Investigating Person Identification in Security Settings with Virtual Reality

A thesis submitted for the degree of PhD in Cognitive Psychology / Neuropsychology in the Faculty of Social Sciences at the University of Kent

Hannah Margaret Tummon

School of Psychology

University of Kent

September 2019

Abstract

Person identification at airports requires the comparison of a passport photograph with its bearer. In psychology, this process is typically studied with static pairs of face photographs that require identity match (same person shown) versus mismatch (two different people) decisions, but this approach provides a limited proxy for studying how environment and social interaction factors affect this task. This thesis explores the feasibility of virtual reality (VR) as a solution to this problem, by examining the identity matching of avatars in a VR airport. In Chapter 2, facial photographs of real people are successfully rendered into VR avatars in a manner that preserves image and identity information (Experiments 1 to 3). Furthermore, identity matching of avatar pairs reflects similar cognitive processes to the matching of face photographs (Experiments 4 and 5), a pattern which holds when assessed in a VR airport (Experiments 6 and 7). Chapter 3 then examines whether a simulation of a passport control task in VR can provide a useful tool for selecting personnel for real-world tasks (Experiment 8). The classification of identity mismatches, the detection of which is of paramount importance in security settings, correlated across conventional laboratory face matching tests and the VR passport control task. Social interaction factors, such as body language, may further influence face matching performance, which was explored in Chapter 4. Whilst performance was unaffected when observers were not instructed explicitly to utilise body language (Experiments 9 and 10), when instructed body language enhanced detection of identity mismatches yet also increased false classification of matches (Experiments 11 to 13). This effect was driven by increased activity levels rather than body language that simply differed from normal behaviour, and occurred independently of individuals' face-matching ability (Experiment 14). This thesis concludes with a summary of how VR can open up many avenues for face-matching research, by facilitating the study of new environment and social interaction factors that may be relevant in real-world operational settings.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Markus Bindemann, for his supportive guidance and advice, and to John Allen for his technical assistance in bringing the virtual reality airport to life. I also thank my family and friends for their support, especially to my parents for their continuous encouragement and Natalie Gentry, Jaimee Mallion and Matt Plummer for all the memories we have made along the way. Finally, to the staff and students within the School of Psychology who have created a fantastic community in which to work during my seven years at the University of Kent.

This research was supported by an Economic and Social Research Council South East Doctoral Training Centre Studentship (no. ES/J500148/1).

Declaration

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

Hannah M. Tummon

Publications

Chapter 2 of this thesis has been published

Tummon, H. M., Allen, J., & Bindemann, M. (2019). Facial identification at a virtual reality airport. *i-Perception*, *10*(4), 1-26. doi:10.1177/2041669519863077

Chapter 4 has been submitted for publication

Tummon, H. M., Allen, J., & Bindemann, M. (submitted). Body language influences person identification at a virtual reality airport. *Journal of Experimental Psychology: Human Perception and Performance*

TABLE OF CONTENTS

Chapter 1:
General Introduction

1.1 Introduction

Security settings critically rely on the accurate identification of a large volume of people. This often involves comparing a photo from an identity document, such as a passport, to the document's bearer and determining if they are an identity match or mismatch. This is an example of forensic face matching, a routinely performed task at international borders. Face matching performed in high stakes environments almost always consists of unfamiliar faces and incorrect judgements could potentially have serious consequences, for example a terrorist gaining access to a country using a stolen passport. It is therefore important to study the classification of identity mismatches to simulate this real-world problem of impostors, who travel on legitimate identity documents of someone that is similar in facial appearance to avoid detection at passport control (Bindemann, Fysh, Cross, & Watts, 2016; Meissner, Susa, & Ross, 2013; Susa, Michael, Dessenberger, & Meissner, 2019). Contemporary passports are difficult to forge due to the multitude of security checks they contain and so criminal organisations are more likely to stockpile stolen genuine passports to be issued to other persons attempting to conceal their identity (Hoepman, Hubbers, Jacobs, Oostdijk, & Wichers Schreur, 2006; Schouten & Jacobs, 2009). Therefore, if the authenticity of the passport is not questionable, passport control officers must detect their fraudulent use through face matching alone.

In psychology, this task has been studied extensively as unfamiliar face matching (Fysh & Bindemann, 2017a; Jenkins & Burton, 2008a, 2011; Robertson, Middleton, & Burton, 2015). In experiments in this field, observers are typically required to compare pairs of face photos, which are presented in isolation on blank backgrounds, and decide whether these depict the same person or two different people. However, the complexity of operational settings cannot be adequately captured through such methods; passport control officers compare a live person with their purported passport photo taken up to ten years previously, making decisions as quickly and accurately as possible in order to clear the growing queues of passengers they face.

Furthermore, these paradigms provide a limited proxy for studying how the environment and social interaction might affect this task. In real-life environments, passport officers may, for example, resort to non-facial cues, such as body language, to support identification decisions (Rice, Phillips, Natu, An, & O'Toole, 2013; Rice, Phillips, & O'Toole, 2013).

This thesis explores the use of virtual reality (VR) as a means of simulating a passport control environment in order to investigate the impact of such real-world factors on face-matching decisions. This has not previously been attempted and so the first stage was to investigate the use of virtual figures (avatars) as a replacement for real people; this was to establish whether avatar faces are processed in the same way as real faces and therefore validates the use of VR for face-matching experiments. Second, it was investigated whether face-matching tasks could predict performance in a passport control scenario simulated in VR to explore the potential of such methods as an assessment tool for selection of personnel at border security. Finally, the impact of body language on face matching decisions in an airport environment was examined through the manipulation of avatar animation, as an example of how VR can be used to investigate the influence of real-world factors. This chapter reviews what is known so far about unfamiliar face matching and its difficulties, factors which influence face matching in the real world, and the potential for investigation in the virtual world.

1.2 Matching unfamiliar faces

Face matching involves the comparison of facial images and determining whether or not they have the same identity. A combination of cognitive processes may be used whilst completing this task. Firstly, *perception* is required to detect the faces to be matched within a scene. Once the to-be-matched faces have been selected, *attention* is then directed towards meaningful facial cues which can be used to establish the identity of the person. This has been

demonstrated with eye-tracking studies showing that when observers' viewing is constrained to a single location, subsequent recognition accuracy decreases (Henderson, Williams, & Falk, 2005). Finally, *evaluation* processes are required to decide whether the two faces are sufficiently similar to be a match or have enough distinctive characteristics which suggest they are different people. When faces are familiar to the observer, matching is an easy task to perform with near-perfect accuracy even if the image quality is substandard (Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999; Megreya & Burton, 2007). Since the matching of familiar faces is performed with ease, there is no expectation for unfamiliar face matching to be difficult (Burton, Kramer, Ritchie, & Jenkins, 2016; Ritchie et al., 2015). In fact, performance is highly error-prone, with matching accuracy falling by approximately 20% when the faces are unfamiliar to the observer (Fysh & Bindemann, 2017a; Hancock, Bruce, & Burton, 2000; Jenkins & Burton, 2011; Johnston & Edmonds, 2009; Robertson, Middleton, & Burton, 2015). This difficulty for matching unfamiliar faces has been demonstrated by numerous experiments conducted both in the laboratory and in the field. How accurately this task is performed can be substantially impacted by the characteristics of the stimuli as well as the content of the images themselves, such as the relative importance of internal and external features and their variability across images.

1.2.1 Stimuli characteristics

Controlled laboratory experiments have been employed to study how the characteristics of facial stimuli affect face matching. Factors which can have a substantial impact on the observer's ability to process the face include image quality (Bindemann, Attard, Leach, & Johnston, 2013; Lander, Bruce, & Hill, 2001), variation in viewpoint (Estudillo & Bindemann, 2014), camera distance (Noyes & Jenkins, 2017) and lighting (Hill & Bruce, 1996). These

factors could have a significant influence on face matching at passport control. Electronic facial recognition gates (eGates) operate by scanning the passport bearer's face using a camera and comparing it to the image stored on the passport (Vine, 2012). Passport officers monitoring the eGate screens view larger scale versions these images, which could result in pixelation, and the eGate image is likely to be taken under different lighting conditions and from a different distance to the original passport image. Pixelation significantly impairs face processing both for matching and recognition tasks, yet the impact can be lessened through the reduction of image size and motion (Bindemann et al., 2013; Lander et al., 2001). Furthermore, when stimuli characteristics are variable across images to be matched, their impact can be considerable; whilst matching is error-prone across standardised and optimal conditions (e.g., Burton, White, & McNeill, 2010; Megreya & Burton, 2006a), when the images are taken from alternative viewpoints (Estudillo & Bindemann, 2014), lighting conditions (Hill & Bruce, 1996), or camera distances (Noyes & Jenkins, 2017) the task becomes even more difficult and the faces are more likely to be perceived as different people. These factors may also interact, for example different viewpoints of faces are more accurately matched when lit from above (Hill & Bruce, 1996).

1.2.2 Internal versus external feature matching

Whilst the stimuli characteristics have a considerable effect, it is also important to consider how facial features may affect face processing. For example, the negative effect of viewpoint changes may be minimised by focusing on the nose and mouth region of the face (Royer et al., 2016), yet the removal of external facial features has also been shown to impair face matching independently of viewpoint (Estudillo & Bindemann, 2014). External features, such as hair style and additional paraphernalia (e.g., glasses, headwear) are easily altered and

can drastically alter the appearance of a face, thus are misleading cues to identity (Kemp, Caon, Howard, & Brooks, 2016).

Unfamiliar face recognition and matching can be prone to relying on external features, resulting in identification errors. When faces are unfamiliar, observers match external and internal facial features with similar accuracy, with a possible advantage for external features (Bruce et al., 1999; Megreya, Bindemann, & Havard, 2011; Nachson & Shechory, 2002; Want, Pascalis, Coleman, & Blades, 2003). With increasing familiarity, faces are recognised faster and more accurately by their internal features (Bonner, Burton, & Bruce, 2003; Clutterbuck & Johnston, 2002; Ellis, Shepherd, & Davies, 1979; Want et al., 2003; Young, Hay, McWeeny, Flude, & Ellis, 1985). This dissociation of feature salience is highlighted by the construction of facial composites, whereby external features are constructed more accurately than internal features, leading to low naming rates by those familiar with the target (Frowd, Bruce, McIntyre, & Hancock, 2007). Furthermore, perceptual expertise by those who typically observe women wearing headscarves (Megreya & Bindemann, 2009; Megreya, Memon, & Havard, 2012; Wang et al., 2015) also demonstrates that identifying faces from internal features alone may also be influenced by a cultural bias. Passport control officers frequently match photographic documentation to persons of different cultural backgrounds to themselves, and so such biases may result in identification errors in this context.

1.2.3 Within-face variability

Faces become familiar as the invariant characteristics of the face, such as the internal features, can successfully be attributed to an identity. This may be explained as the development of stable representations, or prototypes, for identities following multiple encounters with their face (Bruce, 1994; Burton, Jenkins, Hancock, & White, 2005; Burton,

Jenkins, & Schweinberger, 2011). Observers learn over time how a particular face varies under different conditions, and a new face image can be compared to existing representations to assess whether it is already known. For unfamiliar faces, no such representations exist, therefore when comparing images of unfamiliar people it is more challenging to assess if they have the same identity. To compound this difficulty, face images of the same person can vary widely in their appearance; this within-face variability can be caused by physiological changes (e.g., ageing, health), the changing of external features (e.g., hairstyle), the inclusion of paraphernalia (e.g., glasses), and facial movement. As a demonstration of this, Kramer and Ritchie (2016) presented observers with two faces which may have been wearing glasses. When only one of the faces wore glasses, observers were more likely to report an identity mismatch and were less accurate than when both or neither had glasses. Furthermore, Armann, Jenkins and Burton (2016) show that observers familiar with a target identity successfully report whether a new image depicts a previously seen individual, since they understand how that person varies across images, whereas unfamiliar observers are more accurate at reporting if they have seen a specific image.

In order to fully understand the process of unfamiliar face recognition and matching, it is important to take this within-face variability into account when conducting research (Burton, 2013). Photos taken on the same day with the same image capture device do not reflect the natural variability of faces. This has been highlighted by card sorting tasks in which ambient images of the same identity have to be grouped together. When presented with images belonging to two identities to sort into piles, with no direction given regarding the correct number, those familiar with the individuals successfully sort the cards into two piles with near perfection (Jenkins, White, Van Montfort, & Burton, 2011); on the contrary, those unfamiliar with the identities sort the cards into many more piles (on average 6-8; Andrews, Jenkins, Cursiter, & Burton, 2015; Jenkins, White, et al., 2011). Errors rarely occurred whereby images

of different people were thought to have the same identity, suggesting the difficulty lay in telling people *together*. However, these identity judgements are seemingly fragile; Sauerland et al. (2016) asked participants to sort a group of 50 faces by identity before presenting them with a pair of faces and asking them to explain their reasoning for putting the faces into the same or different group. On manipulated trials the experimenter gave false information regarding the decision made, for example asking why two faces were reported to be the same when in fact the participant had placed them in the different group. Only 21% of the manipulated trials were detected and participants freely gave justifications for decisions they had not made.

1.2.4 Multiple images and face averages

It is clear that ambient images are difficult to match for unfamiliar people; compared to same-day low variance images, errors in matching images taken months apart are approximately 20% higher (Megreya, Sandford, & Burton, 2013). This suggests that current methods of verifying identity through photographic documentation, such as with passports which could be valid for up to 10 years, are flawed by observers' inability to recognise how people vary across different instances. The use of multiple images may therefore assist this task; accuracy can improve from near chance levels to between 80-90% when multiple images known to be of the same person are provided for comparison to the target (Bindemann & Sandford, 2011; Dowsett, Sandford, & Burton, 2016). The amount of variation between the images is of particular importance in order to be beneficial. High variability images better capture the idiosyncratic ways in which people vary, thus providing a better overall representation of such persons (Burton et al., 2016; Menon, White, & Kemp, 2015; Ritchie & Burton, 2017). Repeated exposure to the same images (Murphy, Ipser, Gaigg, & Cook, 2015)

and abnormal variation, such as negative or contrast-reversed images (Kramer, Jenkins, Young, & Burton, 2016), provide no benefit thus highlighting the need for natural variability to enhance recognition and matching performance.

An alternative means of capturing within-face variability is to average face images to filter out image-specific properties, such as lighting and pose, and retain the diagnostic characteristics of the face (Burton et al., 2005, 2011). The use of an average face image outperforms a single image for matching both familiar and unfamiliar faces (White, Burton, Jenkins, & Kemp, 2014), with improving accuracy as the number of images contributing to the average increases (Burton et al., 2005; Jenkins, Burton, & White, 2006). Furthermore, when learning new faces with multiple exemplars, observers also report having seen a novel average image of the exemplars in a subsequent recall task, further suggesting average images successfully capture the inherent characteristics of faces (Kramer, Ritchie, & Burton, 2015; Neumann, Schweinberger, & Burton, 2013). The use of face averages therefore has potential applications in identity verification settings, such as automated face recognition at security control points (Jenkins & Burton, 2008b) and smartphone authentication (Robertson, Kramer, & Burton, 2015).

1.3 Photo-to-photo face matching

A principal application of face-matching research concerns the identification of individuals in forensic settings, such as imposters at passport control points. This typically involves the comparison of face photos and deciding if the identities match. Photo-to-photo face matching is a surprisingly difficult task for unfamiliar faces (Fysh & Bindemann, 2017a; Hancock et al., 2000; Robertson, Middleton, & Burton, 2015). Since matching familiar faces appears to be trivially easy, unfamiliar face matching is not expected to be difficult, which is

clearly not the case. This expectation is perpetuated by the fact feedback is rarely available in real life when identifying unfamiliar faces (Jenkins & Burton, 2011) and one assumes others are generally as equally competent as oneself (Ritchie et al., 2015). Unfamiliar face matching performance has been examined extensively using a multitude of stimuli, including photo-to-photo matching with an array and pairwise matching tasks.

