings. Previous automated FR system research has consistently shown that imagery type, characteristics of imagery, and, especially, quality can substantially affect FR algorithm performance (Blackburn, et al., 2000; Lui et al., 2009; McLindin, 2005). To better understand the differences and similarities in human and automated performance, further research should focus more on evaluating the performance of both under the same conditions.

### *7.1.2 Comparison of Live and Laboratory One-to-One Face Matching Findings*

The primary aim of this research was to compare Human Operator one-to-one face matching performance in *live* and laboratory settings. To address this aim, Human Operator one-to-one face matching performance was first evaluated in a simulated *live* access control setting, Experiment 2, which was subsequently replicated in the form of laboratory, Experiment 3. The results of these two evaluations are compared (shown in Table 29 and Figure 28) and offer a potentially important insight about the extent to which laboratory findings replicate applied access control settings.[38]

[38] For the purposes of comparison with Experiment 3, only the results of trained Human Operators from Experiment 2 are considered. Also, this comparison is based on the same 100 stimuli which were used in both experiments.

*Table 29: Comparing live Experiment 2 and laboratory Experiment 3 findings*

| | ***Live Experiment 2*** | ***Lab Experiment 3*** |
|---|---|---|
| | ***Trained***<br>*M (SD)* | ***Trained***<br>*M (SD)* |
| **Overall Accuracy** | .93 (.04) | .93 (.05) |
| **Hit Rate** | .84 (.13) | .92 (.08) |
| **False Alarm Rate** | .03 (.03) | .07 (.09) |

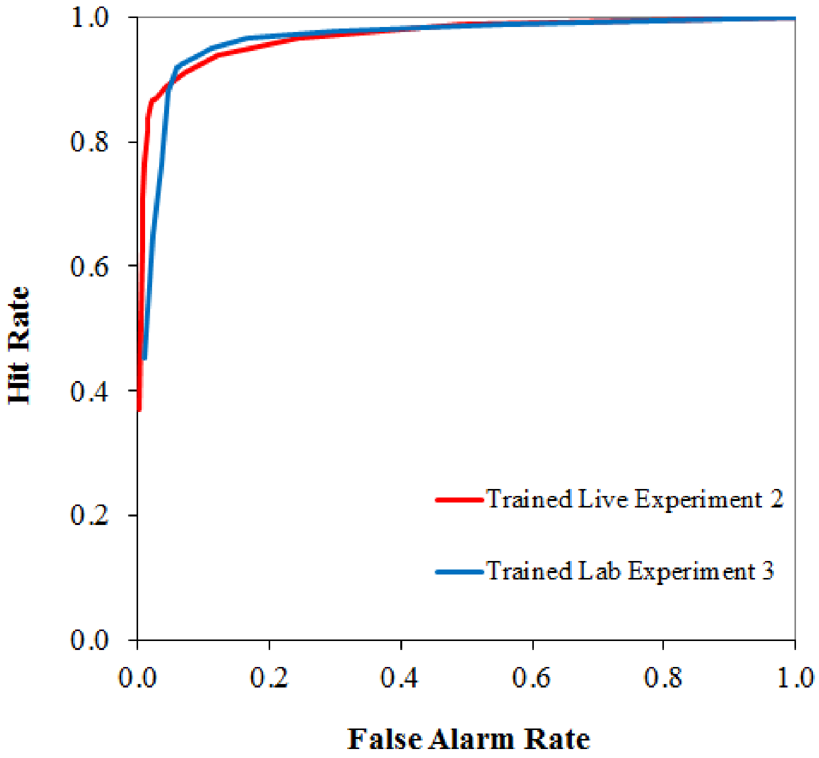


*Figure 28: ROCs comparing live Experiment 2 and laboratory Experiment 3 findings for trained Human Operators*

In considering specific performance rates, it can be seen that as part of *live* Experiment 2, Human Operators did not detect 16% of impostors compared to 8% in the laboratory Experiment 3. Human Operators also correctly matched 97% of legitimate presenters in the *live* Experiment 2, compared to 93% in its laboratory replication, Experiment 3. This result suggests that in the *live* setting Human Operators were more likely to state that an individual presenting an ID and the image on the presented ID were a match, compared to in the laboratory setting where stimuli were presented in form of a video and a still image. This suggests a confirmation bias in the *live* condition for participants to claim that two stimuli are the same.

Current findings are similar to previous results by Megreya and Burton (2008) who assessed the impact of liveness on one-to-one face matching.[39] In the *live* condition, Megreya and Burton (2008) simultaneously presented Human Operators with live participants and a high quality digital photograph projected on a screen. In Megreya and Burton's (2008) static condition, equivalent to the current laboratory Experiment 3, Human Operators were simultaneously presented with a static video image and a high quality digital photograph. A summary of current and Megreya and Burton's (2008) results is presented in Table 30.

[39] Considered is Megreya and Burton's (2008) third experiment as it focused on one-to-one face matching performance.

*Table 30: Comparison of live and laboratory one-to-one face matching performance: Current and Megreya and Burton's (2008) findings*

| | | ***Current Research*** *(%)* | ***Megreya & Burton (2008)*** *(%)* |
|---|---|---|---|
| ***Live*** | **Hits** | 84 | 77 |
| | **False Alarms** | 3 | 11 |
| ***Laboratory*** | **Hits** | 92 | 85 |
| | **False Alarms** | 7 | 15 |

Comparisons between performance rates in the *live* and laboratory conditions reveal that results of the current and Megreya and Burton's (2008) research follow the same pattern. These findings indicate a potential distinction in what occurs in a *live* or applied as opposed to a laboratory setting. In the *live* condition, participants are more biased to claim that two stimuli are the same. Essentially, Human Operators were more likely to conclude that live individuals presenting a facial image on an ID card were a match compared to when they viewed moving or a static video of an individual and compared it to a facial image presented on a computer screen. One explanation for this result may come from considering access control settings, as this is the most likely situation where this type of one-to-one face matching task may be performed.

In applied access control settings, the occurrence of impostors, although not known, is believed to be very low. As a result Human Operators may have a higher tendency to positively confirm identities. It may therefore be possible that when presented with *live* individuals as opposed to image stimuli on a computer screen, Human Operator decision making reflected that employed in the real world where the presence of impostors is believed to be low. Another explanation for this finding may be that the sheer presence of a *live* individual as opposed to image stimuli on a computer screen makes Human Operators less inclined to "reject" individuals and declare them impostors in person. Thus, Human Operators may find it easier to "reject" identities and declare impostors when stimuli are shown on a computer screen, in a form of a video or still image. In such settings, Human Operators do not have direct contact with individuals being verified and may be inclined to, as a result, differently perceive the impact of their decision.

It is important to reiterate a number of differences between the current work (i.e., Experiments 2 and 3), and Megreya and Burton's (2008) experiments. Experiment 2 and 3 were designed with the aim to closely simulate what may occur in applied access control settings and therefore differ substantially to the design adopted by Megreya and Burton (2008). For example, in the *live* Experiment 2, Human Operators matched a live person to a high quality coloured photograph presented in the form of an ID card. In order to simulate the live setting in the laboratory (i.e., Experiment 3) Human Operators matched a high quality coloured video to a still facial image. Also, the Human Operator sample comprised either trained/experienced individuals or lay individuals. However, Megreya and Burton's (2008) *live* condition involved Human Operators comparing a live person to a facial image projected on a screen. In the laboratory condition, Human Operators were presented with a still video frame and a still image. Also, the Human Operator sample comprised of undergraduate university population. Nevertheless, despite these differences the pattern of results in terms of what may happen in the *live* and laboratory settings is the same.

Finally, in respect to the main aim of this research, current results support the conclusion that it is feasible to extrapolate face matching performance findings from laboratory settings to the real world access control environment. The laboratory setting however, needs to closely resemble the real world setting in question and it is important to be cognisant of the confirmation bias tendency within the *live* setting.

## *7.2 Concluding Remarks*

This research was motivated by an applied problem concerned with evaluating and understanding human face matching performance in applied settings. However, evaluating face matching performance in surveillance and access control applied settings is extremely difficult if not impossible due to numerous experimental and logistical difficulties. As a result, human face matching performance has been evaluated in laboratory experiments, where variables can be controlled and monitored. However, the extent to which the results from controlled laboratory experiments explain and inform what happens in the real world is not known. This may be especially challenging when wanting to determine the extent to which laboratory results can be used to inform real world applications. Consequently, the principal aim of this research was motivated by this applied problem and posed the question of whether results from controlled laboratory experiments are representative of what happens in the real world.

In summary, it was encouraging to discover that the main findings indicated that there were little or no differences in *overall* face matching performance between the evaluations conducted in a simulation of *live* and controlled laboratory setting. This suggests that findings from laboratory settings can be generalised to applied access control environments. However, it should also be noted that specific performance rates revealed that in the simulated *live* access control setting, Human Operators were more inclined to indicate that two stimuli were a match, suggesting a confirmation bias. Perhaps unsurprisingly, this serves to indicate that laboratory experiments do not sufficiently capture all aspects of applied face matching performance. This is not to say that all future face matching performance evaluations must involve real world simulations of environments and face matching tasks of interest. Rather, it is important for researchers to be aware of and appropriately consider, and account for, any impact that this finding may have on their research conduct, and even more importantly, the appropriate application of their findings.

Further important results relate to the impact of a number of applied factors on face matching performance. The focus on one-to-one face matching within access control settings helped identify numerous factors that may affect face matching performance within these settings. As a result, this research evaluated the impact of impostors, Human Operator expertise and individual differences on one-to-one face matching performance. In relation to the specific factors, performance was significantly affected by impostor type, but not frequency. The effect of expertise was small and mainly confined to a shift in discrimination ability – experts were statistically better. There were small but reliable effects of individual differences – face matching performance correlated moderately with measures of perceptual speed, face matching ability and rational-experiential thinking styles. The automated face matching system was found to exhibit near perfect performance, exceeding that of the Human Operators. However, like with any automated system the generalisability of these findings to all automated systems should be considered with caution.

Finally, from a methodological perspective the aim of this research has been to develop an ecologically motivated methodology that could be used for future performance assessments. Consequently, the methodology adopted during this research was closely dependant on what happens when performing the one-to-one face matching task within access control settings. The current methodological approach serves as an example and a concept demonstrator for how to address applied research questions while appropriately considering an applied setting of interest. As such, it is anticipated that this methodological approach becomes commonly adopted as part of future research. Finally, it is anticipated that an ecologically motivated methodology facilitates the applicability and relevance of laboratory findings to inform the real world (as the title of this thesis suggests).

# *References*

Adler, A., & Maclean, J. (2004, September). *Performance comparison of human and automatic face recognition.* Paper presented at the Biometrics Consortium Conference, Washington, DC, USA.

Adler, A., & Schuckers, M. E. (2007). Comparing human and automatic face recognition performance. *IEEE: Transactions on Systems Man and Cybernetics*.

Anastasi, J. S., & Rhodes, M. G. (2006). Evidence for an own-age bias in face recognition. *North American Journal of Psychology, 8*, 237-252.

Ariely, D., & Zakay, D. (2001). A timely account of the role of duration in decision making. *Acta Psychologica, 108*(2), 187-207.

Ashbourn, J. (2005). Biometric system integration. In J. Wayman, A. K. Jain, D. Maltoni & D. Maio (Eds.), *Biometric Systems: Technology, Design and Performance Evaluation*: Springer.

Australian Associated Press Pty Limited (AAP). (2010, January 29). SmartGate passport hits 1m milestone, *The Sydney Morning Herald*. Retrieved from http://news.smh.com.au/breaking-news-national/smartgate-passport-hits-1m-milestone-20100129-n2ud.html

Australian Customs and Border Protection Service. (2010). Annual Report 2009-10. Canberra: Australian Customs and Border Protection Service.

Australian Customs Service. (2004). Overview of SmartGate trial (current at February 2004).

Australian Customs Service. (2007). Fact Sheet 4: SmartGate series 1, from http://www.customs.gov.au/webdata/resources/files/FS_SmartGate_Series1.pdf

Australian Government: Department of Immigration and Citizenship. (2009, 9 March 2011). Fact Sheet 71 - SmartGate automated border processing, from http://www.immi.gov.au/media/fact-sheets/71smartgate.htm

Beurskens, A. J. H. M., Bültmann, U., Kant, I., Vercoulen, J. H. M. M., Bleijenberg, G., & Swaen, G. M. H. (2000). Fatigue amongst working people: validity of a questionnaire measure. *Occupational and Environmental Medicine, 57*, 353-357.

Bindemann, M., Avetisyan, M., & Blackwell, K.-A. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied, 16*(4), 378-386.

Bird, C., Found, B., & Rogers, D. (2010). Forensic document examiners' skill in distinguishing between natural and disguised handwriting behaviors *Journal of Forensic Science, 55*(5), 1291-1295.

Biswas, S., Bowyer, K. W., & Flynn, P. J. (2011). *Study of face recognition of identical twins by humans*. Paper presented at the International Workshop on Information Forensics and Security (WIFS 2011), Foz do Iguacu, Brazil.

Blackburn, D. M., Bone, M., & Phillips, J. P. (2000). Facial Recognition Vendor Test 2000: Evaluation report: DoD Counterdrug Technology Development Program Office, Defense Advanced Research Projects Agency, National Institute of Justice.

Bolle, R. M., Connell, J. H., Pankanti, S., Ratha, N. K., & Senior, A. W. (2004). *Guide To Biometrics*. New York: Springer.

Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual Cognition, 10*, 527-536.

Bronstein, M. A., Bronstein, M. M., & Kimmel, R. (2006). Expression-invariant 3D face recognition. In W. Zhao & R. Chellappa (Eds.), *Face Processing: Aadvanced Modelling and Methods* (pp. 159-184): Academic Press.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*, 105-116.

Bruce, V. (1988). *Recognising faces*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bruce, V., Hancock, P. J. B., & Burton, A. M. (1998). Human face perception and identification. In H. Wechsler, J. P. Phillips, V. Bruce, F. Soulie & T. Huang (Eds.), *Face Recognition: From Theory to Applications*. Berlin: Springer-Verlag.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology, 5*(4), 339-360.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207-218.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305-327.

Buckhout, R., & Regan, S. (1988). Explorations in research on the other-race effect in face recognition. In M. M. Gruneberg, P. E. Morris & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues: Volume 1. Memory in everyday life* (pp. 40-46). New York: Wiley.

Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: A comparison across image formats. *Vision Research, 41*(24), 3185-3195.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods, 42*, 286-291.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor quality video: Evidence from security surveillance. *Psychological Science, 10*, 243-248.

Butavicius, M., Mount, C., Macleod, V., Vast, R., Graves, I., & Sunde, J. (2008). An experiment on human face recognition performance for access control *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 141-148): Springer Berlin / Heidelberg.

Caldara, R., & Abdi, H. (2006). Simulating the "other-race" effect with autoassociative neural networks: Further evidence in favor of the face-space model. *Perception, 35*, 659-670.

Calic, D. (2007). Human and machine facial recognition performance *Postgraduate Research Expo*. Adelaide, SA.

Calic, D. (2008). Methodology for the assessment of human facial recognition performance *Postgraduate Research Expo*. Adelaide, SA.