1.3.1 Matching to an array

The 1-in-10 task devised by Bruce et al. (1999) has been used in multiple studies to investigate factors which contribute to the difficulty of matching unfamiliar faces. This task consists of determining whether a target identity is present (and who it is) or absent from a simultaneously presented 10 person array, as depicted in Figure 1.1.

When viewing conditions are the same for both the target face and those in the array, accuracy is approximately 10% greater than when the viewing conditions are altered (Bruce et al., 1999), for instance when the target has a different facial expression or the photo is taken from a different angle compared to the array images. However, even under the same viewing conditions correct responses are given on only 70% of trials, both for target-present and target-absent arrays (Bruce et al., 1999), an accuracy rate which has been replicated many times with these stimuli (e.g., Megreya & Burton, 2006b, 2007). A left-to-right bias has been demonstrated when making an identification in this lineup task, resulting in more false positive responses on the left side of a target-absent lineup (Megreya, Bindemann, Havard, & Burton, 2012). Accuracy is not substantially improved when the 1-in-10 task is constrained to a forced choice, such as when the target is known to be present (Bruce et al., 1999), nor when the number of distractors in the array is reduced to a 1-in-2 forced choice (Henderson, Bruce, & Burton, 2001). This highlights the challenging nature of this seemingly simple task.

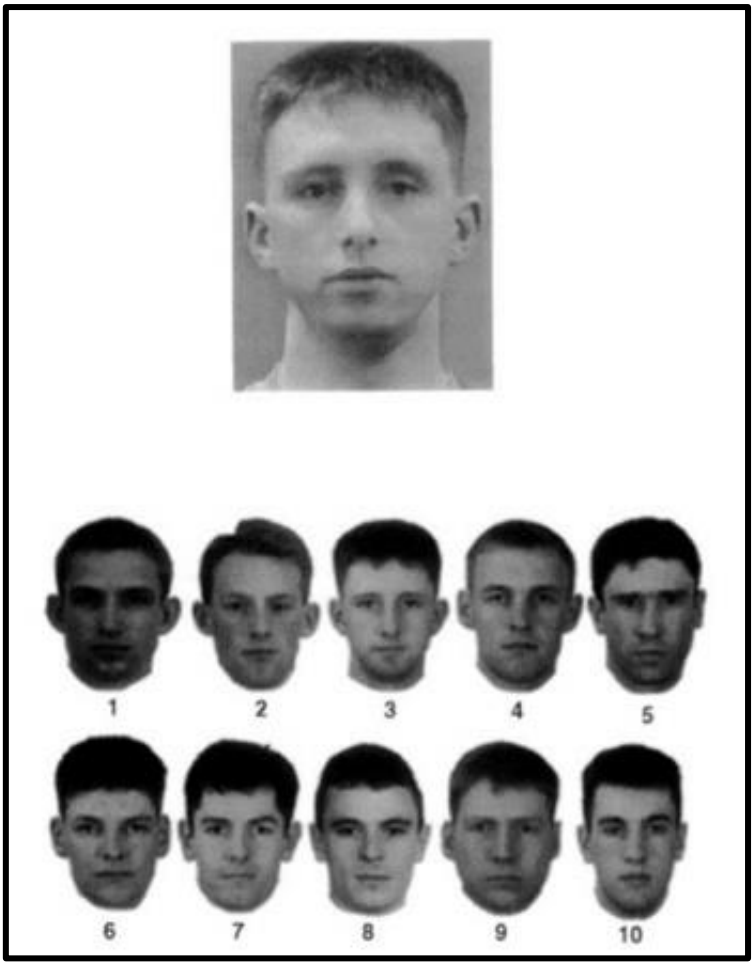

Figure 1.1. Example of the 1-in-10 task of Bruce et al. (1999) in which the person in the top image may or may not match one of the identities in the 10 person array.

This 1-in-10 task has also been used to examine observer characteristics which may influence accuracy. Megreya, White, and Burton (2011) demonstrated an own-race bias in this lineup matching task, with British and Egyptian faces compared by same- and other-race observers. There were universally high error rates, however British observers were more liberal when matching Egyptian faces whilst Egyptian observers were more conservative for British faces. For unfamiliar faces, there is no relationship between hit rate accuracy and false positives (Megreya & Burton, 2007), and observers widely vary in their ability to perform this task, for example between 50-96% accuracy (Megreya & Burton, 2006a).

1.3.2 Pairwise face matching

One explanation why the 1-in-10 task is so difficult is the presence of multiple distractors. Essentially 10 decisions are made per trial, determining if the target is the same or different to each person in the array before resolving on a response. It is clear that with increasing numbers of distractors, performance on this task continues to decline. Accuracy is comparable when observers are tasked with a 1-in-10 or a 2-in-5 lineup, yet with a 2-in-10 task when up to 20 comparisons are made, accuracy falls further still, even when one of the two targets is cued (Bindemann, Sandford, Gillatt, Avetisyan, & Megreya, 2012; Megreya & Burton, 2006b). Pairwise matching somewhat reduces these task demands; only two faces are presented for comparison and the sole decision required is whether or not they have the same identity, such as what would be performed in a passport control context. Despite this simplified task, face matching errors continue to be prevalent.

The Glasgow Face Matching Test (GFMT; Burton et al., 2010) consists of pairs of facial images taken from a frontal view displaying a neutral expression. The two images in a face pair are taken with different cameras and, in the case of identity matches, approximately 15 minutes apart. Each face image is cropped to show the head only and converted to greyscale, with examples of these stimuli seen in Figure 1.2. The GFMT has proven to be a very useful task for assessing different factors which may influence face matching ability. This includes the effects of manipulations to the stimuli, such pixelation (Bindemann et al., 2013) and facial wipes (Strathie & McNeill, 2016), and operational task demands at passport control, such as prolonged task duration (Alenezi, Bindemann, Fysh, & Johnston, 2015; Bindemann, Avetisyan, & Rakow, 2012), infrequent mismatches (Bindemann, Avetisyan, & Blackwell, 2010) and time pressure (Bindemann et al., 2016; Özbek & Bindemann, 2011). Furthermore, the GFMT has successfully demonstrated individual performance differences (Megreya, Bindemann, & Havard, 2011; Noyes, Hill, & O'Toole, 2018; White, Rivolta, Burton, Al-

Janabi, & Palermo, 2017), including between those with professional expertise (Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016; White, Kemp, Jenkins, Matheson, & Burton, 2014), and how this may be improved by feedback (Alenezi & Bindemann, 2013) and the formation of groups (Balsdon, Summersby, Kemp, & White, 2018; Dowsett & Burton, 2015; White, Burton, Kemp, & Jenkins, 2013). These factors will be discussed further when considering how face matching is performed in real-world tasks.

Figure 1.2. Example stimuli of match (left) and mismatch (right) face pairs from the GFMT (Burton et al., 2010).

A second pairwise face-matching task is the Kent Face Matching Test (KFMT; Fysh & Bindemann, 2018). These face pairs consist of an image from a student ID card presented alongside a portrait photo. The student ID photos were taken at least three months prior to the face portraits and were not constrained by pose, facial expression, or image-capture device. The portrait photos depict the target's head and shoulders from a frontal view whilst bearing a neutral facial expression and were captured with a high-quality digital camera. Example stimuli are shown in Figure 1.3.

Figure 1.3. Example stimuli of match (left) and mismatch (right) face pairs from the KFMT (Fysh & Bindemann, 2018).

The KFMT is markedly more challenging than the GFMT, with accuracy on both match and mismatch trials approximately 66% (Fysh & Bindemann, 2018). This reflects the difficulty of processing the within-face variability apparent in comparison images taken at different points in time relative to the highly controlled same-day GFMT stimuli (see Megreya, Sandford, & Burton, 2013). Nonetheless, face matching accuracy on these two tasks correlate highly, r = .67 (Fysh & Bindemann, 2018), despite their differences in difficulty. In everyday circumstances, facial comparisons rarely involve same-day images, rather a present-day target image to be compared to a previously taken image, such as to previously obtained Closed Circuit Television (CCTV) video footage or to photographic documentation at security checkpoints.

1.4 Photo-to-video face matching

In everyday life faces are typically viewed in motion, and it is this motion which may provide a cue for identity. With increasing familiarity, facial motion may become incorporated into its representation (Christie & Bruce, 1998), with distinctive motion providing the most

benefit (Butcher & Lander, 2017; Lander & Chuang, 2005). Dynamic information obtained from motion contributes to building an accurate representation by integrating a characteristic motion signature (Lander & Butcher, 2015); viewing multiple static images may show the face from different viewpoints, but not *how* it moves and therefore are less useful than viewing in motion (Pike, Kemp, Towell, & Phillips, 1997; Thornton & Kourtzi, 2002). Non-rigid motion, such as talking and expressing, may capture attention and lead to better learning (Lander & Bruce, 2003; Lander & Chuang, 2005), however such movement must be natural in order for the information to be retained (Kramer, Jenkins, et al., 2016; Lander, Chuang, & Wickham, 2006).

Viewing faces in motion is particularly useful when facial cues are degraded in some way. If identity cannot be established from static cues, knowing how the face moves is a useful supplementary cue for discerning identity (Bennetts et al., 2013; Lander & Butcher, 2015; O'Toole, Roark, & Abdi, 2002). For example, pixelated or blurred images of familiar faces are better recognised if viewed in motion than when static (Lander et al., 2001). This could have useful implications when attempting to identify individuals from poor-quality CCTV footage (Lander, Christie, & Bruce, 1999).

1.4.1 Matching to CCTV footage

The identification of individuals from video footage is a common application of face matching in real-life settings, such as law enforcement. When comparing a high quality photo to a poorer quality image obtained from CCTV, familiar faces are matched with relative ease (over 90% accurate) compared to matching unfamiliar faces, for which accuracy remains poor at approximately 75% (Bruce et al., 2001). In this pairwise matching task, no difference in accuracy was found when matching the individual to the video compared to when matching to

a still image taken from the footage. However, when identifying an individual from the video footage with an eight person lineup, considerable errors were made by selecting a similar-looking individual, on more than 50% of occasions even when high quality target images were used for the lineup (Henderson et al., 2001); at passport control, imposters are therefore more likely to successfully evade detection by using genuine documentation of someone with sufficiently similar facial appearance. Although passport officers match photographic documentation to live people rather than to video footage, this does not markedly improve face matching accuracy (80%) and decreases by a further 20% if presented alongside a photo taken one year apart (Davis & Valentine, 2009), a timeframe clearly possible for a passport valid for up to 10 years.

Identifying unfamiliar people from video footage continues to be difficult under optimised conditions. The use of high quality colour video with close-up full face target images induces 13% identification errors (Davies & Thasen, 2000), and error rates remain high with close-up video footage, especially if the footage is a week old (Davis & Valentine, 2009). In addition, individuals captured by CCTV footage may attempt to disguise their appearance, for example by wearing a hat to obscure their external features, which reduces matching accuracy by a further 20% (Henderson et al., 2001; Lee, Wilkinson, Memon, & Houston, 2009). Trained police officers are also shown to be no more accurate at person identification from CCTV footage than laypersons (Burton et al., 1999; Lee et al., 2009). Furthermore, contemporary police work is increasingly using unmanned aerial vehicles (UAVs), known as drones, for surveillance and yet matching high-quality drone-captured images to photos of unfamiliar people is highly error-prone, even when only attempting to identify a target's sex, age or race from drone stills (Bindemann, Fysh, Sage, Douglas, & Tummon, 2017). Additional operational demands may influence face-matching performance in real-world security settings, which will be discussed further.

1.5 Face matching in the real world

Aside from policing and surveillance, pairwise comparison of unfamiliar faces is commonplace in a multitude of other real-world settings, such as during sales of age-restricted products, business identity cards, and identity verification at border control points where admission of entry relies critically on the routine identification of a large volume of passengers. This is typically achieved by identification from photographic documentation, by comparing the article image with its bearer. However, matching a photo to a live person does not appear to have any advantage over matching two photos, with accuracy in both cases around 85% (Megreya & Burton, 2008); such an accuracy rate is unlikely to be considered acceptable in security settings, which depend upon the accurate performance of this task. In most cases there will be a valid identity match, however on a small number of occasions an imposter may be present (see, e.g., Bindemann et al., 2010; Papesh & Goldinger, 2014). Modern passports are difficult to forge therefore those who attempt to evade recognition are likely to use legitimate documents of a similar-looking individual (Bindemann et al., 2016; Lander, Bruce, & Bindemann, 2018; Meissner et al., 2013; Susa et al., 2019). Since those individuals are likely to have a criminal motive (Johnston & Bindemann, 2013) such cases are vital to detect.

Despite the difficulty of unfamiliar face matching, finding a suitable replacement for photographic documentation that does not rely on human performance is problematic. Electronic facial recognition gates (eGates) were introduced at Heathrow Airport in 2008. Rejections by these automated systems are referred to human operators, thus human performance will always be relied upon for the most difficult comparisons (Jenkins & Burton, 2008a). The eGates are also not infallible; on one occasion a woman was able to pass through the eGate with her husband's passport (Vine, 2012), and high rejection rates can be caused by sunlight shining into the camera lens or the faces of passengers, making the facial comparison difficult (Vine 2014). Furthermore, passengers may also elect not to use automated systems;

an inspection at Manchester airport in 2015 found that take-up rates of eGates by eligible passengers ranged from 30-50% across the three terminals (Bolt, 2016). Security personnel will therefore always be required for identity verification and so their capabilities of performing the task to a suitable standard must be assessed.

1.5.1 Individual differences in unfamiliar face matching performance

Face matching accuracy is typically in the region of 80% (Fysh & Bindemann, 2017a; Hancock et al., 2000; Jenkins & Burton, 2011; Johnston & Edmonds, 2009; Robertson, Middleton, & Burton, 2015), yet there is considerable range across individuals, from 50% accuracy to near perfection, both for lineup and pairwise face matching tasks (Burton et al., 2010; Megreya & Burton, 2006a), and also significant ranges in the Cambridge tests of face memory and perception (CFMT: Duchaine & Nakayama, 2006; CFPT: Duchaine, Germine, & Nakayama, 2007). These tasks have been useful to classify abilities; prosopagnosia patients show severe face processing impairments on these tasks (e.g., Duchaine et al., 2007; Duchaine & Nakayama, 2006; White et al., 2017) whilst some individuals demonstrate superior abilities and consistently outperform normative performance levels, i.e., super-recognisers (Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Bobak, Hancock, & Bate, 2016; Bobak, Pampoulov, & Bate, 2016; Russell, Duchaine, & Nakayama, 2009). Generally, overall performance is stable across tasks (e.g., Burton et al., 2010; Fysh & Bindemann, 2018) but can be inconsistent over time. Bindemann, Avetisyan and Rakow (2012) presented 40 different GFMT trials to participants for five days and demonstrated that whilst overall face-matching accuracy correlated well on consecutive days, this was much weak over extended periods.

Other differences in face matching include those across ages, sex and personality. Older adults are less effective at holistic processing and have shown to be less accurate at face

identification than younger adults (Konar, Bennett, & Sekuler, 2013) and observers are better at matching faces of their own age (Ritchie et al., 2015). Female observers are more accurate on identity match trials of their own sex than male faces, whilst male observers match faces of either sex as accurately. For mismatch trials, female observers are more accurate than male observers, regardless of sex of the target faces (Megreya, Bindemann, & Havard, 2011). In terms of personality, stable and relaxed participants are more accurate on target present trials of the 1-in-10 task than those who are reactive and tense (Megreya & Bindemann, 2013), though Lander and Poyarekar (2015) found no relationship between extraversion and face matching, only for face recognition.

Face matching by professionals

Observers believe that generally others will be as good at matching faces as themselves, yet expect those with experience such as passport officers to have superior abilities (Ritchie et al., 2015). However, given the notable variation in individual ability, such personnel are likely to widely vary in their ability to perform this task (Lander et al., 2018). Despite years of professional experience verifying identification documents, notaries and bank tellers are no more accurate at matching a face photo with a student ID photo on a mock driving license (Papesh, 2018). Similarly, those untrained in facial identification are as accurate at face recognition from low quality CCTV footage as those who have received formal training (Lee et al., 2009), although forensic experts are more likely to give careful conclusions (and prefer unresolved decisions than being incorrect) in such cases (Norell et al., 2015).