Calic, D., Macleod, V., McLindin, B., & Dunn, J. (2010, 8-10 April). *Face recognition: The ability of trained and untrained humans to detect impostors*. Paper presented at the 37th Australian Experimental Psychology Conference (EPC), The University of Melbourne.

Calic, D., & McLindin, B. (2009). *Human and automated facial recognition performance: A preliminary assessment*. Paper presented at the 31st Annual Conference of the Cognitive Science Society, Amsterdam.

Calic, D., McLindin, B., & MacLeod, V. (2009). *Methodology for the comparative assessment of trained and untrained Human Operators with automated facial recognition systems*. Paper presented at the Biometrics Institute Annual Conference, Sydney. presentation retrieved from

Calic, D., McLindin, B., & Macleod, V. (2010, 5-9 September). *Facial recognition: The ability of humans and algorithms to detect impostors*. Paper presented at the 20th International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society (ANZFSS), Sydney.

Carey, S., & Diamond, R. (1977). Science. *From piecemeal to configurational representation of faces, 195*, 312-314.

Cellerino, A., Borghetti, D., & Sartucci, F. (2004). Sex differences in face gender recognition in humans. *Brain Research Bulletin, 63*, 443-449.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology, 48*(4), 879-894.

Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory and Cognition, 26*, 780-790.

Cinque, S. (2009). MPEG Streamclip. Rome: Squared 5 srl, http://www.squared5.com/.

Clutterbuck, R., & Johnson, R. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception, 31*, 985-994.

Crookes, L., & McKone, E. (2009). Early maturity of face recognition: No childhood development of holistic processing, novel face encoding, or face-space. *Cognition, 11*, 219-247.

Dancey, C. P., & Reidy, J. (2002). *Statistics without maths for Psychology: Using SPSS for Windows* (Second ed.): Pearson Education Limited.

Department of Foreign Affairs and Trade (DFAT). (2005). Photograph guidelines: Your passport photos Retrieved 16 August, 2007, from https://www.passports.gov.au/Web/Requirements/Photos.aspx

Dewhurst, T., Found, B., & Rogers, D. (2008). Are expert penmen better than lay people at producing simulations of a model signature? *Forensic Science International, 180*, 50-53.

Ding, L., Shu, C., Fang, C., & Ding, X. (2010). *Computers do better than experts matching faces in a large population.* Paper presented at the 9th IEEE International Conference on Cognitive Informatics (ICCI), Beijing

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

Dror, I. E., Charlton, D., & Pe´ron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International, 156*, 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*.

Dyer, A. G., Found, B., & Rogers, D. (2006). Visual attention and expertise for forensic signature analysis. *Journal of Forensic Science, 51*(6), 1397-1404.

Edmond, G. (2010). Impartiality, efficiency or reliability? A critical response to expert evidence law and procedure in Australia. *Australian Journal of Forensic Sciences*, 1-17.

Edmond, G., Kemp, R., Porter, G., Hamer, D., Burton, A. M., Biber, K., & Roque, M. S. (2010). Case Note: Atkins v The Emperor: the 'cautious' use of unreliable 'expert' opinion. *The International Journal of Evidence and Proof, 14*, 146-166.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception, 8*(4), 431-439.

Facial Identification Scientific Working Group (FISWG). (2011). Guidelines and recommendations for facial comparison training to competency

Facial Identification Scientific Working Group (FISWG). (2012a). Guidelines for facial comparison methods.

Facial Identification Scientific Working Group (FISWG). (2012b). Recommendations for a training program in facial comparison.

Fletcher, K., Butavicius, M. A., & Lee, M. D. (2008). Attention to internal face features in unfamiliar face matching. *British Journal of Psychology, 99*, 379-394.

Fraser, F. (2004). *Security in government: SmartGate – successful use of biometrics*. Paper presented at the Security in Government Conference. www.ag.gov.au/...SmartGate_Security.../Conferencepapers_2004Fiona+Fraser_SmartGate_Security+in+Gov+Presentation.pdf

Fraser, I. H., Craig, G. L., & Parker, D. M. (1990). Reaction time measures of feature saliency in schematic faces. *Perception, 19*(5), 661-673.

Frias, C. M. D., Nilsson, L.-G., & Herlitz, A. (2006). Sex differences in cognition are stable over a 10-year period in adulthood and old age. *Aging, Neuropsychology, and Cognition, 13*, 574–587.

Frowd, C., Bruce, V., McIntyre, A., & Hancock, P. A. (2007). The relative importance of external and internal features of facial composites. British Journal of Psychology. *British Journal of Psychology, 98*, 61-77.

Furl, N., Phillips, J. P., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science, 26*, 797-815.

Georgia Institute of Technology. (1999). Georgia tech face database Retrieved September, 2008, from http://www.anefian.com/face_reco.htm

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition, 118*, 201-210.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.

Godard, O., & Fiori, N. (2010). Sex differences in face processing: Are women less lateralized and faster than men? *Brain and Cognition, 73*(3), 167-175

Goldstone, R. L. (2003). Do we all look alike to computers? *TRENDS in Cognitive Sciences, 7*(2), 55-57.

Graves, I. (2008). [Discussion relating to Customs Primary Line processing].

Graves, I., Johnson, R., & McLindin, B. (2003). *Problems with false accept rate in operational access control systems*. Paper presented at the 4th Australian Information Warfare and IT Security Conference, Adelaide.

Grill-Spector, K., Golarai, G., & Gabrieli, J. (2008). Developmental neuroimaging of the human ventral visual cortex. *Trends in Cognitive Sciences, 12*(4), 152-161.

Haga, S., Shinoda, H., & Kokubun, M. (2002). Effects of task difficulty and time-on-task on mental workload. *Japanese Psychological Research, 44*, 134-143.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (1998). A comparison of two computer-based face identification systems with human perceptions of faces. *Vision Research, 38*(15-16), 2277-2288.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*(9), 330-337.

Havard, C. (2007). *Eye movement strategies during face matching.* PhD, University of Glasgow, Glasgow.

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology, 15*, 445-464.

Heyer, R., MacLeod, V., Calic, D., Kuester, N., & McLindin, B. (2009). *Security through science: Human sciences contributions to biometrics research*. Paper presented at the Defence Human Sciences Symposium, Melbourne.

Heyer, R., MacLeod, V., Hopley, L., Semmler, C., & Ma-Wyatt, A. (2011). Profiling the facial identification practitioner in Australia: Report on the Human Operator Capability Project survey. Edinburgh, South Australia: DSTO.

Heyer, R., Semmler, C., MacLeod, V., Calic, D., McLindin, B., Ma-Wyatt, A., & Hopley, L. (2011). *Towards an understanding of facial identification practitioners in Australia: The Human Operator Capability Project*. Paper presented at the 12th Biometrics Institute Conference, Sydney.

Heyer, R., Semmler, C., & McLindin, B. (2010). *Identification using an automated facial recognition system: towards an understanding of human operator decision making*. Paper presented at the International Symposium of the Forensic Sciences (ANZFSS'10), Sydney.

Hill, H., & Bruce, V. (1996). The effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 986-1004.

Hillstrom, A. P., Sauer, J., & Hope, L. (2011). Training methods for facial image comparison: A literature review: University of Portsmouth, Department of Psychology.

Hole, G. (1994). Configurational factors in the perception of unfamiliar faces. *Perception, 23*, 65-74.

Introna, L. D., & Nissenbaum, H. (2009). Facial recognition technology: A survey of policy and implementation issues. New York: The Center for Catastrophe Preparedness & Response.