Expertise is a critical factor for high face-matching performance by professionals. Whilst passport officers have been shown to perform this task to a similar level as untrained students (White, Kemp, Jenkins, Matheson, & Burton, 2014), more specialised facial examiners

outperform facial review passport staff and controls (White, Dunn, Schmid, & Kemp, 2015; White, Phillips, Hahn, Hill, & O'Toole, 2015). Fingerprint specialists, with matching expertise in an alternative forensic discipline, are also more accurate at matching faces than students (Phillips et al., 2018). Although fingerprint specialists as a group could not outperform face specialists, this was not the case at an individual level. No relationship between employment duration and accuracy has been found (see Figure 1.4.; White, Dunn, et al., 2015; White, Kemp, Jenkins, Matheson, & Burton, 2014), with some junior officers performing unfamiliar face matching more accurately than their more experienced colleagues (Robertson et al., 2016; Wirth & Carbon, 2017). It has therefore been investigated whether face matching could be improved with training, or whether recruitment should instead focus on hiring individuals with a predisposed talent for face processing.

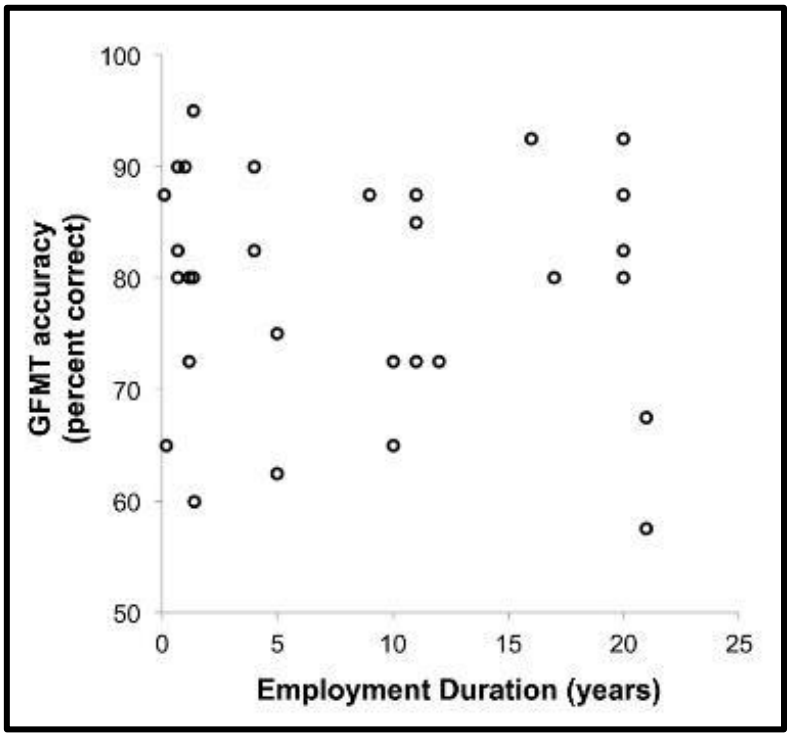

Figure 1.4. Passport officer data plotting employment duration against their score on the GFMT (White, Kemp, Jenkins, Matheson, & Burton, 2014).

Facial comparison training

Training personnel in face matching seems an intuitive method for improving task performance, yet there is mixed evidence for its efficacy. Feature-based training whereby observers are directed to focus on particular features has produced varied results. The classification of face shape is a common strategy encouraged by training programmes, however Towler, White and Kemp (2014) demonstrated such a strategy is unreliable; different images of the same person are frequently classified as having different face shapes, both between- and within-observers. Attending to specific facial features which appear to be more stable across images may enhance face matching, but need to be selected carefully as some may be detrimental to performance (Megreya & Bindemann, 2018; Towler, White & Kemp, 2017). Thus, it is imperative that suitable features are selected should such strategies be employed, although this approach does not appear to cross over to other-race faces (Megreya & Bindemann, 2018), an important consideration in real-world settings such as passport control. Professional facial image comparison training courses used in a variety of real-world security settings globally contain such components, training staff in feature comparison (Towler et al., 2019). When using representative examples of these courses to train novices, however, Towler and colleagues did not find any notable improvement in facial identification accuracy.

Another training approach is providing feedback after decisions have been made so observers can monitor their performance. In real-life settings one could expect face-matching decisions to be performed for long periods of time (Alenezi et al., 2015) and trial-by-trial feedback successfully inhibits a decline in accuracy over extended durations (Alenezi & Bindemann, 2013). Furthermore, low-aptitude performers can be brought up to the level of high-aptitude performers when provided with trial-by-trial feedback, although this has not been shown to assist high-aptitude performers to improve further (White, Kemp, Jenkins, & Burton, 2014). Similarly, when face matching is performed in pairs or groups, accuracy is higher than

when the task is carried out by individuals (Dowsett & Burton, 2015; White et al., 2013), with further benefit for low-aptitude performers who perform better after having worked in a pair. However, mismatches appear infrequently in passport control settings (see, e.g., Bindemann et al., 2010; Fysh & Bindemann, 2017b, 2018; Papesh & Goldinger, 2014; Susa, Michael, Dessenberger, & Meissner, 2019) and despite feedback appearing to draw attention to low prevalence stimuli, face-matching judgements remain conservative thus resulting in a high miss rate under these conditions (Papesh, Heisick, & Warner, 2018). Given the mixed evidence supporting the efficacy of training strategies, the selection of high-aptitude observers may instead be a viable alternative for improving task performance (Lander et al., 2018; Robertson et al., 2016).

Personnel selection

As discussed, it is clear that there are substantial individual differences in observers' face-processing abilities (e.g., Bindemann, Avetisyan, & Rakow, 2012; Burton et al., 2010; Duchaine & Nakayama, 2006; White et al., 2013, 2017). Moreover, some individuals are capable of performing person identification tasks far better than average, i.e. super-recognisers (Bobak, Bennetts, et al., 2016; Bobak, Hancock, & Bate, 2016; Russell et al., 2009). Super-recognisers typically outperform controls on both recognition and perceptual tasks, such as face matching whereby no demands are placed on memory. It appears that they are not simply on the opposite end of a spectrum to those with face-processing difficulties since they adopt qualitatively different strategies, for instance attending longer to central face regions (Bobak, Parris, Gregory, Bennetts, & Bate, 2017). However, at an individual level, not all super-recognisers are superior to controls on matching tasks, suggesting recognition superiority is dissociable from general perceptual expertise for faces (Bobak, Bennetts, et al., 2016; Bobak,

Dowsett, & Bate, 2016); potentially, there may be instances of super-matchers who are not super-recognisers (Bobak, Dowsett, & Bate, 2016), akin to how the latter are dissociable from super-memorisers (Ramon et al., 2016).

In order to evaluate an individual's face-processing ability, assessments should consist of multiple tests because single tests may not be representative. This has been highlighted by recent studies examining how different tasks tap into different aspects of face processing. McCaffery, Robertson, Young and Burton (2018) compared performance on the GFMT (Burton et al., 2010), the CFMT (Duchaine & Nakayama, 2006), and the Before They Were Famous task (BTWF: Russell et al., 2009). Performance on these tasks of unfamiliar face matching, unfamiliar face recognition, and familiar face recognition all correlated with one another, yet a maximum of 25% of the variance in performance across these tasks could be accounted for by a general face-processing ability, suggesting some task-specific demands. Similarly, Balsdon et al. (2018) used the GFMT, CFMT, a real-world passport task and a self-report questionnaire, the 20-item Prosopagnosia Index (PI20: Shah, Gaule, Sowden, Bird, & Cook, 2015) as "pre-screening" for high-performing individuals to predict performance in a second run of the passport task. Moderate correlations were found between all tasks and modest improvements were made on an individual level for high performers, whilst aggregating the accuracy of groups provided substantial performance gains.

Balsdon and colleagues (2018) further note the challenge for organisations to improve performance through selecting personnel based on the current tests available, due to the relatively modest gains in performance possible by these methods. Recruitment tasks need to provide a closer proximate to the real-world task expected to be faced by recruits, since current laboratory paradigms do not adequately capture their complexities (Ramon, Bobak, & White, 2019); for example, at present, it is unknown whether laboratory face-matching tasks correlate strongly with a passport control task found in real-life security environments (Lander et al.,

2018). This could perhaps be achieved through increasing task difficulty in order to establish the very top performers (Balsdon et al., 2018), and comparing this to a simulation of a real-world task.

1.5.2 Environmental factors

In real-world settings other factors within the environment may influence the face matching task to be performed. This includes the operational demands of the environment, such as prolonged task durations and completing other identification checks, as well as performing under time pressure and detecting the infrequent imposters disguised amongst the majority of legitimate matches. Such factors are likely to have a significant impact on an already difficult task.

Operational demands

At passport control, officers perform the repetitive task of identity verification for long durations at a detriment to their accuracy over the course of the day. Alenezi et al. (2015) noted a persistent decrease in accuracy over five blocks of 200 GMFT trials separated by 5 minute breaks whereby observers lost the ability to tell faces apart, which could not be alleviated by changing desks. Furthermore, aside from checking whether the passport bearer matches the photo, officers examine the information contained in the passport. When a face is contained within a passport-style frame, observers are more biased to report a match; yet if an identity match is indeed present, invalid information is unlikely to be detected (McCaffery & Burton, 2016). This suggests that faces draw attention and are processed first, therefore the presence of passenger queues may also impair identification by competing for attention (see, e.g.,

Bindemann, Burton, & Jenkins, 2005; Bindemann, Sandford, et al., 2012; Megreya & Burton, 2006b).

This passport face-matching task is further complicated by the self-selection of passport images by the owner of the document. Those unfamiliar with the individual do not judge self-selected images as most representative of the person, which are matched with 7% more errors (White, Burton, & Kemp, 2016). This has also been seen with familiar faces, with some photos of celebrities rated as capturing their likeness better than others (Jenkins, White, et al., 2011). The poorer likeness of self-selected images may be as a result of a person not needing to regularly recognise their own face and so their expectation of their current appearance is distorted (White et al., 2016). In addition, consider that passports can be valid for up to 10 years; facial appearance can change dramatically over this time period and so face-matching accuracy rates would further decrease (see Megreya, Sandford, & Burton, 2013).

Time pressure

Passport officers are also required to complete their identity checks without excessive delay to the passengers. Staff are expected to process 95% of EEA passengers within 25 minutes of them joining the queue, whilst non-EEA passengers should be cleared within 45 minutes (Bolt, 2015). This can be a strain on the system, for example at Stansted Airport in 2013 enough resources were available to process 3300 passengers per hour yet the number of arrivals could exceed 4000 per hour (Vine, 2014). To minimise disruption to passengers, since August 2017 the number of eGates monitored at a time by Stansted passport control staff doubled from five to 10 (Bolt, 2018). Under low time pressure and without additional tasks to complete, 10% of mismatches may go undetected with false rejections at a rate of approximately 50%; the addition of one of these factors has little impact on performance,

however with both task pressures accuracy further deteriorates (Lee, Vast, & Butavicius, 2006). Strict time constraints impair accuracy and over extended durations can lead to a match bias, i.e. difficulty in telling people apart (Bindemann et al., 2016; Fysh & Bindemann, 2017b). It has been suggested that time constraints direct attention to internal facial features and should therefore improve face matching as the more changeable external features exert less influence (Fletcher, Butavicius, & Lee, 2008), however undoubtedly optimal performance is obtained under unconstrained conditions (Özbek & Bindemann, 2011). The detection of imposters is of primary concern in security settings and so the impact of factors such as time pressure needs careful consideration.

Infrequent and disguised imposters

Imposters are unlikely to appear frequently in settings such as passport control and will take measures to avoid detection. A person can disguise themselves by attempting evasion, i.e. trying to not look like themselves, or by attempting to impersonate someone else. Evasion is a more effective disguise producing 35% errors, although impersonation can still induce 9% more matching errors than no disguise (Noyes & Jenkins, 2019). Impersonation is the type of disguise one may expect to find at passport control, with an imposter attempting to replicate the facial appearance of someone else's passport image.

This provides an additional challenge to security officers; in the majority of cases they will be processing legitimate identity matches, yet careful attention is required throughout to detect difficult infrequent mismatches which are intended to be similar in appearance to their purported passport. For same-day GFMT stimuli, mismatches are detected at similar rates when present on both 50% and 2% of trials, and also when the frequency of mismatches was known in advance with unconstrained viewing time of all face pairs (Bindemann et al., 2010).

However, when photos are taken at different time points, a low-prevalence effect occurs; with photos pairs of an average difference of 1.5 years, infrequent mismatches (10% of trials) are more likely to be missed than high prevalence mismatches (50% of trials) even when given the opportunity to correct uncertain responses (Papesh & Goldinger, 2014). Both disguise and the low-prevalence of mismatches exacerbate an own-race bias (Meissner et al., 2013; Susa et al., 2019), with poorer accuracy and increased overconfidence in other-race decisions compared to own-race faces.

Body language

Passport officers may rely on alternative cues to assist their detection of imposters. When faced with a long queue to process, someone behaving unusually relative to their fellow passengers may raise suspicion. Though facial information is the primary contributor towards person identification, body information also appears to have valuable input (Robbins & Coltheart, 2012), especially in the identity matching of unfamiliar people. For example, although the face outperforms the body in identity matching tasks when these types of stimuli are presented in isolation, accuracy is best when both sources of information are available (Rice, Phillips, & O'Toole, 2013). This effect appears to be amplified by increasing viewing distance, which shifts observers' reliance on identity information further towards the body (Hahn, O'Toole, & Phillips, 2016). The utility of combining facial and body information has also been highlighted in identity sorting tasks, where intra-personal variability is easier to distinguish for whole persons than faces and bodies in isolation (Balas & Pearson, 2017). Remarkably, however, observers' self-reports of usage are much lower for body features than internal facial features when making identifications. This suggests that observers often remain

unaware of their reliance on body information as identity cues when facial information is insufficient (Rice, Phillips, Natu, et al., 2013).

Further evidence for the integration of the body with facial information in person identification comes from paradigms that present people in motion. This research shows that facial information is prioritised over body cues when static stimuli are observed, but both are utilised in a more balanced manner when dynamic stimuli are used, resulting in superior person identification accuracy (O'Toole et al., 2010). This effect persists when moving footage from video clips is compared with multiple static images (Simhi & Yovel, 2016), thus providing converging evidence that it is motion itself that enables information from multiple cues, such as the face and body, to be combined to enable accurate person identification (Yovel & O'Toole, 2016).

These findings highlight the role of the body in person identification and suggest also that the progression of research in this field is limited by the use of static stimuli in investigations of such non-facial cues. Nonetheless, whilst previous studies have looked at the impact of the body on person identification, they do not address how specific body language, that is not indicative of identity *per se* but may reflect a hidden motivation, might affect identification in security settings. People seeking to avoid detection at airports may, for example, betray their intention through common non-verbal cues of anxiety, such as restless fidgeting (Ekman & Friesen, 1969). With regard to face matching, the impact of such factors is difficult to study. Only a few studies have examined face matching in real-world interaction (e.g., Kemp, Towell, & Pike, 1997; White, Kemp, Jenkins, Matheson, & Burton, 2014), but such experiments are logistically challenging, and variables such as non-verbal behaviour are difficult to control systematically. Consequently, additional measures, such as double-blind procedures, are taken to *prevent* intrusion of such variables. Equally, such factors are difficult

to study systematically in occupational field settings, such as at passport control, due to the security-sensitive nature of this task.

1.6 Face matching in a virtual world

Investigating face matching as it occurs in real-world environments is practically impossible due to the security sensitive nature of the task. As a compromise, studies in the field have typically involved simplified variations of the real-world task to isolate a variable of interest (e.g., Kemp et al., 1997; Megreya & Burton, 2008; White, Kemp, Jenkins, Matheson, & Burton, 2014). Whilst such paradigms have been useful to assess face-matching ability by persons who routinely complete such tasks in their occupation, the complexities of their working environment are difficult to capture by this method. Furthermore, social interaction factors such as body language are difficult to manipulate accurately with experimental control. Virtual reality (VR) may provide a potential solution to these problems, by immersing observers into interactive reconstructions of real-world environments.

1.6.1 Virtual reality as a research method

A principal advantage of conducting experimental research in VR is the diminished trade-off between experimental control and ecological validity compared to traditional laboratory methods (see Figure 1.5). Unlike in real-world field studies, researchers have complete control over the environment, both in terms of its visual appearance and what the observers experience (Blascovich et al., 2002; Bombari, Schmid Mast, Canadas, & Bachmann, 2015; de la Rosa & Breidt, 2018; Loomis, Blascovich, & Beall, 1999). Social interaction experiments demonstrate this trade-off and difficulty of replication; ideally, interaction partners would behave

realistically and identically for every participant. Laboratory experiments may opt for standardisation by using vignettes or for realism with trained actors (Bombari et al., 2015), however asking participants to imagine an interaction does not capture a realistic communication whilst actors are difficult to standardise (Blascovich et al., 2002). With VR, on the other hand, researchers precisely control the social situation (the people involved, how they interact, etc.) which is replicated with every iteration of the experiment (de la Rosa & Breidt, 2018; Fox, Arena, & Bailenson, 2009; Pan & Hamilton, 2018). For example, investigating the impact of unusual body language on face matching would be difficult to standardise in real life as different actors may interpret "unusual" differently and not perform the same to each observers, yet with VR virtual humans (avatars) can be coded with animation and a set behaviour can be replicated many times over.

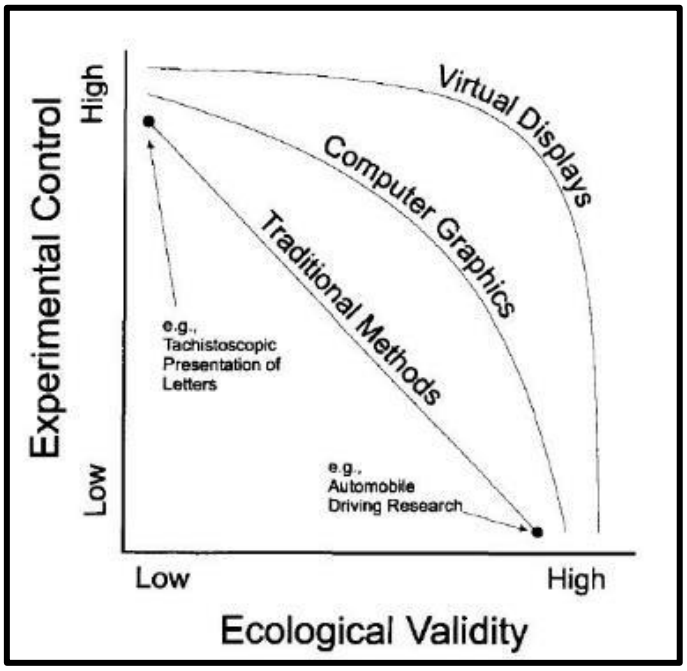

Figure 1.5. Comparison of trade-off between ecological validity and experimental control for laboratory methods (Loomis et al., 1999).

Similarly, VR enables precise replication of experimental conditions by different researchers. Once coded, providing colleagues have the appropriate equipment, experiments can be shared across research groups. This allows for data collection from distal locations and increases potential sample size and variability (Fox et al., 2009; Pan & Hamilton, 2018). This is one of many practical benefits of VR, as such replication by research groups would not be possible in field experiments.

Furthermore, factors which are impossible to study in real life can be recreated in VR to answer novel research questions (de la Rosa & Breidt, 2018; Fox et al., 2009; Loomis et al., 1999; Wilson & Soranzo, 2015). For example, Rosenberg, Baughman and Bailenson (2013) investigated how prosocial behaviour could be encouraged by participants taking on the role of a superhero in VR. They found that those who could fly like Superman to find a missing child in a city were subsequently more helpful to the experimenter than those who searched by helicopter. VR can also be used to simulate situations which would be dangerous or unethical in real life (Bombari et al., 2015; Pan & Hamilton, 2018; Wilson & Soranzo, 2015); Slater and colleagues (2006) recreated Milgram's obedience studies in VR and, despite knowing the virtual characters were not real and that they could withdraw from the experiment, participants were stressed by giving shocks to the learner and exhibited "caring" behaviour by delaying shocks for incorrect responses, yet generally continued until the end.

Experimentation in VR presents its own challenges which need to be addressed. First, it can be time-consuming and costly to set up, with technical expertise required to program the experiments (de la Rosa & Breidt, 2018; Loomis et al., 1999). Head-mounted displays (HMDs) are becoming cheaper and more accessible as the technology develops (Wilson & Soranzo, 2015) however additional equipment may be required for eye- and body-tracking experiments. The HMDs must also be carefully calibrated to minimise risk of after effects such as motion sickness and reduced hand-eye coordination (de la Rosa & Breidt, 2018; Loomis et al., 1999).

This can be alleviated by slowing any essential motion and reducing the intensity of optics (Pan & Hamilton, 2018). However, the main challenge facing VR is the visual realism and creation of avatars (Bombari et al., 2015; Loomis et al., 1999; Pan & Hamilton, 2018). Avatars need to be believable in order to encourage realistic interactions, which can be time-consuming and difficult to develop. Ideally avatars would have high photorealism and smooth motion akin to humans. As VR technology continues to develop, this is becoming increasingly possible; Figure 1.6 illustrates how the visual quality of virtual environments has improved in the past decade.

It is clear that VR has the potential to be a highly useful methodological tool. It has previously been investigated as a treatment method for phobias and post-traumatic stress disorder (e.g., Parsons & Rizzo, 2008; Nelson, 2013) and also for training professionals in potentially hazardous situations, such as surgery (e.g., Seymour et al., 2002) and firefighting (e.g., Cha, Han, Lee, & Choi, 2012). Furthermore, topographical difficulties are associated with developmental prosopagnosia and spatial navigation can be improved with VR cognitive map training, with subsequent improvement on face memory tasks (Bate, Adams, Bennetts, & Line, 2017). Face perception research is an emerging discipline for VR experimentation, with research to date focusing on eyewitness lineup procedures.

Figure 1.6. The development in visual quality of virtual environments; the top pane shows a casino environment (Bailenson, Blascovich, Beall, & Noveck, 2006), the bottom pane shows the immersive virtual airport to be used in the current series of experiments.

1.6.2 Face perception in virtual environments

When immersed in VR, people respond to avatars in a similar manner to how they would interact with others in real life. For example, the shape and size of a personal space area given

to avatars closely resembles that typically afforded to humans (Bailenson, Blascovich, Beall, & Loomis, 2003). Although a more difficult task than recognition from photographs, observers can discriminate old and new virtual faces to an above chance level (Bailenson, Beall, Blascovich, & Rex, 2004); similar to the processing a real faces, this virtual face recognition task becomes more challenging when viewing angle differs from learning to test (e.g., Bruce et al., 1999). However, if one considers the context of eyewitness identification, using VR can overcome these challenges by recreating the interpersonal distance and viewing angle the eyewitness observed at the time. Bailenson et al. (2008) explored the potential of virtual identification lineups for context reinstatement of the crime scene, demonstrating that when the distance between the eyewitness and suspect was matched for the lineup, and providing the eyewitness with unlimited viewing angles, identification accuracy was improved. Such reconstructions also have possible courtroom applications (e.g., Bailenson et al., 2006) by allowing lawyers and jurors to understand the subjective perspective of eyewitnesses and defendants. Virtual lineups can be manipulated so that location, clothing, viewpoint and distance can be controlled (see Figure 1.7) and have the practical advantage of not needing to recruit real people as foil identities; fillers could be selected from a database and chosen for resemblance to the suspect, making a fair lineup more feasible (Segovia, Bailenson, & Leonetti, 2012).

In addition, police officers blind to the identity of the suspect cannot influence eyewitnesses' selection, which reduces the risk of false identifications. This has practical issues owing to the demand on resources, yet the implementation of a virtual police officer to carry out the lineup procedure could provide a solution. Eyewitness accuracy is highly similar for virtual officers administering the lineup through guided conversation compared to a real police officer, and the virtual officer was also deemed to be less confusing (Cutler, Daugherty, Babu, Hodges, & Van Wallendael, 2009; Daugherty, Van Wallendael, Babu, Cutler, & Hodges,

2008). Future developments may also permit changes to characteristics of the virtual officer such as matching their gender to the victim in assault cases or their language for non-native speakers (Daugherty, Babu, Cutler, & Hodges, 2007).

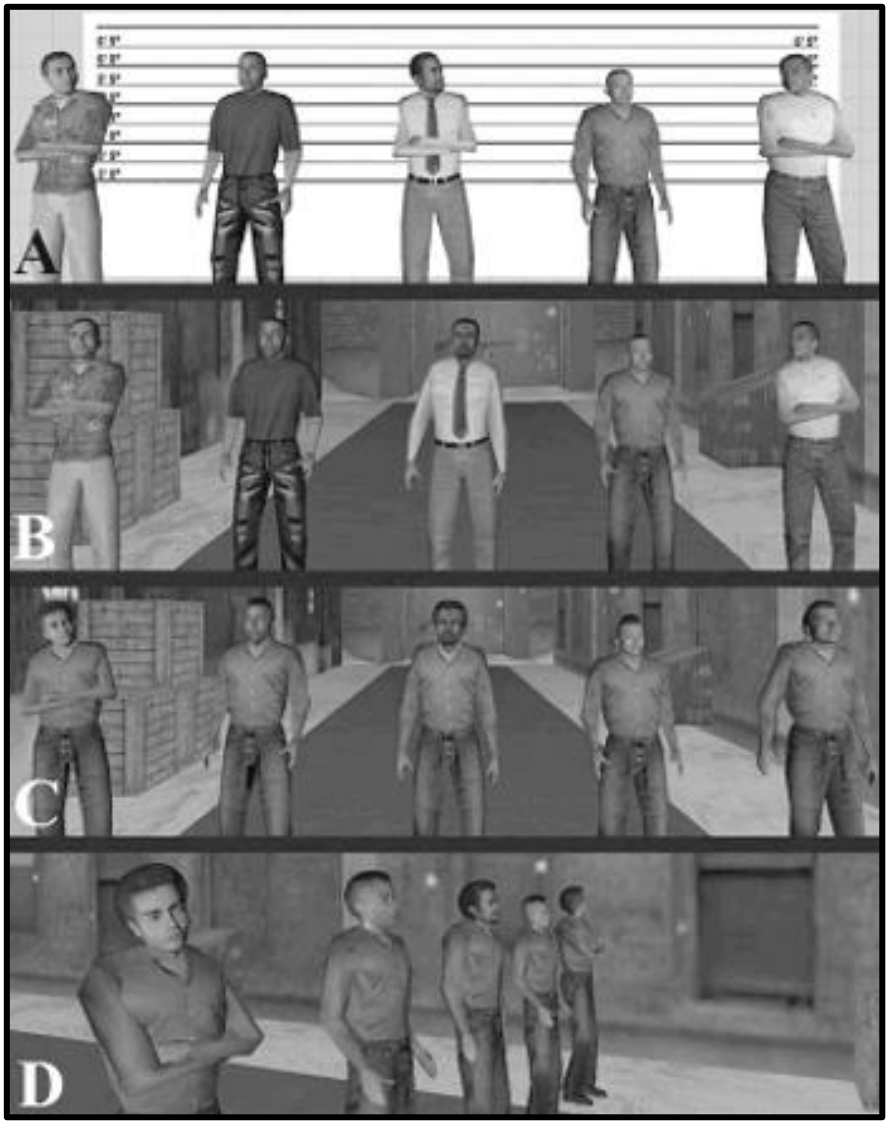

Figure 1.7. Virtual lineup for eyewitness identification (Segovia et al., 2012). A lineup is created (A) which can then be placed in a reconstruction of the crime scene (B). Features such as clothing can be standardised (C) and eyewitnesses may view all persons close up and at any angle (D).

There is clear potential for the implementation of VR in forensic settings. The simulation of complex environments provides the opportunity to develop new research which cannot feasibly be investigated in the field. The security sensitive nature of face matching at passport control prevents the manipulation of variables likely to have a profound impact on the task, such as body language, in real-world experimentation. Using VR as an exploratory research method is an innovative approach in the face-matching domain; its efficacy and potential applications and will be assessed during this thesis.

1.7 Structure of this thesis

The overall aim of this thesis is to provide a foundation for further face-matching research with VR, by demonstrating that this approach can capture the face processes that are currently studied with more simplistic laboratory approaches. Chapter 2 begins by assessing the construction process of rendering photographs of real faces onto avatars from an existing database to create identity pairs. Face portraits of the new avatars are compared to the source photographs from which they were derived to confirm identity is successfully captured (Experiments 1 to 3). The second phase of experimentation compares avatar matching with two established laboratory tests of face matching to explore whether the same face processes are utilised (Experiments 4 and 5). In the final phase, avatar matching is assessed in an immersive VR airport environment simulating passport control (Experiments 6 and 7). This series of validation experiments aims to demonstrate avatars can provide a suitable substrate to study face-identification processes in VR.

Chapter 3 then explores the use of the VR passport control task (VRPC) as an assessment tool for personnel in real-world security settings. Performance on the VRPC task is compared with aptitude on laboratory facial identity comparison tasks and a self-report measure of face

processing (Experiment 8). The selection of individuals with an aptitude for facial identification may provide a viable alternative to training, and the VRPC task provides a suitable trial of the face-matching task recruits would be expected to perform in their working environment.

A further application of the VRPC task is to assess the impact of social interaction factors on face matching. Chapter 4 investigates whether alternate displays of body language would be perceived as unusual in a passport control context and lead to enhanced scrutiny of facial identity. To manipulate body language, the majority of passengers were programmed to have small shifts in body posture, whilst for a small number of passengers these shifts were exaggerated to simulate more restless body language. It was examined whether identity mismatches would be detected more frequently when exhibiting unusual body language, and whether these behavioural differences are consciously observed (Experiments 9 and 10) or direction to inspect unusual behaviour is required (Experiments 11 to 13). Finally, it was investigated whether individual face-processing ability, assessed with laboratory tasks, attenuates the influence of body language on face matching (Experiment 14). This thesis concludes by discussing the possible future applications of VR for investigating person identification in security settings.

Chapter 2:

Facial Identification at a Virtual Reality Airport

Introduction

Passport officers at airports and national borders are widely required to verify the identity of passengers by comparing their faces to passport photographs. People seeking to avoid detection at such security controls may attempt to do so by acting as impostors, using valid identity documents that belong to other persons who are of sufficiently similar facial appearance. In psychology, this task has been studied extensively as unfamiliar face matching (for reviews, see Fysh & Bindemann, 2017a; Jenkins & Burton, 2008a, 2011; Robertson, Middleton, & Burton, 2015). In experiments in this field, observers are typically required to match pairs of face photographs, which are presented in isolation on blank backgrounds, and have to decide whether these depict the same person or two different people.

This general approach has been successful for isolating and understanding a range of important factors, such as *observer* characteristics. For example, pairwise face-matching experiments have been used to assess individual differences in performance (e.g., Bindemann, Avetisyan, & Rakow, 2012; Bobak, Dowsett, & Bate, 2016; Bobak, Hancock, & Bate, 2016; Megreya & Burton, 2006a), to compare untrained observers with passport officers (White, Kemp, Jenkins, Matheson, & Burton, 2014; Wirth & Carbon, 2017) and different groups of professionals, such as facial review staff and facial examiners (White, Dunn, Schmid, & Kemp, 2015; see also Phillips et al., 2018; White, Phillips, Hahn, Hill, & O'Toole, 2015), and to assess observers familiar and unfamiliar with the target identities (Bruce, Henderson, Newman, & Burton, 2001; Ritchie et al., 2015), as well as those with impairments in face matching (White, Rivolta, Burton, Al-Janabi, & Palermo, 2017). Similarly, such controlled laboratory experiments have been employed to study how the characteristics of *stimuli* affect face matching, by exploring factors such as image quality (e.g., Bindemann, Attard, Leach, & Johnston, 2013; Strathie & McNeill, 2016), the addition of paraphernalia and disguise (Henderson, Bruce, & Burton, 2001; Kramer & Ritchie, 2016; Wirth & Carbon, 2017), and

variation in viewpoint (Estudillo & Bindemann, 2014), camera distance (Noyes & Jenkins, 2017), and facial appearance across photographs (e.g., Bindemann & Sandford, 2011; Megreya, Sandford, & Burton, 2013).