Itier, R. J., & Taylor, M. J. (2004). Face inversion and contrast-reversal effects across development: In contrast to the expertise theory. *Developmental Science, 7*(2), 246-260.

iTWire. (2011). Facial recognition - Emerging as the fastest growing segment. Retrieved from http://www.itwire.com/press-release/44541-facial-recognition-emerging-as-the-fastest-growing-segment

Jain, A. K., Flynn, P. J., & Ross, A. A. (Eds.). (2008). *Handbook of biometrics*: Springer.

Jain, V., & Mukherjee, A. (2002, September 2008). The Indian face database, from http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/

Johnson, K. J., & Fredrickson, B. L. (2005). "We all look the same to me": Positive emotions eliminate the own-race bias in face recognition. *Psychological Science, 16*(11), 875-881.

Kemp, R. (2009). *Some limitations of human face processing: Implications for the design of automated face recognition systems*. Presented at the Face Recognition User Advisory Group (FRUAG) Meeting

Kemp, R., & Howard, M. (2007, 7-8 June ). *The importance of considering the performance of human operators when designing facial biometrics systems.* Paper presented at the The Biometrics Institute Australia, Sydney.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11*, 221-222.

Keval, H., & Sasse, M. A. (2008). *Can we ID from CCTV? Image quality in digital CCTV and face identification performance.* Paper presented at the the SPIE Conference on Mobile Multimedia/Image Processing, Security and Applications.

Kirby, D. (2008, August 19). Face recognition takes off, *Manchester Evening News*. Retrieved from http://menmedia.co.uk/manchestereveningnews/news/s/1063254_facial_recognition_takes_off

Knight, B., & Johnston, A. (1997). The role of movement in face recognition. *Visual Cognition, 4*(3), 265-273.

Krouse, F. L. (1981). Effects of pose, pose change, and delay on face recognition performance. *Journal of Applied Psychology 66*(5), 651-654.

Kuefner, D., Cassia, V. M., Picozzi, M., & Bricolo, E. (2008). Do all kids look alike? Evidence for an other-age effect in adults. *Journal of Experimental Psychology: Human Perception and Performance, 34*(4), 811-817.

Lamont, A. C., Stewart-Williams, S., & Podd, J. (2005). Face recognition and aging: Effects of target age and memory load. *Memory & Cognition, 33*(6), 1017-1024.

Lander, K., & Bruce, V. (2003). The role of motion in learning new faces. *Visual Cognition, 10*(8), 897-912.

Lee, M. D., Vast, R. L., & Butavicius, M. (2006). *Face matching under time pressure and task demands.* Paper presented at the 28th Annual Conference of the Cognitive Science Society, Vancouver.

Lee, W.-J., Wilkinson, C., Memon, A., & Houston, K. (2009). Matching unfamiliar faces from poor quality closed-circuit television (CCTV) footage: An evaluation of the effect of training on facial identification ability. *Axis: The Online Journal of CAHId, 1*(1), 19-28.

Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology, 22*(6), 1364-1382.

Levin, D. T. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *General Journal of Experimental Psychology, 129*(4), 559-574.

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition - women's faces make the difference. *Brain Cognition, 50*(1), 121-128.

Li, J., Tian, M., Fang, H., Xu, M., Li, H., & Liu, J. (2010). Extraversion predicts individual differences in face recognition. *Communicative & Integrative Biology, 3*(4), 295-298.

Liu, C., & Wechsler, H. (2005). Face Recognition. In J. Wayman, A. K. Jain, D. Maltoni & D. Maio (Eds.), *Biometric Systems: Technology, Design and Performance Evaluation* (pp. 97-114). London: Springer.

Lui, Y. M., Bolme, D., Draper, B., Beveridge, J. R., Givens, G., & Phillips, J. P. (2009). *A meta-analysis of face recognition covariates.* Paper presented at the 3rd IEEE Intternational Conference, Washington D.C.

MacLeod, V. (2010). *Humans and biometrics: How well-intentioned humans can ruin a perfectly good biometric system*. Paper presented at the Biometrics Institute Conference, Sydney.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race faces. *Journal of Personality and Social Psychology, 13*, 330-334.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Science, 6*, 255-260.

McKelvie, S. J., Standing, L., Jean, D. S., & Law, J. (1993). Gender differences in recognition memory for faces and cars: Evidence for the interest hypothesis. *Bulletin of the Psychonomic Society, 31*(5), 447-448.

McKone, E., Crookes, K., & Kanwisher, N. (2009). The cognitive and neural development of face recognition in humans. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences (4th ed.).* Cambridge: MIT Press.

McLindin, B. (2002). *Biometrics Systems Analysis: Canada and USA overseas visit report*. (DSTO-OR-0497). Adelaide: Defence Science and Technology Organisation.

McLindin, B. (2005). *Improving the performance of two dimensional facial recognition systems - The development of a generic model for biometric technology variables in operational environments.* PhD, University of South Australia, Adelaide.

McNicol, D. (2005). *A Primer of Signal Detection Theory* New Jersey: Lawrence Erlbaum Associates.

Megreya, A. M., & Bindemann, M. (2009). Revisiting the processing of internal and external features of unfamiliar faces: The headscarf effect. *Perception, 38*, 1831-1848.

Megreya, A. M., Bindemann, M., & Havard, C. (2011). Sex differences in unfamiliar face identification: Evidence from matching tasks. *Acta Psychologica, 137*, 83-89.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory and Cognition, 34* (4), 865-876.

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception and Psychophysics, 69*(7), 1175-1184.

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

Megreya, A. M., Memon, A., & Havard, C. (2011). The headscarf effect: Direct evidence from the eyewitness identification paradigm. *Applied Cognitive Psychology*. doi: 10.1002/acp.1826

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3-35.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36*(4), 630-633.

National Academy of Sciences. (2010). Biometric recognition: Challenges and opportunities. Washington, D.C.

National Science and Technology Council (NSTC). (2006). Biometrics history (pp. 27).

NorPix: Digital Video Recording Software. (2009). StreamPix (Version 4). Montreal: www.norpix.com.

Nowosielski, A. (2006). *Face recognition methods integration for visitor identification systems.* PhD, Szczecin University of Technology.

O'Toole, A. J., Bülthoff, H. H., Troje, N. F., & Vetter, T. (1995). *Face recognition across large viewpoint changes.* Paper presented at the the International Workshop on Automatic Face and Gesture Recognition.

O'Toole, A. J. (2004). Psychological and neural perspectives on human face recognition. In S. Z. Li & A. K. Jain (Eds.), *Handbook of Face Recognition*: Springer-Verlag.

O'Toole, A. J., Abdi, H., Jiang, F., & Phillips, J. P. (2007). Fusing face verification algorithms and humans. *IEEE: Transactions on Systems, Man & Cybernetics, 37*, 1149-1155.

O'Toole, A. J., Jiang, F., Roark, D., & Abdi, H. (2006). Predicting human performance for face recognition. In W. Zhao & R. Chellappa (Eds.), *Face Processing: Advanced Modelling and Methods* (pp. 293-319). New York: Academic Press.

O'Toole, A. J., Phillips, J. P., Cheng, Y., Ross, B., & Wild, H. A. (2000). *Face recognition algorithms as models of human face processing.* Paper presented at the IEEE Fourth International Conference on Face and Gesture Recognition, Los Alamitos, CA.

O'Toole, A. J., Phillips, J. P., Jiang, F., Ayyad, J., Penard, N., & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces across changes in illumination. *IEEE: Transactions on Pattern Analysis and Machine Intelligence, 29*, 1642-1646.

O'Toole, A. J., Phillips, J. P., & Narvekar, A. (2008). *Humans versus algorithms: Comparisons from the Face Recognition Vendor Test 2006.* Paper presented at the 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, The Netherlands.

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences, 6*(6), 261-266.