While this research has advanced understanding of face matching considerably, these paradigms provide a limited proxy for studying how the environment and social interaction might affect this task. In real-life environments, passport officers may, for example, resort to non-facial cues, such as body language, to support identification decisions (Rice, Phillips, Natu, An, & O'Toole, 2013; Rice, Phillips, & O'Toole, 2013). Similarly, environmental factors, such as the presence of passenger queues, might impair identification by exerting time pressure on passport officers (see, e.g., Bindemann, Fysh, Cross, & Watts, 2016; Fysh & Bindemann, 2017b; Wirth & Carbon, 2017) or competition for attention (see, e.g., Bindemann, Burton, & Jenkins, 2005; Bindemann, Sandford, Gillatt, Avetisyan, & Megreya, 2012; Megreya & Burton, 2006b). The impact of such factors is likely to be huge but not captured by current laboratory paradigms, and practically impossible to study in real life owing to the importance of person identification at passport control.

As a compromise, a few studies have moved beyond highly controlled laboratory paradigms to study this task in simplified field settings (e.g., Kemp, Towell, & Pike, 1997; Megreya & Burton, 2008; White, Kemp, Jenkins, Matheson, & Burton, 2014). White and colleagues, for example, examined passport officers' matching accuracy under live conditions, in which target identities were presented in person and compared with a face photograph on a computer screen. Such paradigms are valuable for assessing whether limitations in face-matching accuracy are also observed in interpersonal interaction, but are logistically challenging. Moreover, such set-ups do not adequately capture the complexity of real-life passport control environments, and cannot provide the control that experimenters might desire

to manipulate environment and social interaction factors accurately for psychological experimentation.

In this chapter, a potential solution to these problems is proposed, by examining face matching in virtual reality (VR). In recent years, this technology has developed rapidly to provide affordable high-capability VR equipment. With VR, viewers can be immersed in detailed, interactive, and highly controllable three-dimensional (3D) environments that conventional laboratory experiments cannot provide. However, this approach is completely new to face matching. This chapter reports an exploratory series of experiments to investigate the potential of VR for increasing our understanding of face matching. The overall aim is to provide a foundation for further face-matching research with VR, by demonstrating that this approach can capture the face processes that are currently studied with more simplistic laboratory approaches.

In VR, people are represented by animated 3D avatars, on which the two-dimensional (2D) faces of real persons are superimposed. In the first phase of experimentation, the quality of the resulting person avatars is assessed in a tightly controlled laboratory task, in which these 3D avatars are presented back as isolated 2D images, to establish that these capture the identities from which they were derived (Experiments 1 to 3). In the second phase of the chapter, identity-matching of these avatars is compared with two established laboratory tests of face matching (Experiments 4 and 5). In the final phase, identification of avatars is then assessed in an immersive 3D VR airport environment (Experiments 6 and 7).

Phase 1: Avatar face construction and validation

Phase 1 begins with a description of the construction of the person avatars for experimentation. The initial stimulus sets consisted of 129 male and 88 female professional German sportspeople. As these identities were required to be unfamiliar to participants, a pre-

test was carried out to ensure these people were not generally recognizable to UK residents. A list of the identities was presented to 20 students who were asked to cross the names of anyone who they would recognise. Identities familiar to two or more people were excluded. From those who remained, 50 male and 50 female identities were selected for avatar creation. Two full-face portrait photographs were employed for each of these sportsmen and women, which were obtained via Google searches.

The person avatars for this VR paradigm were created by combining these face photographs with an existing database of person avatars (see www.kent.ac.uk/psychology/downloads/avatars.pdf) with graphics software (Artweaver Free 5). The internal features of the face were cut as a selection from the photograph and overlaid onto the base avatar's graphics file. The size of the selection was altered to best map the features onto the positions of the base avatar's features. This was then smoothed around the edges and skin colour adjusted to blend in with the base avatar. Note that the 3D structure of the avatar faces could not be adapted to that of the face photographs, as extraction of such shape information is limited from 2D images. This may be sub-optimal for modelling face recognition, to which both texture and shape information contribute (e.g., O'Toole, Vetter, & Blanz, 1999). However, face recognition is also tolerant to dramatic manipulations of shape (see Bindemann, Burton, Leuthold, & Schweinberger, 2008; Hole, George, Eaves, & Rasek, 2002) and texture appears to be more diagnostic for face identification and face matching (see, e.g., Calder, Burton, Miller, Young, & Akamatsu, 2001; Hancock, Burton, & Bruce, 1996; Itz, Golle, Luttmann, Schweinberger, & Kaufmann, 2017). Therefore, this method for combining the 2D photographs with animated 3D avatars captures the most diagnostic information for identification. In addition, to mitigate for the fact that original shape information could not be incorporated, the same base avatar was employed for both face photographs of each identity. However, avatar elements such as clothing were changed to create two unique appearances for each instance of a person.

Therefore, for each of the 100 identities retained, two avatars were created. For the experiments reported here, this pool of avatars provided sufficient stimuli to create identity-match pairs consisting of two avatars of the same person, and identity-mismatch pairs consisting of two avatars from different people.

As an initial step, confirmation was required that the resulting avatars adequately capture the identities of the face set. For this purpose, a 2D face portrait of each finished identity avatar was recorded. These images were constrained to reveal the internal facial features only (i.e., not hairstyle) and sized to 438 (w) x 563 (h) pixels at a resolution of 150 ppi. In addition, a 2D full-body image, which showed a frontal view of the avatar with arms outstretched, was also recorded and sized to 751 (w) x 809 (h) pixels at a resolution of 150 ppi. The procedure for avatar construction is illustrated in Figure 2.1.

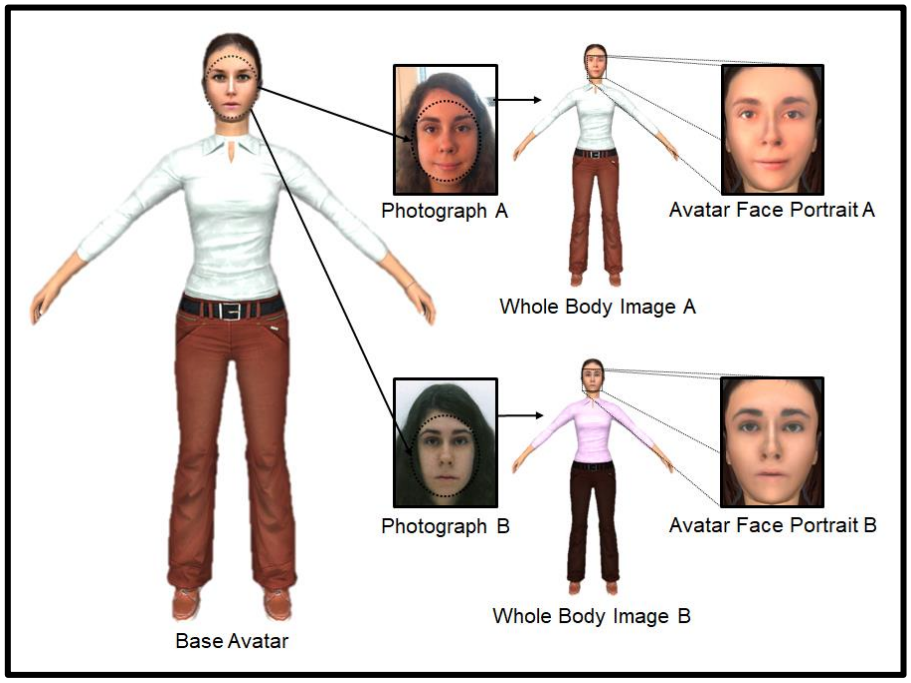

Figure 2.1. An illustration of avatar construction. 2D face photographs were superimposed on animated 3D avatar bodies, whose clothing could be adapted for different identities. 2D face portraits and full-body images were then derived from the 3D avatars for initial experimentation.

Experiment 1

The aim of Experiment 1 was to assess whether the production process of the avatar faces sufficiently captures the images and identities on which these are based. If so, then observers should be able to match these identities in a pairwise comparison. This was assessed with a face photograph-to-avatar matching test with three conditions. These comprised trials on which an avatar face portrait was paired with the original source face photograph (same-image identity-match), trials on which an avatar face portrait was paired with a different face photograph of the same person (different-image identity-match), and trials on which the avatar face portrait was paired with a face photograph of a different person (identity-mismatch). Participants were asked to match these stimulus pairs according to whether they depicted the same person or two different people. The detection of matches and mismatches are dissociable processes and individuals' ability to match these different stimuli types does not tend to correlate (Megreya & Burton, 2007). As such, an overall accuracy score will often not be an appropriate measure as it would not capture how individuals perform the task. For example, an overall accuracy of 50% could be obtained by perfect match accuracy and a failure to detect any mismatches, or by accurately detecting half of the both match and mismatch trials. Therefore, for this and all subsequent experiments, data will be separated by trial type when analysed unless where otherwise stated.

Method

Participants

Thirty Caucasian participants (12 male, 18 female) with a mean age of 21.6 years (*SD* = 3.7 years), who reported normal or corrected-to-normal vision, were recruited at the University of Kent for course credit or a small payment. This sample size is directly comparable

to face matching studies using a broad range of paradigms (e.g., Bindemann et al., 2013; Megreya & Burton, 2007; White et al., 2017).

Stimuli and Procedure

Each participant was presented with 80 trials across two blocks, with each block comprised of the following image-type trials. Firstly, 10 same-image identity-match pairs were produced, which consisted of a 2D avatar face portrait and the high-quality face photograph used to create that avatar. Secondly, 10 different-image identity-match trials were included, in which the 2D avatar face portrait was combined with a different photograph of the same person. These trials did not consist of any of the identities shown in the same-image identity-match trials. Finally, 20 mismatch trials were created. In these, the 2D avatar face portrait was paired with a photograph of a different person, which was chosen by the experimenter for its general visual similarity.

The stimuli of the second block consisted of the same identity pairings as the first block (i.e., 10 same-image identity-match, 10 different-image identity-match, 20 mismatch) but with the reverse image-type pairings, as demonstrated in Figure 2.1. For example, if an observer saw avatar face portrait A paired with photograph B for an identity in Block 1, in Block 2 for the same identity avatar face portrait B was paired with photograph A. Thus, all participants saw each identity twice during the course of the experiment but each image (avatar face portrait or face photograph) only once. All of these images were presented on a white background, with the avatar face portrait to the left and the face photograph to the right of centre. Both images were sized to 70mm (w) x 90mm (h) and were presented 50 mm apart.

In the experiment, each trial began with a 1-second fixation cross, followed by a stimulus pair, which remained on screen until a matching decision had been made. Participants were asked to decide as accurately as possible whether a stimulus pair depicted the same person

or two different people, by pressing one of two corresponding buttons on a standard computer keyboard. Participants were instructed to decide whether the *identity* of the two faces was a match as opposed to whether the *images* matched. It is important to make this distinction because unfamiliar faces may invoke image-comparison techniques instead of using face-specific processes when solving the task, and therefore the similarity of the images could guide responses instead (Burton, 2013). Participants were not informed about the ratio of match-to-mismatch trials in order to not bias their responses. The experiment was presented using PsychoPy (Peirce, 2007) and stimulus identities were rotated around the conditions across observers. Block order was counterbalanced.

Results

The percentage of accurate responses was calculated for all conditions and compared. This is shown in Figure 2.2, which also illustrates individual performance. A one-factor ANOVA of these data showed an effect of trial type, $F(2,58) = 37.83$, $p < .001$, $\eta_p^2 = .57$, with paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) indicating higher accuracy on same-image identity-match trials ($M = 92.3\%$, $SD = 9.4$) than different-image identity-match trials ($M = 53.3\%$, $SD = 18.3$) and mismatch trials ($M = 64.9\%$, $SD = 18.7$), $t(29) = 13.73$, $p < .001$, $d = 2.65$ and $t(29) = 6.58$, $p < .001$, $d = 1.83$, respectively. The difference in accuracy between different-image identity-match trials and mismatch trials was not reliable, $t(29) = 1.87$, $p = .07$, $d = 0.62$.

Considering the low accuracy for different-image identity-match trials and mismatch trials, a series of one-sample *t*-tests was also conducted to determine whether accuracy was above chance (i.e., 50%) for the conditions. This was the case for same-image identity-matches, $t(29) = 24.79$, $p < .001$, $d = 6.32$, and identity mismatches, $t(29) = 4.38$, $p < .001$, $d = 1.12$, but not for different-image identity-matches, $t(29) = 1.00$, $p = .33$, $d = 0.25$.

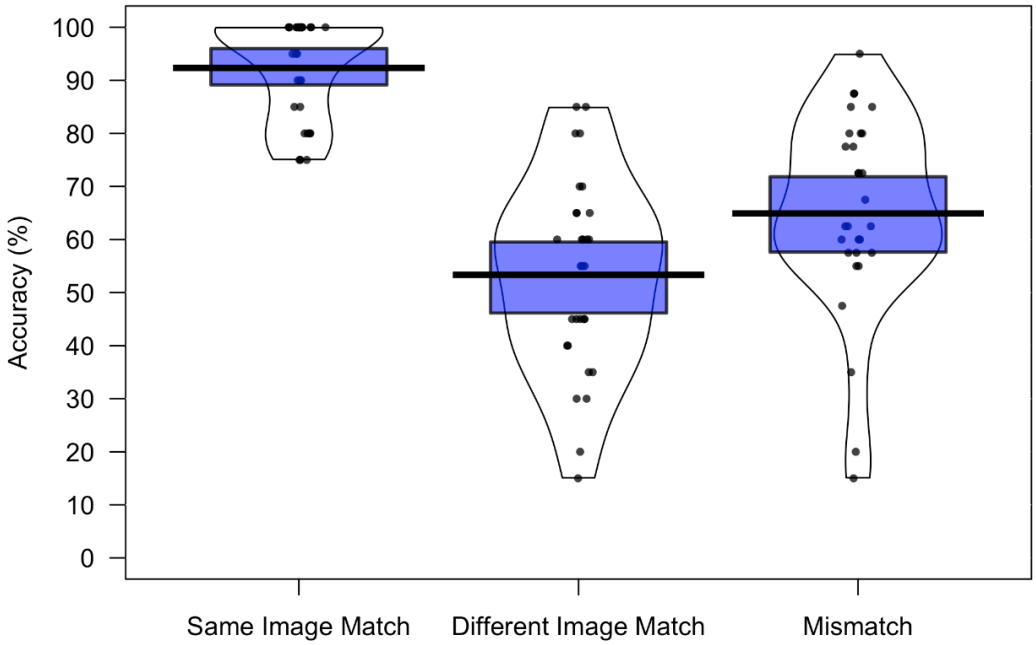

Figure 2.2. Percentage accuracy data for Experiment 1. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

Discussion

This experiment shows that matching of avatar faces to their source face photographs is highly accurate, which indicates that image-specific identity information from these source images is captured well. By contrast, matching of avatar faces to a different photograph of the same person was difficult and did not reliably exceed the chance benchmark of 50%. Accuracy was also fairly low for identity mismatches, comprising pairings of avatar faces with face photographs of a different person. The low accuracy in these conditions is potentially problematic for adopting VR to study unfamiliar face matching, but it is possible that this is caused by the inclusion of same-image identity-matches. Whilst this condition was included here to assess the production process of the stimuli, it is typically not included in face matching experiments (see, e.g., Fysh & Bindemann, 2018). Considering that these same-image stimulus

pairs inevitably display much greater similarity than different-image identity-matches and mismatches, the inclusion of this condition may have served to attenuate the perceived differences between these critical identity conditions, resulting in a reduction in accuracy. To address this possibility, only different-image identity-matches and mismatches were employed in Experiment 2.

Experiment 2

This experiment further assesses whether the production process of the avatars captures the identities on which these are based. In contrast to Experiment 1, this was assessed with only two conditions, comprising different-image identity-matches and identity mismatches, to minimise the influence that same-image identity-matches might exert on the classification of these conditions.

Method

Participants

Thirty Caucasian participants from the University of Kent (10 male, 20 female), with a mean age of 19.6 years (*SD* = 1.5 years), participated in exchange for a small fee or course credit. None of these had participated in Experiment 1.

Stimuli and Procedure

Stimuli, procedure and task instructions were identical to Experiment 1, except that same-image identity-matches were excluded. All observers completed two blocks of 40 trials, comprising 20 different-image identity-matches and 20 mismatches pairs in each block. As was the case in Experiment 1, Block 2 consisted of the reverse image-type stimulus pairings for the identities in Block 1. Once again, all trials began with a 1-second fixation cross and were

presented in a randomised order, block order was counterbalanced, and accuracy of response was emphasised.

Results

The percentage accuracy data for Experiment 2 are illustrated in Figure 2.3. A paired-sample *t*-test of these data showed that accuracy was comparable for different-image identity-match trials (M = 57.9%, SD = 16.4) and mismatch trials (M = 59.3%, SD = 15.4), $t(29) = 0.25$, $p = .80$, $d = 0.08$. In addition, one-sample *t*-tests revealed that performance in both conditions was above the chance level of 50%, with $t(29) = 2.64$, $p = .01$, $d = 0.67$ and $t(29) = 3.28$, $p = .003$, $d = 0.84$ for match and mismatch trials, respectively.

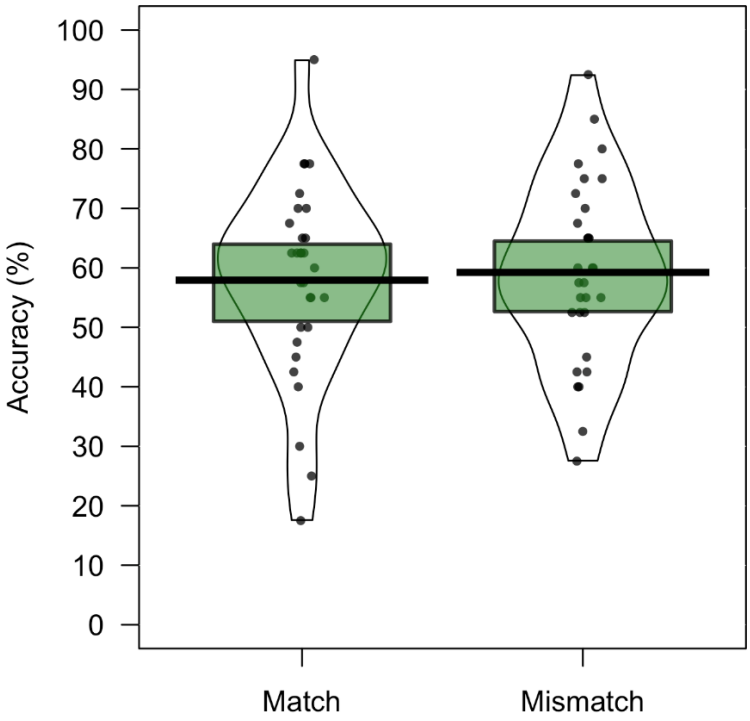

Figure 2.3. Percentage accuracy data for Experiment 2. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

Discussion

Experiment 1 revealed that the avatars capture the face source photographs sufficiently for accuracy on same-image identity-match trials to be high. Experiment 2 complements these findings by showing that accuracy for different-image identity-matches and mismatches exceeds chance when these same-image trials are excluded. Different-image identity-matches are a fundamental requirement for studying the identification of unfamiliar faces, to ensure that this task is not solved by using simple image-matching strategies (see, e.g., Burton, 2013; Jenkins & Burton, 2011). The data from Experiment 2 therefore provide initial evidence that avatar stimuli have the potential to provide a suitable substrate to study face identification processes in VR.

Experiment 3

The two preceding experiments in this initial avatar validation phase have compared avatar face portraits with source photographs. These demonstrate that such avatar portraits capture the facial characteristics of their respective source photographs, and can also be matched to a different photograph from which they were created to an above-chance level. This final validation experiment separates these two image types to investigate whether performance of avatar-to-avatar facial comparisons are consistent with performance of photograph-to-photograph comparisons.

Method

Participants

Thirty Caucasian participants from the University of Kent (1 male, 29 female), with a mean age of 19.2 years (*SD* = 2.0 years), participated in exchange for course credit. None had participated in any of the preceding experiments.

Stimuli and Procedure

The stimuli for this experiment consisted of the same 20 match and 20 mismatch identity pairings of Experiment 2, presented in two blocks (80 trials in total). However, rather than combining an avatar face portrait with a source photograph, avatar face portraits A and B were paired together in one block of trials, while source photographs A and B were paired together in a second block. As with the previous experiments, all trials began with a 1-second fixation cross and were presented in a randomised order. Furthermore, participants were instructed to match for identity rather than image and were not informed about the ratio of match-to-mismatch trials. Block order was counterbalanced across participants, and accuracy of response emphasised.

Results

To compare performance across image type, the mean percentage accuracy of correct match and mismatch responses was calculated for all conditions. These data are illustrated in Figure 2.4. For avatar-to-avatar comparisons, accuracy was higher for match trials (M = 66.2%, SD = 19.1) than mismatch trials (M = 56.0%, SD = 15.4). The opposite pattern was observed for photograph-to-photograph comparison trials, with higher accuracy for mismatch trials (M = 87.0%, SD = 10.3) than for match trials (M = 83.2%, SD = 13.7). A 2 (image type: source photograph, avatar) x 2 (trial type: match, mismatch) within-subjects ANOVA of these data did not show a main effect of trial type, $F(1,29) = 0.55$, $p = .47$, $\eta_p^2 = .02$, but revealed a main effect of image type, $F(1,29) = 219.55$, $p < .001$, $\eta_p^2 = .88$, and an interaction between factors, $F(1,29) = 13.67$, $p < .001$, $\eta_p^2 = .32$. A simple main effect of image type was found for match, $F(1,29) = 54.31$, $p < .001$, $\eta_p^2 = .65$, and mismatch trials, $F(1,29) = 135.51$, $p < .001$, $\eta_p^2 = .82$, due to higher accuracy for photograph than avatar matching. No simple main effects of trial

type were found within avatar matching, $F(1,29) = 3.29$, $p = .08$, $\eta_p^2 = .10$, and photograph matching, $F(1,29) = 1.17$, $p = .29$, $\eta_p^2 = .04$.

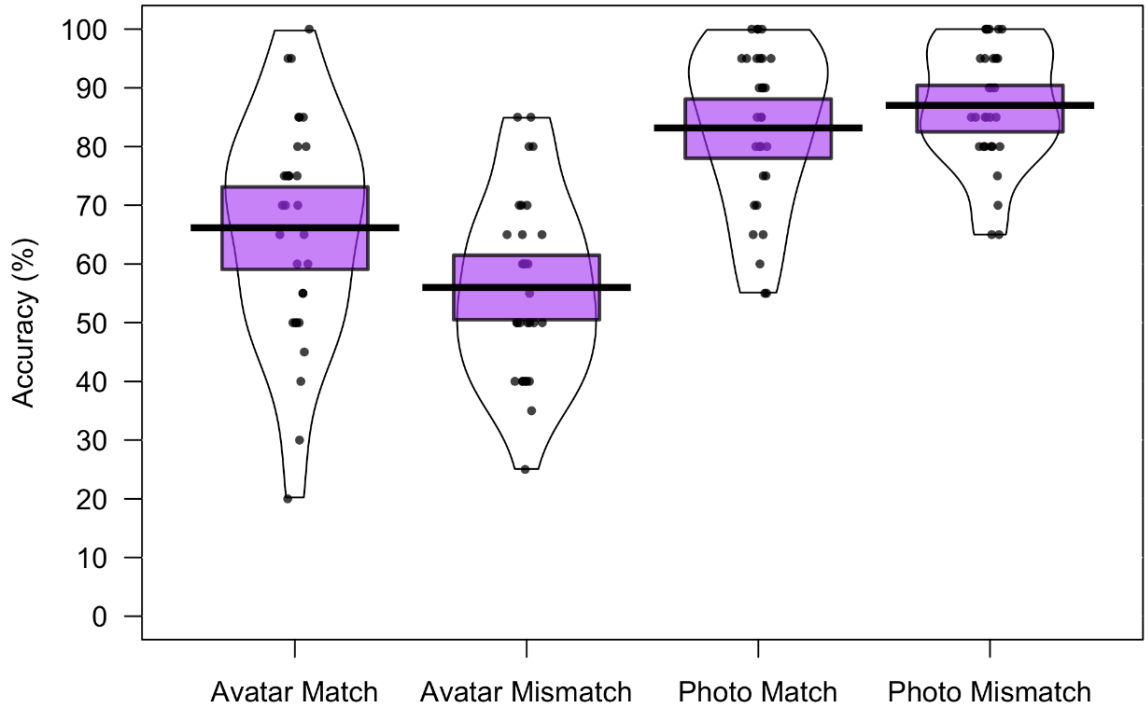

Figure 2.4. Percentage accuracy data for Experiment 3. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

One-sample *t*-tests showed that match and mismatch accuracy for photographs exceeded chance (50%), $t(29) = 13.22$, $p < .001$, $d = 3.37$, and $t(29) = 19.67$, $p < .001$, $d = 5.01$, respectively. Importantly, this was also the case for match and mismatch trials with avatar portraits, $t(29) = 4.62$, $p < .001$, $d = 1.18$ and $t(29) = 2.13$, $p = .04$, $d = 0.54$.

Finally, accuracy for source photographs and avatar faces correlated on both match trials, $r = .752$, $p < .001$, and mismatch trials, $r = .415$, $p < .05$, indicating that matching of both stimulus types reflects the same underlying cognitive processes.

Discussion

In contrast to Experiment 1 and 2, which examined photograph-to-avatar matching, the current validation experiment demonstrates that avatar faces also can be successfully matched to each other. Avatar matching was more difficult than matching pairs of face photographs, but this is unsurprising considering that the photographs reflect the original identity images. In addition, identities for mismatches were paired up based on avatar similarity, which should increase the difficulty of this task relative to matching of photographs also. Despite this, performance for avatar-to-avatar and photograph-to-photograph matching correlated well, indicating that both reflect the same underlying processes. The next phase of this chapter will explore this further, by comparing avatar matching with two established tests of face matching.

Phase 2: Matching avatars versus matching face photographs

The experiments of phase 1 demonstrate that avatar identification is a difficult task, but indicate also that avatar matching reflects similar processes to matching of face photographs. To examine this further prior to implementation in a VR environment, correlations were sought between matching of avatar face pairs with two tests of unfamiliar face matching in phase 2, comprising the widely-used GFMT (Burton et al., 2010) and the newer KFMT (Fysh & Bindemann, 2018). Of these tests, the GFMT represents a best-case scenario to assess face-matching accuracy, by providing highly-controlled, same-day photographic pairs of faces. The KFMT, on the other hand, provides a more challenging matching test, in which face pairs consist of a controlled face portrait and an uncontrolled image. Despite these differences, performance on the GFMT and KFMT correlates well. Here it was investigated whether such correlations exist also between these tests and the matching of avatar face pairs.

Experiment 4

This experiment compared performance on the GFMT and KFMT, which required matching of photographs of faces, with the matching of pairs of avatar faces. Overall, performance should be best with the optimised stimuli of the GFMT than the more challenging KFMT. In addition, accuracy for the KFMT should be similar to avatar-to-avatar face matching, considering that both tests are based on different-day face images. The main aim here, however, was to correlate performance on these tasks to explore whether these capture the same identification processes.

Method

Participants

The participants consisted of 30 Caucasian individuals (8 male, 22 female), with a mean age of 21.2 years (*SD* = 3.3 years), who were paid a small fee or given course credit. None of these had participated in the preceding experiments.

Stimuli and Procedure

The GFMT: The GFMT face pairs consist of images of faces taken from a frontal view displaying a neutral expression. Both images in a face pair are taken with different cameras and, in the case of identity matches, approximately 15 minutes apart. Each face image is cropped to show the head only and converted to greyscale with a resolution of 72 ppi. The dimensions of the faces range in width from 70 mm to 90 mm and in height from 85 mm to 125 mm, and are spaced between 40 mm and 55 mm apart on screen. This study employed 20 identity match and 20 mismatch trials from the GFMT (for more information, see Burton et al., 2010). Example stimuli are shown in the top row of Figure 2.5.

The KFMT: Face pairs in the KFMT consist of an image from a student ID card, presented at a maximal size of 35 mm (w) x 47 mm (h), and a portrait photo, sized at 70 mm (w) x 82 mm (h) at a resolution of 72 ppi, spaced 75 mm apart. The student ID photos were taken at least three months prior to the face portraits and were not constrained by pose, facial expression, or image-capture device. The portrait photos depict the target's head and shoulders from a frontal view whilst bearing a neutral facial expression and were captured with a high-quality digital camera. In this study, 20 identity match and 20 mismatch trials from the KFMT were employed (for more information, see Fysh & Bindemann, 2018). Example stimuli are shown in the second row of Figure 2.5.

Avatar face pairs: These stimuli are the same as those shown in Block 1 of Experiment 3 and consisted of 40 face pairs (20 identity matches, 20 mismatches), each depicting two avatar face portraits. For identity-match trials, the avatar faces in a pair were based on different source photographs, whereas two different identities were shown in identity mismatch pairs. These faces were cropped to remove external features, such as hairstyle, and shown at a size of 70 mm (w) x 90 mm (h) and spaced 50 mm apart. Example stimuli are shown in the third row of Figure 2.5.

These three face-matching tasks (GFMT, KFMT, avatar) were administered in separate blocks of 40 trials, which were presented in a counterbalanced order across participants. The procedure for all tasks was identical and presented using PsychoPy (Peirce, 2007). Thus, each trial begun with a 1-second fixation cross presented on a computer screen and was followed by a face pair, which participants were asked to classify as an identity match or mismatch as accurately as possible. As with the previous experiments, for all tasks participants were

instructed to match for identity rather than image and were not informed about the ratio of match-to-mismatch trials. Trial order was randomised within the blocks.