O'Toole, A. J., & Tistarelli, M. (2010). Face recognition in humans and machines. In M. Tistarelli, S. Z. Li & R. Chellappa (Eds.), *Handbook of Remote Biometrics, Advances in Pattern Recognition*: Springer-Verlag London Limited.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*(6), 972-987.

Parasuraman, R. (1986). Vigilance, monitoring, and speech. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance* (Vol. 2). Toronto, Canada: John Wiley & Sons.

Parasuraman, R., Visser, E. d., Clarke, E., McGarry, W. R., Hussey, E., Shaw, T., & Thompson, J. C. (2009). Detecting threat-related intentional actions of others:Effects of image quality, response mode, and target cuing on vigilance.

Phillips, J. P., Grother, P., Micheals, R., Blackburn, D. M., Tabassi, E., & Bone, J. M. (2003). Face Recognition Vendor Test 2002: Evaluation report: National Institute of Standards and Technology (NIST).

Phillips, J. P., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions of Pattern Analysis and Machine Intelligence, 22*(10), 1090-1104.

Phillips, J. P., Scruggs, W. T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L., & Sharpe, M. (2007). FRVT 2006 and ICE 2006 large-scale results. Gaithersburg: National Institute of Standards and Technology (NIST).

Phillips, J. P., Wechsler, H., Huang, T., & Rauss, P. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing, 16*, 295-306.

Pike, G., Kemp, R., & Brace, N. (2000, March). *The psychology of human face recognition.* Paper presented at the IEE Electronics and Communications: Visual Biometrics, London.

Pike, G., Kemp, R., Towell, N. A., & Phillips, K. C. (1997). Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition, 4*(4), 409-437.

Porter, G. (2009). CCTV images as evidence. *Australian Journal of Forensic Sciences, 41*(1), 11-25.

Prosapognosia Research Centres. Retrieved 01 February, 2011, from http://www.faceblind.org/research/

Rehnman, J. (2007). *The role of gender in face recognition.* Doctor of Phylosophy, Stockholm University, Stockholm.

Rehnman, J., & Herlitz, A. (2007). Women remember more faces than men do. *Acta Psychologica, 124*, 344-355.

Rhodes, G., Brake, S., Taylor, K., & Tan, S. (1989). Expertise and configural coding in face recognition. *British Journal of Psychology, 80*, 313-331.

Roark, D. A., Barrett, S. E., Spence, M., Abdi, H., & O'Toole, A. J. (2003). Memory for moving faces: Psychological and neural perspectives on the role of motion in face recognition. *Behavioral and Cognitive Neuroscience Reviews, 2*(1), 15-46.

Robert, G., & Hockey, J. (1986). *Changes in operator efficiency as a function of environmental stress, fatigue and circadian rythmns* (Vol. 2). Toronto, Cananda: John Wiley & Sons.

Rossion, B. (2002). Is sex categorization from faces really parallel to face recognition? *Visual Cognition, 9*(8), 1003–1020.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*(2), 252-257

Sadr, J., Jarudi, I., & Sinha, P. (2003). The role of eyebrows in face recognition. *Perception, 32*, 285-293.

Salthouse, T. (2004). What and when of cognitive aging. *Current Directions in Psychological Science, 13*, 140-144.

School of Natural Sciences (Psychology). (accessed 2008). The psychological image collection at Stirling (PICS). Retrieved 2008, from University of Stirling Psychology Department http://pics.psych.stir.ac.uk/

Schretlen, D. J., Pearlson, G. D., Anthony, J. C., & Yates, K. O. (2001). Determinants of Benton Facial Recognition Test performance in normal adults. *Neuropsychology, 15*(3), 405-410.

Schwaninger, A., Carbon, C.-C., & Leder, H. (2003). Expert face processing: Specialization and constraints. In G. Schwarzer & H. Leder (Eds.), *Development of face processing* (pp. 81-97). Göttingen: Hogrefe.

Schwaninger, A., Lobmaier, J. S., Wallraven, C., & Collishaw, S. (2009). Two routes to face perception: Evidence from psychophysics and computational modeling. *Cognitive Science.*

Schweitzer, N. J., & Saks, M. J. (2007). The CSI Effect: Popular Fiction About Forensic Science Affects the Public's Expectations About Real Forensic Science. *Jurimetrics, 47*, 357-364

Semmler, C., Ma-Wyatt, A., Heyer, R., & Macleod, V. (2012, 23-27 September). *The impact of individual differences and eye movements on facial comparison performance.* Paper presented at the International Symposium of the Forensic Sciences (ANZFSS), Hobart.

Shepherd, J. (1981). Social factors in face recognition. In G. Davies, H. Ellis & J. Shepherd (Eds.), *Perceiving and Remembering Faces* (pp. 55-80). San Diego: Academic Press.

Sinha, P., Balas, B. J., Ostrovsky, Y., & Russell, R. (2006a). Face recognition by humans. In W. Zhao & R. Chellappa (Eds.), *Face Recognition: Advanced Modeling and Methods*: Academic Press.

Sinha, P., Balas, B. J., Ostrovsky, Y., & Russell, R. (2006b, November 2006). *Face recognition by humans: Nineteen results all computer vision researchers should know about.* Paper presented at the Proceedings of the IEEE.

Slone, A. E., Brigham, J. C., & Meissner, C. A. (2000). Social and cognitive factors affecting the own-race bias in whites. *Basic & Applied Social Psychology, 22*(2), 71-84.

Spacek, L. (last updated 2008). University of Essex face recognition database, from http://cswww.essex.ac.uk/mv/allfaces/index.html

Spaun, N. A. (2009, June 2 - 5). *Facial comparisons by subject matter experts: Their role in biometrics and their training.* Paper presented at the International Conference on Biometrics, Alghero, Italy.

Staal, M. A. (2004). Stress, cognition, and human performance: A literature review and conceptual framework. California: Moffett Field: Ames Research Centre.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149.

Sunde, J., Butavicius, M., Graves, I., Hemming, D., Ivancevic, V., Johnson, R., . . . Meaney, K. (2003, 2-5 July). *A methodology for evaluating the operational effectiveness of facial recognition systems.* Paper presented at the the 4th EUROSIP conference focused on Video/Image Processing and Multimedia Communications, Zagreb, Croatia.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition *The Quarterly Journal of Experimental Psychology, 46*(2), 225-245

Tarr, M. J., Kersten, D., & Bulthoff, H. H. (1998). Why the visual recognition system might encode the effects of illumination. *Vision Research 38*, 2259–2275.

Tarres, F., & Rama, A. (2005). GTAV Face Database available, from http://gps-tsc.upc.es/GTAV/ResearchAreas/UPCFaceDatabase/GTAVFaceDatabase.htm

Tay, L. (2010). Gold Coast Airport deploys SmartGate, *Itnews*. Retrieved from http://www.itnews.com.au/News/171566,gold-coast-airport-deploys-smartgate.aspx

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception, 9*, 483-484.

Thornton, I. M., & Kourtzi, Z. (2002). A matching advantage for dynamic human faces. *Perception, 31*(1), 113-132.

U.S. Department of Homeland Security. (2011). *e-Passports*. Retrieved from http://www.dhs.gov/files/crossingborders/gc_1161636133959.shtm

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43 A*, 161-204.

Valentine, T., Chiroro, P., & Dixon, R. (1995). *An account of the own-race bias and the contact hypothesis in terms of a face space model of face recognition*. London: Routledge.

Vast, R. (2004). *Face matching performance under time pressure and additional task demands.* Master of Psychology (Organisational and Human Factors), The University of Adelaide, Adelaide, SA.

Vast, R., & Butavicius, M. (2005). *A literature review of face recognition for access control: Human versus machine solutions*. (DSTO-TR-1747). Edinburgh, SA: Defence Science and Technology Organisation.