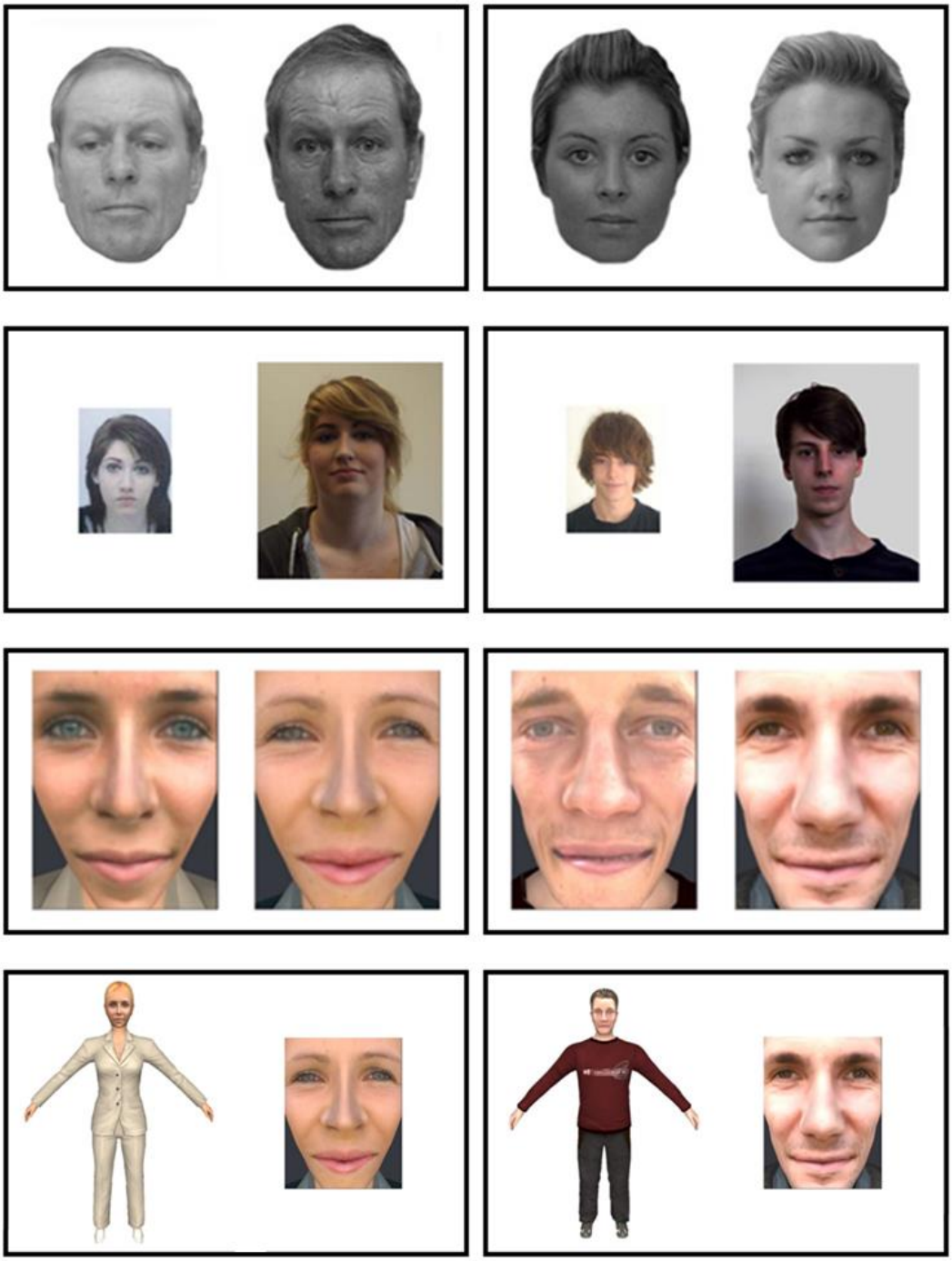

Figure 2.5. Example stimuli of match (left) and mismatch (right) trials for the GFMT (top row), KFMT (second row), avatar face portraits (third row) and whole avatar image to avatar face matching (bottom row).

Results

To compare performance across the three face-matching tasks, the mean percentage of correct match and mismatch responses was calculated for each participant. These data are illustrated in Figure 2.6. For match trials, the cross-subject mean accuracy was higher for the GFMT (*M* = 78.7%, *SD* = 13.2) than the KFMT (*M* = 67.8%, *SD* = 14.6) and the avatar face pairs (*M* = 68.7%, *SD* = 13.3). The same pattern was observed for mismatch trials, with higher accuracy for the GFMT (*M* = 71.8%, *SD* = 18.4) than the KFMT (*M* = 59.0%, *SD* = 14.4) and the avatar face pairs (*M* = 52.5%, *SD* = 16.6).

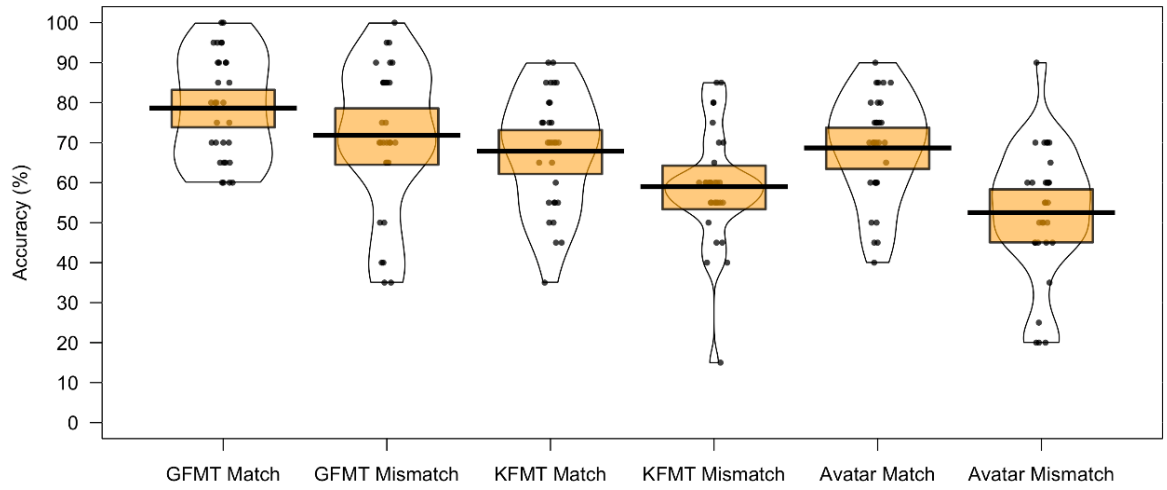

Figure 2.6. Percentage accuracy data for the GFMT, KFMT and avatar face pairs in Experiment 4. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

A 3 (task: GFMT, KFMT, avatar) x 2 (trial type: match, mismatch) within-subjects ANOVA of these data confirmed a main effect of trial type, $F(1,29) = 8.83$, $p = .006$, $\eta_p^2 = .23$, due to higher accuracy on match than mismatch trials. A main effect of task was also found,

$F(2,58) = 34.70$, $p < .001$, $\eta_p^2 = .55$. Paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) showed that accuracy was higher on the GFMT than both the KFMT, $t(29) = 6.09$, $p < .001$, $d = 1.24$, and the avatar pairs, $t(29) = 7.87$, $p < .001$, $d = 1.57$. There was no difference in accuracy between the KFMT and avatar pairs, $t(29) = 1.58$, $p = .13$, $d = 0.32$. The interaction of task and trial type was not significant, $F(2,58) = 2.35$, $p = .11$, $\eta_p^2 = .08$.

A series of one-sample *t*-tests was also conducted to determine whether accuracy was above chance (i.e., 50%) for the conditions. This was the case for match and mismatch trials on the KFMT, $t(29) = 6.69$, $p < .001$, $d = 1.70$ and $t(29) = 3.42$, $p = .002$, $d = 0.87$, and on the GFMT, $t(29) = 11.90$, $p < .001$, $d = 3.03$ and $t(29) = 6.51$, $p < .001$, $d = 1.66$. For avatar face pairs, accuracy was also above chance for match trials, $t(29) = 7.68$, $p < .001$, $d = 1.96$, but not for mismatch trials, $t(29) = 0.83$, $p = .42$, $d = 0.21$. A by-item inspection of these data shows a very broad range in accuracy for avatar mismatch face pairs, which suggests that mean chance performance masks items that are consistently classified correctly and also items that are classified consistently as incorrect. Further analysis of these data is returned to after Experiment 7, to demonstrate that these by-item differences for avatar stimuli are stable.

Overall, the mean percentage accuracy data show that accuracy on the GFMT is higher than for the KFMT and the avatar faces, which appear to be more evenly matched. While such general differences between these tasks were expected, the question of main interest in this experiment was whether performance on these tests is correlated. For match trials, Pearson's correlations were obtained for the GFMT and KFMT, $r = .580$, $p < .001$, the GFMT and the avatar faces, $r = .406$, $p = .03$, and the KFMT and the avatar faces, $r = .336$, $p = .05$. Similarly, mismatch accuracy correlated for the GFMT and avatar faces, $r = .550$, $p = .002$, and the KFMT and the avatar faces, $r = .407$, $p = .03$. The correlation for mismatch trials on the GFMT and the KFMT did not reach significance, $r = .333$, $p = .07$.

Discussion

This experiment correlated matching of avatar faces directly with two laboratory tests of face matching to determine whether identification of the avatars taps into the same processes as identification of real faces. Overall, accuracy was best with the highly-optimised face pairs of the GFMT, and comparable for the KFMT and the avatar faces. This finding makes good sense considering that the stimuli of the KFMT and those that were used to create the avatar face pairs captured identities across different days and more variable ambient conditions. Moreover, the similarity in performance across these tests suggest that low accuracy with the avatars reflects a difficulty in face matching that is comparable to the matching of challenging different-day face pairs (see Fysh & Bindemann, 2018; see also Megreya, Sandford, & Burton, 2013). Despite these differences in accuracy between the GFMT, KFMT and the avatar faces, performance correlated well across the three tasks. This indicates that such avatar face pairs can provide a substitute to the matching of real faces for experimentation in virtual reality.

Experiment 5

The preceding experiments examine the matching of isolated face pairs. In contrast, identity matching in the VR environment requires comparison of a *person* with a face photograph. The inclusion of such body information reduces face size. This may affect identification, though it is unclear whether this would attenuate (see, e.g., Bindemann, Fysh, Sage, Douglas, & Tummon, 2017) or improve accuracy (see Bindemann, Attard, Leach, & Johnston, 2013). To explore this question under strictly controlled conditions, a further experiment was conducted in which the avatar matching stimuli comprised a whole person and a face photograph. As in Experiment 4, performance on this task was also compared with the GFMT and KFMT.

Method

Participants

Thirty Caucasian participants from the University of Kent (11 male, 19 female), with a mean age of 21.0 years (*SD* = 2.9 years), participated for a small fee or course credit. None had participated in any of the preceding experiments.

Stimuli and Procedure

Stimuli, procedure and task instructions were identical to Experiment 4, except for the following changes. The avatar matching stimuli comprised the same identities but now consisted of the image of a whole avatar (i.e., showing the entire body and the face) and an avatar face (for an illustration, see the bottom row of Figure 2.5). The whole avatar was sized to a height of 155 mm, with a body width of 35 mm (from hand to hand, 115 mm). This resulted in the face on the whole avatar to have dimensions of 20 mm (w) x 30 mm (h). By comparison, the isolated avatar face image in each stimulus pair measured 70 mm (w) x 90 mm (h) and was presented 30 mm apart from the whole avatar.

Results

The percentage accuracy data for this experiment are presented in Figure 2.7. For match trials, accuracy was higher for the GFMT (*M* = 89.3%, *SD* = 10.1) than the KFMT (*M* = 66.5%, *SD* = 20.5) and the avatar stimulus pairs (*M* = 53.8%, *SD* = 18.1). This pattern was also observed with identity mismatches, with highest accuracy for GFMT pairs (*M* = 72.7%, *SD* = 23.6), followed by the KFMT (*M* = 67.2%, *SD* = 15.4) and the avatar pairs (*M* = 52.2%, *SD* = 15.1).

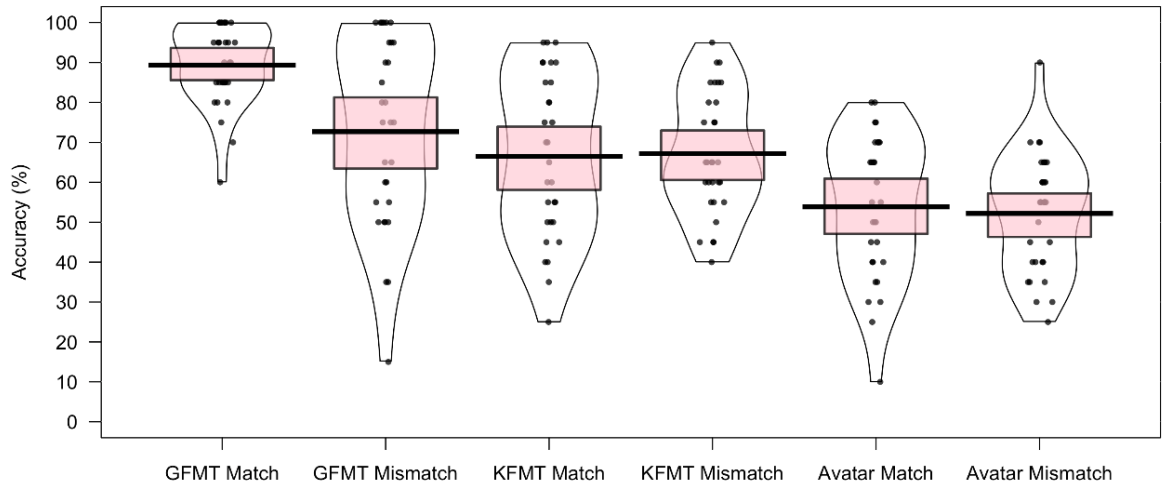

Figure 2.7. Percentage accuracy data for the GFMT, KFMT and avatar stimulus pairs in Experiment 5. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

A 3 (task: GFMT, KFMT, avatar) x 2 (trial type: match, mismatch) within-subjects ANOVA did not reveal a main effect of trial type, $F(1,29) = 1.47$, $p = .24$, $\eta_p^2 = .05$, but showed a main effect of task, $F(2,58) = 75.27$, $p < .001$, $\eta_p^2 = .72$, and an interaction, $F(2,58) = 9.32$, $p < .001$, $\eta_p^2 = .24$. Simple main effects analysis was carried out to interpret this interaction. A simple main effect of trial type within the GFMT task was found, $F(1,29) = 9.53$, $p = .004$, $\eta_p^2 = .25$, due to higher match than mismatch accuracy. There was no simple main effect of trial type within the KFMT, $F(1,29) = 0.01$, $p = .91$, $\eta_p^2 < .01$, or avatar tasks, $F(1,29) = 0.10$, $p = .76$, $\eta_p^2 < .01$.

In addition, a simple main effect of task within match trials was found, $F(2,28) = 98.89$, $p < .001$, $\eta_p^2 = .88$. Paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) showed accuracy on the GFMT was higher than for both the KFMT and the avatar task on match trials, $t(29) = 7.51$, $p < .001$, $d = 1.39$ and $t(29) = 13.39$, $p < .001$, $d = 2.39$

respectively. The KFMT was also performed more accurately than the avatar task on match trials, $t(29) = 3.49$, $p = .002$, $d = 0.65$.

Similarly, a simple main effect of task within mismatch trials was also found, $F(2,28) = 32.84$, $p < .001$, $\eta_p^2 = .70$. Paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) showed accuracy was higher on the GFMT and KFMT than the avatar task for this trial type, $t(29) = 6.48$, $p < .001$, $d = 1.02$ and $t(29) = 5.99$, $p < .001$, $d = 0.97$, respectively. There was no difference in mismatch trial accuracy between the GFMT and KFMT, $t(29) = 1.47$, $p = .15$, $d = 0.27$.

Finally, a series of one-sample *t*-tests was also conducted to determine whether accuracy was above chance (i.e., 50%) for the conditions. This was the case for match and mismatch trials on the GFMT, $t(29) = 21.41$, $p < .001$, $d = 5.46$ and $t(29) = 5.26$, $p < .001$, $d = 1.34$, and the KFMT, $t(29) = 4.41$, $p < .001$, $d = 1.12$ and $t(29) = 6.13$, $p < .001$, $d = 1.56$. In contrast, accuracy for the avatar pairs did not exceed chance for match trials, $t(29) = 1.16$, $p = .26$, $d = 0.30$, nor mismatch trials, $t(29) = 0.79$, $p = .43$, $d = 0.20$. However, a by-item inspection of these data again shows a very broad range in accuracy, suggests that mean performance masks consistent correct and incorrect classifications of avatar items (further analysis provided after Experiment 7). Moreover, Pearson correlations revealed that match accuracy correlated across all combinations of the GFMT, KFMT and the avatar stimuli, all rs $\geq .474$, all ps $\leq .008$, as did accuracy for mismatch trials, all rs $\geq .514$, all ps $\leq .004$.

Discussion

This experiment replicates the main findings of Experiment 4, by revealing that performance for matching GFMT, KFMT and avatar faces correlates consistently. This provides further evidence that identification across these tasks is based on similar processes. However, in contrast to Experiment 4, which displayed only avatar faces, matching avatar faces

to whole persons was more difficult in Experiment 5 and accuracy was low. This poor performance is attributed to the size of the whole body stimuli, which resulted in a compression of the facial information (see bottom row of Figure 2.5). This raises the question of whether these avatars provide sufficient information for person identification during immersion in a VR airport environment. This was examined in the final phase of this chapter.