Walker, P. M., & Hewstone, M. (2006). A perceptual discrimination investigation of the own-race effect and intergroup experience. *Applied Cognitive Psychology, 20*, 461-475.

Walker, P. M., & Hewstone, M. (2008). The influence of social factors and implicit racial bias on a generalized own-race effect. *Applied Cognitive Psychology, 22*, 441-453.

Walker, P. M., & Tanaka, J. W. (2003). An encoding advantage for own-race versus other-race faces. *Perception, 32*, 1117-1125.

Wayman, J. (1999). Technical testing and evaluation of biometric identification devices. In A. K. Jain, R. Bolle & S. Pankanti (Eds.), *Biometrics: Personal Identification in Networked Society*: Kluwer Academic Press.

Wayman, J. (2001). Fundamentals of biometric authentication technologies. *International Journal of Image and Graphics, 1*(1), 93-113.

Wayman, J. (2008). *Facial recognition from e-Passports: Australian Customs SmartGate*. Paper presented at the ROBUST 2008 Conference on Biometrics, Honolulu.

Wayman, J., Jain, A. K., Maltoni, D., & Maio, D. (2005). An introduction to biometrics authentication systems. In J. Wayman, A. Jain, D. Maltoni & D. Maio (Eds.), *Biometric Systems: Technology, Design and Performance Evaluation*: Springer.

Weber, M. (1999). Caltech Faces: Frontal face dataset, from http://www.vision.caltech.edu/html-files/archive.html

Weyrauch, B., Huang, J., Heisele, B., & Blanz, V. (2004). *Component-based face recognition with 3D morphable models.* Paper presented at the First IEEE Workshop on Face Processing in Video, Washington, D.C.

Wiese, H., Schweinberger, S. R., & Hansen, K. (2008). The age of the beholder: ERP evidence of an own-age bias in face memory. *Neuropsychologia, 46*, 2973-2985.

Wilkinson, C., & Evans, R. (2009). Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science and Justice, 49*, 191-196.

WISE (Web Interface for Statistics Education). Signal Detection Theory tutorial Retrieved 20h December 2010, from http://wise.cgu.edu/sdtmod/index.asp

Woodhead, M. M., Baddeley, A. D., & Simmonds, D. C. V. (1979). On training people how to recognize faces. *Ergonomics, 22*(3), 333 - 343

Woodward Jr., J. D., Horn, C., Gatune, J., & Thomas, A. (2003). Biometrics: A look at facial recognition. Santa Monica: RAND Public Safety and Justice.

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta Psychologica, 114*, 101-114.

Wright, D. B., & Stroud, J. N. (2002). Age differences in lineup identification accuracy: People are better with their own age. *Law & Human Behavior, 26*, 641-654.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception, 16*, 747-759.

Young, S. G., Bernstein, M. J., & Hugenberg, K. (2010). When do own-group biases in face recognition occur? Encoding versus post-encoding. *Social Cognition, 28*(2), 240-250.

Zhao, W., Chellappa, R., Phillips, J. P., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys, 35*(4), 399-458.

# *Appendices*

## *Appendix A* *Image Preparation: Target Participant Information Sheet*

### BIOMETRIC PERFORMANCE TESTING

**Brief Description of the Study:**
This study aims to examine biometrics performance. To be able to assess this performance, we are collecting photographs and videos to generate imagery for use in biometric trials.

**Your Part in this Study:**
Participation in this study is entirely voluntary. You may choose not to participate and are able to withdraw at any time. You participation will involve the following:

1. You will have a series of images taken by DSTO staff (both still and video, inside and outside).
2. You will also be asked to look through a database of images to find a photograph of a person who you believe looks the most similar to you. This will be of relevance to another trial that we will be conducted later this year.

This should not take longer than 30 minutes.

**NOTE:** Employees of the Defence Department are considered 'on duty' during their participation.

**Risks of Participating:**
Health or well being risks as a result of participating in this research are consistent with working in an office environment. Before the commencement of the experiment, participants will be given a short occupational, health and safety brief in which they will be shown the emergency exits.

**Statement of Privacy:**
The facial images, video recordings, and any other information obtained from your involvement in this study will be treated in the strictest confidence. Your personal particulars will not be linked with this data in any form. All collected data, including participants' images, video recordings and ID cards will be stored in accordance with the National Security classification of RESTRICTED. The data will be stored in the form of a database and used only for further testing within the National Security Systems Analysis. If you wish to withdraw your images and recordings from this database, you can advise the experimenter at any stage after the initial testing. Furthermore, once the data is no longer required it will be destroyed as classified waste.

If you are willing to participate, please sign the Consent Form. Should you have any questions or concerns, feel free to contact us using the details provided below.

**Investigators contact information:**
Should you have any complaints or concerns about the manner in which this project is conducted, please do not hesitate to contact the researchers in person, or you may prefer to contact the Australian Defence Human Research Ethics Committee.

*Dragana Calic*
Psychology PhD Candidate
Ph: 8259 4191
Email:
dragana.calic@dsto.defence.gov.au

Executive Secretary
Australian Defence Human Research Ethics Committee
Department of Defence
CANBERRA ACT 2600
Ph: (02) 6266 3837 Fax: (02) 6266 4982
Email: ADHREC@defence.gov.au

## DSTO PROCEDURES FOR RESEARCH INVOLVING HUMAN PARTICIPANTS

Thank you for taking part in Defence Research. Your involvement is much appreciated. This pamphlet explains your rights as a volunteer.

ADF Pamphlet 1.2.5.3 sets out guidelines for Defence research involving human beings. The Australian Defence Human Research Ethics Committee (ADHREC) exists to review Defence proposals for scientific research involving humans. The ADFP 1.2.5.3 does not require ADHREC to review all proposals for experimentation involving human participants. DSTO follows an internal review process to ensure its research conforms to ADF 1.2.5.3 and is reviewed by ADHREC when appropriate.

### What is ADHREC?

- ADHREC is the Australian Defence Human Research Ethics Committee. It was established in 1988, as the Australian Defence Medical Ethics Committee (ADMEC), to make sure that Defence complied with accepted guidelines for research involving human beings.
- After World War II, there was concern around the world about human experimentation. The Declaration of Helsinki was made in 1964, which provided the basic principles to be followed wherever humans were used in research projects.
- The National Health & Medical Research Council in Australia (NHMRC) published a set of guidelines in 1982 for how human research should be carried out.
- ADHREC follows both the Declaration of Helsinki and the NHMRC Guidelines.

### DSTO process

- DSTO has developed an approval process for ensuring that research involving humans complies with the comprehensive guidelines provided in Australian Defence Force Publication (ADFP) 1.2.5.3 entitled “Health and Human Performance Research in Defence – Manual for Researchers”.
- If you are told that the project has DSTO approval, what that means is that the DSTO S&T Activity Review Team has reviewed the research proposal and has agreed that, in accordance with ADFP 1.2.5.3 paragraph 1.13, ethical clearance through ADHREC is not required. In addition, a Research Leader has reviewed the research proposal and is satisfied that safety and ethical issues relating to informed consent, confidentiality and security of data have been addressed.
- DSTO approval does not imply any obligation on commanders to order or encourage their military personnel to participate, or to release military personnel from their usual workplace to participate. Obviously, the use of any particular military personnel must have clearance from their commanders but commanders should not use DSTO approval to pressure military personnel into volunteering.

### Voluntary participation

- As you are a volunteer for this research project, you are under no obligation to participate or continue to participate. You may withdraw from the project at any time without detriment to your military career or to your medical care.
- At no time must you feel pressured to participate or to continue if you do not wish to do so.
- If you do not wish to continue, it would be useful to the researcher to know why, but you are under no obligation to give reasons for not wanting to continue.

**Informed consent**

- Before commencing the project you will have been given an information sheet which explains the project, your role in it and any risks to which you may be exposed.
- You must be sure that you understand the information given to you and that you ask the researchers about anything of which you are not sure.
- If you are satisfied that you understand the information sheet and agree to participate, you should initial every page of the information sheet and keep a copy.
- Before you participate in the project you should also have been given a consent form to sign. You must be happy that the consent form is easy to understand and spells out to what you are agreeing. Again, you should keep a copy of the signed consent form.

**Complaints**

- If at any time during your participation in the project you are worried about how the project is being run or how you are being treated, then you should speak to the researchers.

  Contact details:<br>
  **Ms Dragana Calic**<br>
  **Division: Land Operations Division**<br>
  **Address: Bay 2 Building 75**<br>
  **Telephone number: (08) 8259 4191**<br>
  **Email address: dragana.calic@dsto.defence.gov.au**

- If you don't feel comfortable doing this, you can contact the Executive Secretary of ADHREC.

  Contact details:<br>
  **Executive Secretary**<br>
  **Australian Defence Human Research Ethics Committee**<br>
  **CP2-7-66**<br>
  **Department of Defence**<br>
  **CANBERRA ACT 2600**<br>
  **Ph: 02 62663837 Fax: 02 62664982**<br>
  **E-mail: ADHREC@defence.gov.au**

**More information**

- If you would like to read more about ADHREC, you can look up the following references on the Defence Manager's Toolbox or on DEFWEB
  - DI(G)ADMIN 24-3 *Function, Structure and Procedures for Obtaining Clearance for Research from Australian Defence Medical Ethics Committee* (or as amended)
  - HPD 205 *Australian Defence Medical Ethics Committee* (or as amended)
  - ADFP733 *Health and Human Performance Research in the Australian Defence Organisation – Manual for Researchers*
- Or, visit the ADHREC web site at http://defweb2.cbr.defence.gov.au/dpedhs/default.htm

*Appendix C Consent Form*

## CONSENT FORM

I…………………………………………………………………..give my consent to participate in the project mentioned above on the follow basis:

- I acknowledge that I have read the attached information sheet and that I have been given a copy of this information sheet and the *Guidelines for Volunteers* for my records.
- I have had explained to me the aims of this research project, how it will be conducted and my role in it to my satisfaction.
- It has been explained to me that my involvement in this project is voluntary and will not be detrimental to me.
- I have been informed that, while information gained during the study may be published, the information I provide will be kept private and I will not be individually identified.
- I understand that facial imagery obtained during the trial will not be published without prior written authorisation/approval, but may be stored for further analyses.
- I understand that I am under no obligation to take part in this study and that I am free to withdraw from the project at any time without repercussions.
- The option to be given a copy of the consent form at a later time following the trial has been offered to me.

**……………………………………………………………………………………………**

**(Signature of Volunteer) (Name in Full) (Date)**

***Image Preparation: Target Participant Record Sheet***

This sheet will be used as a checklist for your participation today.

TP number: □

| | |
|---|---|
| **Date** | |
| **Time started** | |
| **Time completed** | |

**Photographic and Video Sessions:**

| **Indoor Portraiture** | | **Video Recordings** | |
|---|---|---|---|
| □ | Still Image | □ | Video |

Please record the following:

Gender □Male □Female

Age______Ethnicity (optional) ___________

Glasses □Yes □No

Please look through the provided booklet of photographs and try to find which face you think most looks like you.

Please record the number of the face that you think most looks like you.

______________________

| How confident are you that you would be able to use this photo as form of an ID? | | | | |
|---|---|---|---|---|
| **Very Confident**<br>□ | **Confident**<br>□ | **Neutral**<br>□ | **Not Confident**<br>□ | **Not Confident at all**<br>□ |

# Human Face Recognition in Photos and Videos

**Brief Description of the Study:**
This study aims to examine the way that humans perform facial recognition. It will assess humans' ability to verify an individual's identity and confirm that they are who they claim to be. This will be done by Human Operators (i.e., you!!!) by comparing a video to a photograph and deciding if the person presented in the video and the person in the photograph are of the same individual.

**Your Part in the Study:**
Participation in this study is entirely voluntary. You may choose not to participate and are able to withdraw at any time.

You participation will involve sitting down in a lab and looking at a series of videos and photographs. Your task will be to make a decision about if the person in the video and the person in the photo are of the same person.

**Risks of Participating:**
There are no risks to your health or well being as a result of participating in this study. Any occupational health and safety issues will be identified on site and appropriate measures will be taken to control risks to participants. Participation is purely voluntary and you are free to withdraw at any time.

**Statement of Privacy:**
The information obtained from your involvement in this study will be treated in the strictest confidence. Your personal particulars will not be linked with this data in any form. Furthermore, once the data is no longer required it will be destroyed as classified waste. You will also have the opportunity to receive a summary of the research's findings.

Should you have any questions or concerns, feel free to contact us using the details provided below.

**Investigators Contact Information:**
Should you have any queries, complaints or concerns about the manner in which this project is conducted, please do not hesitate to contact the researchers in person, or you may prefer to contact the convener of the School of Psychology Human Ethics Subcommittee.

| | |
|---|---|
| *Dragana Calic*<br>*Psychology PhD Candidate*<br>Ph. 0401 688 245<br>dragana.calic@adelaide.edu.au | |
| *Dr. Anna Ma-Wyatt*<br>*Supervisor*<br>Ph. (08) 8303 5660<br>anna.mawyatt@adelaide.edu.au | *Dr. Paul Delfabbro*<br>*Convenor of the School of Psychology Human Ethics Subcommittee*<br>Ph. (08) 8303 4936<br>paul.delfabbro@psychology.adelaide.edu.au |

## *Human Operator Instructions*

*Your role today is to undertake a facial recognition task to verify an individuals' identity.*

*You will be approached by a number of Target Participants who will present to you an envelope containing an ID Card.*
*Your task will be to:*

- *record the Target Participant's ID number in this booklet (which is indicated on the envelope and ID Card);*
- *take the ID Card out of the envelope and assess if the person in front of you is the same as the person in the photograph on the ID Card;*
- *record your decision in this booklet;*
- *indicate how confident you are in your decision;*
- *put the ID Card back into the envelope;*
- *cross out your cubicle number in the table on the back of the envelope; and*
- *return the envelope to the Target Participant.*

*In making your decision you should ensure that:*

- *the Target Participant does* ***NOT*** *see their ID card OR your decision response;*
- *you do* ***NOT*** *ask the Target Participant any questions about their appearance or any other questions that may help you make your decision;*
- *you make your decisions as quickly and as accurately as you can; and*
- *you do* ***NOT*** *discuss your decisions with any other Operators.*

*When you complete this process with one Target Participant the next person will then approach you and you will repeat the same process.*

*Before starting the facial recognition task you will also be asked to complete a set of five short perceptual tests that we will be using as part of this study.*

*You will then complete some demographic questions in this booklet.*

*At the end of the day, once all facial recognition tasks are completed, we will also ask you to complete the last section "After Trial Questions". Those questions relate to how you have gone about conducting the facial recognition task.*

*Please rest assured that your responses will not be individually reported.*

*If you have any questions about the procedure and need some clarification please do not hesitate to ask at any time.*

*Once again, thank you for your participation.*

## *Demographic Questions*

**Age (Years)** □□

**Gender**

□ Male

□ Female

**Ethnicity**

□ White Caucasian

□ Asian

□ Other. Please specify □

**Do you wear glasses/contact lenses to correct your vision?**

□ Yes (all the time)

□ Yes (sometimes)

□ No

**Are you wearing glasses now?**

□ Yes

□ No

**Highest level of education attended/ing (tick the highest)**

□ Secondary Schooling

□ Post Secondary (e.g., TAFE)

□ University Bachelor Level

□ University Honours Level or Higher

□ Other. Please Specify □

**Which organisation are you from?**

□ Agency A

□ Agency B

□ Agency C

□ Agency D

□ The University of Adelaide

## *Face Recognition Training and Experience Related Questions*

### *Training*

**Have you ever received training (e.g., a course or on-the-job) related to how to identify/verify people?**

□ Yes
□ No

**If YES, please provide some details about this training.**

| | |
|---|---|
| **Type of Training** | |
| **Training Description** | |
| **Length of Training** | |
| **Year of Training** | |
| **Training Provider** | |

### *Experience*

**Do you work (or have you ever worked) in an area where you were required to perform face identification/verification tasks (e.g., recognising faces)?**

□ Yes

□ No

**If YES, please provide some details about this.**

| | |
|---|---|
| **How often do/did you work in this role?** | |
| **Do you undertake facial recognition tasks as part of your current role?** | |
| **In what context?** | |
| **How long have/did you work/worked in this role for?** | |
| **Further details as required** | |

# SAMPLE

## ID Card

NOTE:
This figure/table/image has been removed
to comply with copyright regulations.
It is included in the print copy of the thesis
held by the University of Adelaide Library.

## Back of the Envelope

| Check In Time<br>__________________ | | | Check Out Time<br>______________ | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 |
| 25 | 26 | 27 | 28 | 29 | 30 |
| 31 | 32 | 33 | 34 | 35 | 36 |
| 37 | 38 | 39 | 40 | 41 | 42 |

«Participant_ID_Number_»4794

Please complete the following information for each Target Participant.

| | | | | | | |
|---|---|---|---|---|---|---|
| Target Participant Number | **T** | **P** | **4** | **7** | **9** | **4** |

Does the photograph on the ID Card match the presenting individual?

**YES** ☒ **NO** □

How confident are you in your decision?

(0% indicates not confident at all and 100% indicates extremely confident)

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

0% 50% 100%

Please complete the following information for each Target Participant.

| | | | | | | |
|---|---|---|---|---|---|---|
| Target Participant Number | **T** | **P** | | | | |

Does the photograph on the ID Card match the presenting individual?

**YES** □ **NO** □

How confident are you in your decision?

(0% indicates not confident at all and 100% indicates extremely confident)

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

0% 50% 100%

## *Appendix G* *Experiment 2: Target Participant Information Sheet*

# Human Face Recognition Performance: Live Evaluation

### Brief Description of the Study:

In this part of the research we are assessing human facial recognition performance. This will involve you presenting a pre-prepared ID card to each operator in turn who will attempt to determine whether this is a true photograph of you (or not).

### Your Part in this Study:

Participation in this study is entirely voluntary. You may choose not to participate and are able to withdraw at any time. You participation will involve the following:

1. You will be asked to present a pre-prepared ID card to each operator who will attempt to determine whether this is a true photograph of you (or not).
2. We will also acquire still and moving imagery of you (as done earlier this year).

### Risks of Participating:

Health or well being risks as a result of participating in this research are consistent with working in an office environment. Before the commencement of the experiment, participants will be given a short occupational, health and safety brief in which they will be shown the emergency exits.

### Statement of Privacy:

The facial images, video recordings, and any other information obtained from your involvement in this study will be treated in the strictest confidence. Your personal particulars will not be linked with this data in any form. The data will be stored in the form of a database and used only for further testing within the National Security Systems Analysis task. If you wish to withdraw your images and recordings from this database, you can advise the experimenter at any stage after the initial testing. Furthermore, once the data is no longer required it will be destroyed as classified waste.

If you are willing to participate, please sign the Consent Form. Should you have any questions or concerns, feel free to contact us using the details provided below.

### Investigators Contact Information:

Should you have any complaints or concerns about the manner in which this project is conducted, please do not hesitate to contact the researchers in person, or you may prefer to contact the Australian Defence Human Research Ethics Committee.

*Dragana Calic*
Psychology PhD Candidate
Ph:  8259 4191
Email:
dragana.calic@dsto.defence.gov.au

Executive Secretary
Australian Defence Human Research Ethics Committee
Department of Defence
CANBERRA ACT 2600
Ph: (02) 6266 3837    Fax: (02) 6266 4982
Email: ADHREC@defence.gov.au

## Human Face Recognition Performance: Live Evaluation

**Brief Description of the Study:**
This study aims to examine how well humans perform face recognition. It will assess humans' ability to identify individuals and confirm that they are who they claim to be. This will be done by Human Operators (i.e., you!!!) by comparing individuals to a photograph and deciding if they are the same as the individual on the photograph.

**Your Part in the Study:**
Participation in this study is entirely voluntary. You may choose not to participate and are able to withdraw at any time. You participation may be video recorded.
Your task will be to compare an individual (i.e., the target participant) to an ID photograph that they will present to you. You will then record your decision about if the two faces are the same. You will also be asked to record how confident you are about the decision that you have just made. The same process will need to be repeated for all target participants.
Additionally, you will be asked to provide some basic demographic details, complete a set of cognitive and personality tests and participate in a very brief post evaluation survey to gain an insight into how you conduct face recognition tasks.

**Risks of Participating:**
Health or well being risks as a result of participating in this research are consistent with working in an office environment. Before the commencement of the experiment, participants will be given a short occupational, health and safety brief in which they will be shown the emergency exits.

**Statement of Privacy:**
Your decisions, video recordings and any other information obtained from your involvement in this study will be treated in the strictest confidence. Your personal particulars will not be linked with this data in any form. Once the data is no longer required it will be destroyed as classified waste. You will also have the opportunity to receive a summary of the research's findings.

**Investigators Contact Information:**
Should you have any complaints or concerns about the manner in which this project is conducted, please do not hesitate to contact the researchers in person, or you may prefer to contact the Australian Defence Human Research Ethics Committee.

*Dragana Calic*
Psychology PhD Candidate
Ph: (08) 8259 4191
Email:
dragana.calic@dsto.defence.gov.au

Executive Secretary
Australian Defence Human Research Ethics Committee
Department of Defence
CANBERRA ACT 2600
Ph: (02) 6266 3837 Fax: (02) 6266 4982
Email: ADHREC@defence.gov.au

## *Appendix I* *Experiment 2: Trained and Untrained Correlation Analyses*

*Table 31: Pearson's correlations between trained and untrained Human Operator performance and individual differences*

| | ***Trained Human Operators*** | | | ***Untrained Human Operators*** | | |
|---|---|---|---|---|---|---|
| | *Overall Accuracy* | *Hit Rate* | *False Alarm Rate* | *Overall Accuracy* | *Hit Rate* | *False Alarm Rate* |
| **Confidence Ratings** | **.44*** | .19 | .28 | .45 | .60 | **.83*** |
| **Perceptual Speed** | | | | | | |
| Finding A's | .25 | .09 | .38 | .41 | .20 | .33 |
| Identical Pictures | **.43*** | .22 | -.05 | .41 | 45 | .31 |
| Number Comparison | .27 | **.42*** | .00 | .30 | .25 | .47 |
| **Glasgow Face Matching Test** | **.57*** | .29 | .35 | .12 | -.16 | .49 |
| **Rational-Experiential Inventory** | | | | | | |
| Rational Ability | **-.49*** | **-.45*** | -.05 | .35 | -.41 | .15 |
| Rational Engagement | -.29 | **-.41*** | .08 | .41 | -.15 | .60 |
| Experiential Ability | -.28 | -.32 | -.34 | .14 | .17 | .47 |
| Experiential Engagement | -.14 | -.18 | -.24 | -.16 | .48 | -.18 |

* correlation is significant at *p* < .05

** correlation is significant at *p* < .01