Phase 3: Face matching in virtual reality

In the final phase, avatar identification was examined in virtual reality (VR), by constructing a passport control desk in an airport arrivals hall. This environment comprised an airport lounge, with seating and rope queue barriers to channel passengers to a passport control booth. Visual cues were incorporated to convey clearly to participants that this is an airport environment, such as departure boards and a waiting aeroplane within view of the passport control desk area. This environment is illustrated in Figure 2.8.

Participants were immersed in this environment and asked to take on the role of passport officers in the control booth, by processing a queue of passengers by identity-matching a face photograph to an avatar's appearance (see inset of Figure 2.8). Animated avatars queued in line and then approached the booth individually to be processed. After participants made an identification decision, the avatar would then walk away, with stimuli classified as identity matches proceeding past the booth and towards an exit at the back of the airport hall, whilst stimuli classified as mismatches would walk into a waiting area to the side of the control point.

Figure 2.8. An overhead view of the virtual reality airport. Inset (bottom right) displays the viewpoint of the participants from the passport control booth, when processing the queue in Experiment 6.

Experiment 6

In Experiment 6, this airport environment was employed to investigate face matching in VR. The same avatar identities as in the preceding experiments were employed and specifically sought to examine the accuracy levels that participants achieve in this task.

Method

Participants

Thirty Caucasian participants from the University of Kent (7 male, 23 female), with a mean age of 21.6 years (*SD* = 4.1 years), took part for a small fee or course credit. None had participated in the preceding experiments. Owing to the use of virtual reality equipment, no persons with epilepsy or who were liable to motion sickness were recruited. Before immersion

in VR, participants were briefed about potential side effects of using VR, such as discomfort from wearing the headset and symptoms of motion sickness, and health and safety procedures.

Stimuli and Procedure

The stimuli consisted of the same avatar-face pairings that were employed in Experiment 5, comprising 20 matches and 20 mismatches. These were displayed in the VR environment using Vizard 5 and an Oculus Rift DK2 headset, with a resolution of 960 x 1080 pixels per eye with 100° field of view and an image refresh rate of 75 Hz.

On immersion in the VR environment, participants found themselves seated in the passport control booth, which was equipped with a desk and desktop PC. A group of 40 avatars then arrived in the airport hall and queued at the control desk, with one avatar at a time approaching the participants. As each avatar approached, their 'passport photograph' would appear on the screen of the desktop PC. Participants were asked to compare this image with the face of the presenting avatar, and make identity match or mismatch decisions via button presses on a computer mouse. As with the previous experiments, participants were not informed about the ratio of match-to-mismatch trials and were instructed to match for identity rather than image. In this instance, participants were asked to imagine how they would vary to their own passport as an example of how someone could look different to their purported passport image and yet still be the same person. Once a response was registered, the avatar would move past the control desk to exit the airport hall (if classified as a match) or would depart to the side of the airport hall into a waiting area (if classified as a mismatch). At this point, the next avatar would approach the control desk, prompting the start of the next trial. Presentation of avatars was randomised. Accuracy of response was emphasized, and there was no time restriction for task completion.

Results

The percentage accuracy data for this VR experiment are illustrated in Figure 2.9. A paired sample *t*-test showed that accuracy was higher on match trials (M = 59.3%, SD = 13.0) than mismatch trials (M = 39.2%, SD = 12.0), $t(29) = 5.29$, $p < .001$, $d = 1.59$. In addition, one-sample *t*-tests showed that performance was above chance (50%) on match trials, $t(29) = 3.94$, $p < .001$, $d = 1.00$, but below chance on mismatch trials, $t(29) = 4.93$, $p < .001$, $d = 1.26$. However, by-item inspection of these data again shows a very broad range in accuracy for mismatch stimuli (further analysis provided after Experiment 7).

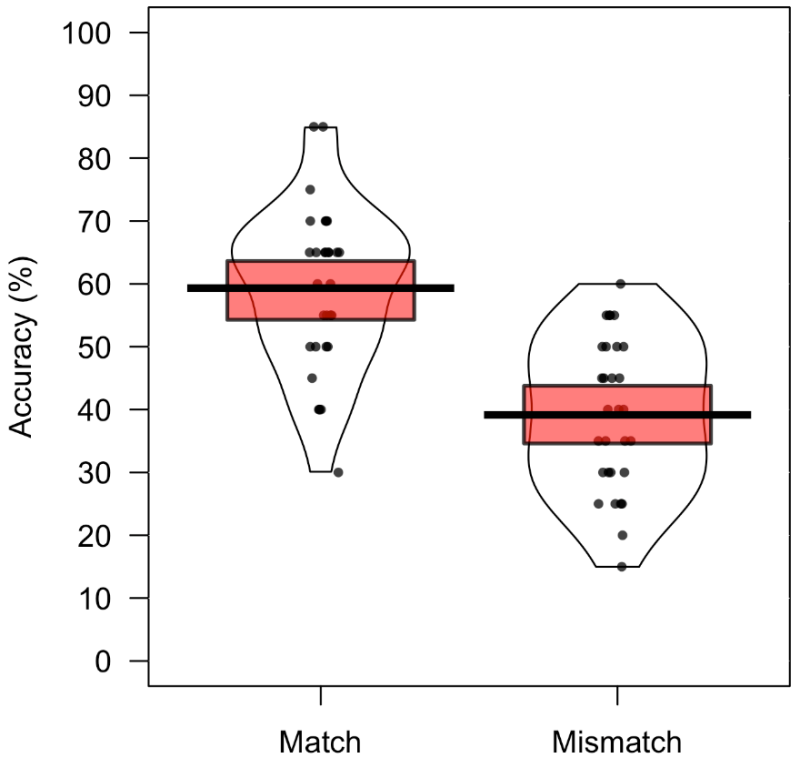

Figure 2.9. Percentage accuracy data for Experiment 6. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

Cross-experiment analyses were conducted to examine how performance for this face-to-avatar matching in VR compared to the still image avatar matching of Experiment 4 (face-to-face matching: match accuracy M = 68.7%, SD = 13.3; mismatch accuracy M = 52.5%, SD = 16.6) and Experiment 5 (face-to-body matching: match accuracy M = 53.8%, SD = 18.1; mismatch accuracy M = 52.2%, SD = 15.1). A 3 (stimulus type: face-to-face, face-to-body, face-to-avatar) x 2 (trial type: match, mismatch) mixed-factor ANOVA showed main effects of trial type, $F(1,87) = 22.73$, $p < .001$, $\eta_p^2 = .21$, and stimulus type, $F(2,87) = 16.25$, $p < .001$, $\eta_p^2 = .27$, and an interaction between these factors, $F(2,87) = 4.47$, $p = .01$, $\eta_p^2 = .09$.

To interpret this interaction, simple main effects analyses were carried out. A simple main effect of trial type was found for face-to-face matching (Experiment 4), $F(1,87) = 12.34$, $p < .001$, $\eta_p^2 = .12$, and face-to-avatar matching (Experiment 6), $F(1,87) = 19.20$, $p < .001$, $\eta_p^2 = .18$, both due to higher match than mismatch accuracy. There was no simple main effect of trial type for face-to-body matching (Experiment 5), $F(1,87) = 0.13$, $p = .72$, $\eta_p^2 < .01$.

In addition, a simple main effect of stimulus type within match trials was found, $F(2,87) = 7.52$, $p < .001$, $\eta_p^2 = .15$. Paired-samples t-tests (with alpha corrected to .017 [.05/3] for three comparisons) showed that face-to-face matching was performed more accurately than both face-to-body matching, $t(58) = 3.62$, $p < .001$, $d = 0.92$, and face-to-avatar matching, $t(58) = 2.75$, $p = .008$, $d = 0.70$. There was no difference in accuracy between these latter two stimulus types on match trials, $t(58) = 1.35$, $p = .18$, $d = 0.34$.

A simple main effect of stimulus type within mismatch trials was also found, $F(2,87) = 8.02$, $p < .001$, $\eta_p^2 = .16$. Paired-samples t-tests (with alpha corrected to .017 [.05/3] for three comparisons) showed accuracy was higher for both face-to-face and face-to-body matching over face-to-avatar matching, $t(58) = 3.56$, $p < .001$, $d = 0.91$ and $t(58) = 3.68$, $p < .001$, $d = 0.94$ respectively. No difference in accuracy was found between face-to-face and face-to-body matching on mismatch trials, $t(58) = 0.08$, $p = .94$, $d = 0.02$.

Discussion

The results from this experiment indicate an increase in task difficulty when face matching is performed in VR. The accuracy of avatar matching, particularly on mismatch trials, was considerably lower in the VR environment than when the same stimuli were presented in 2D and in isolation in Experiments 4 and 5. Considering this low accuracy, the paradigm was modified for a final experiment in an attempt to improve performance.

Experiment 7

In this experiment, it was attempted to optimise the VR paradigm to improve face-matching performance. The Oculus Rift DK2 headset was replaced with an HTC Vive, which provides greater screen resolution (960 x 1080 pixels per eye versus 1080 x 1200 pixels per eye). The HTC Vive is also equipped with handheld controllers to enable participants to interact better with the environment. The controllers were utilised to allow participants to hold the passports of passengers in the VR environment. This enabled participants to bring these closer to their own face, thus increasing the size and resolution of these images for comparison, as well as to hold the passport photos next to the passengers to facilitate face matching (see Figure 2.10). As a final change, the face image for the photo-identities in VR were re-recorded. The software models convexity by elongating face shape as viewing distance decreases. As a result of this, the avatar face stimuli were narrow in appearance in the preceding experiments, particularly near the chin region. These images were re-recorded from greater distance to produce a more natural, rounded appearance (see inset of Figure 2.10). It was then examined whether face-matching performance in the VR environment was improved as a result of these changes.

Method

Participants

Thirty Caucasian participants from the University of Kent (7 male, 23 female) with a mean age of 20.3 years (*SD* = 2.8 years) participated for a small fee or course credit. None had participated in the preceding experiments. No persons with epilepsy or who were liable to motion sickness were recruited. All participants were given a health and safety briefing prior to immersion in the VR.

Stimuli and Procedure

The stimuli consisted of the same avatar identities as in Experiment 6, but the images for the passport photographs were re-recorded at a great viewing distance to produce faces with a more natural, rounded face shape (see inset of Figure 2.10). The size of these images was maintained at 438 (w) x 563 (h) pixels at a resolution of 150 ppi. The procedure was identical to Experiment 6 except that the Oculus Rift DK2 headset was replaced with an HTC Vive, which has an improved resolution of 1080 x 1200 pixels per eye with 110° field of view with a faster image refresh rate of 90Hz. In addition, two handheld controllers were utilised as controls for this experiment.

On each trial, the passport face image was no longer presented on the desktop PC in the control booth but was inserted into a passport-style card, which could be picked up by participants using a hand-held controller. This enabled participants to hold the passport images closer to their own eyes or next to the avatar's head to facilitate identity comparison. The hand-held controllers were also employed to record participants' responses, with button presses on the right-hand controller indicating identity matches and on the left-hand controller indicating mismatches. All other instructions given to participants were the same as in Experiment 6.

Figure 2.10. Improved interactivity of airport environment in Experiment 7. Inset (top right) displays an avatar face portrait from Experiment 6 (left) alongside its updated image for Experiment 7 (right).

Results

As in all preceding experiments, accuracy was higher for match trials (M = 77.3%, SD = 12.6) than mismatch trials (M = 48.2%, SD = 12.6), $t(29) = 7.28$, $p < .001$, $d = 2.28$, as illustrated in Figure 2.11. In addition, match accuracy was reliably above chance level (i.e., 50%), $t(29) = 11.90$, $p < .001$, $d = 3.03$, whereas mismatch accuracy was not, $t(29) = 0.80$, $p = .43$, $d = 0.20$. Again, however, by-item inspection of the mismatch data shows broad differences between items (further analysis provided after this experiment).

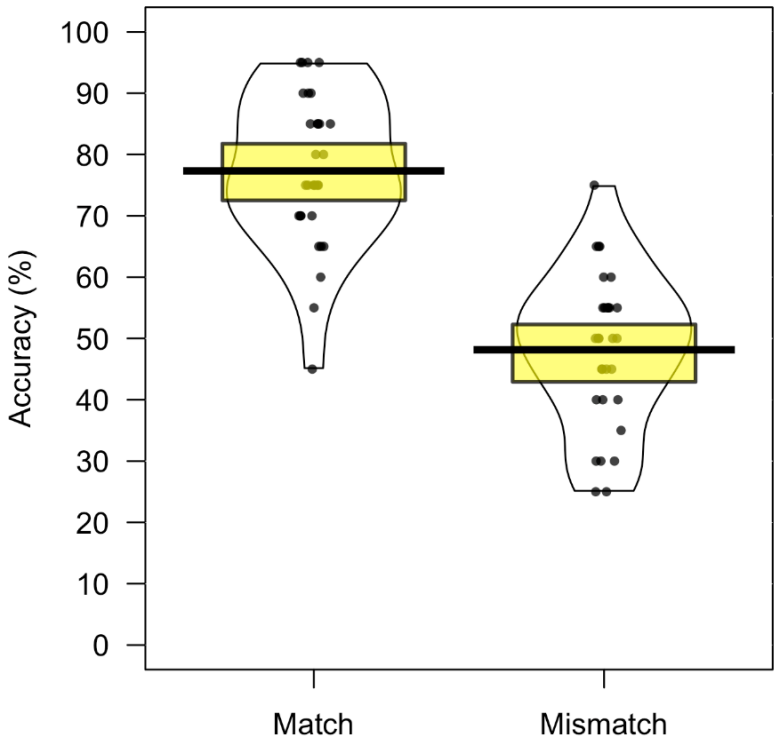

Figure 2.11. Percentage accuracy data for Experiment 7. The mean performance of each trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represent the accuracy of individual participants. The width of each violin represents the expected probability density of performance.

To determine whether the adjustments to the VR paradigm successfully reduced the difficulty of the task, a 2 (environment: Experiment 6, Experiment 7) x 2 (trial type: match, mismatch) mixed-factor ANOVA was conducted. This showed a main effect of trial type, $F(1,58) = 79.67$, $p < .001$, $\eta_p^2 = .58$, due to higher accuracy on match trials than mismatch trials. A main effect of environment was also found, $F(1,58) = 63.27$, $p < .001$, $\eta_p^2 = .52$, reflecting higher accuracy in Experiment 7. The interaction between trial type and experiment was not significant, $F(1,58) = 2.65$, $p = .11$, $\eta_p^2 = .04$.

Discussion

This experiment demonstrates that the improvements to the VR paradigm enhanced accuracy. This improvement was particularly marked on match trials, where accuracy reached 77%. Mismatch performance was enhanced too but remained particularly difficult in the VR paradigm, at 48% accuracy. This is a limiting factor for research on unfamiliar face matching, considering the important role that these trials hold for person identification at passport control in the real-world (see, e.g., Fysh & Bindemann, 2017a). However, previous research on face matching demonstrates that considerable variation in accuracy can exist across items, to the point where some items may be consistently classified incorrectly (see Fysh & Bindemann, 2018). In turn, this raises the possibility that even though mean performance on mismatch trials does not exceed 50%, a substantial proportion of these may nonetheless be classified with high accuracy. A cursory analysis of such by-item differences was provided in Experiments 4 to 7, which revealed broad differences in accuracy between individual items. To explore whether these by-item differences are stable, correlational comparisons across Experiments 4 to 7 were performed.

Comparison of items across experiments

To analyse accuracy for individual items, the mean accuracy for each stimulus pair was compared across experiments (i.e., for face-to-face pairs in Experiment 4, face-to-body in Experiment 5, and face-to-avatar in Experiments 6 and 7). These scores are illustrated in Figure 2.12 and reveal considerable variation in accuracy across items. In Experiment 4, for example, this variation is such that accuracy for individual match items ranges from 40% to 93%, and from 20% to 90% for mismatch items. These differences were even more marked by Experiment 7, in which by-item accuracy ranged from 7% to 97% for match stimuli, and from 3% to 97% for mismatch stimuli.

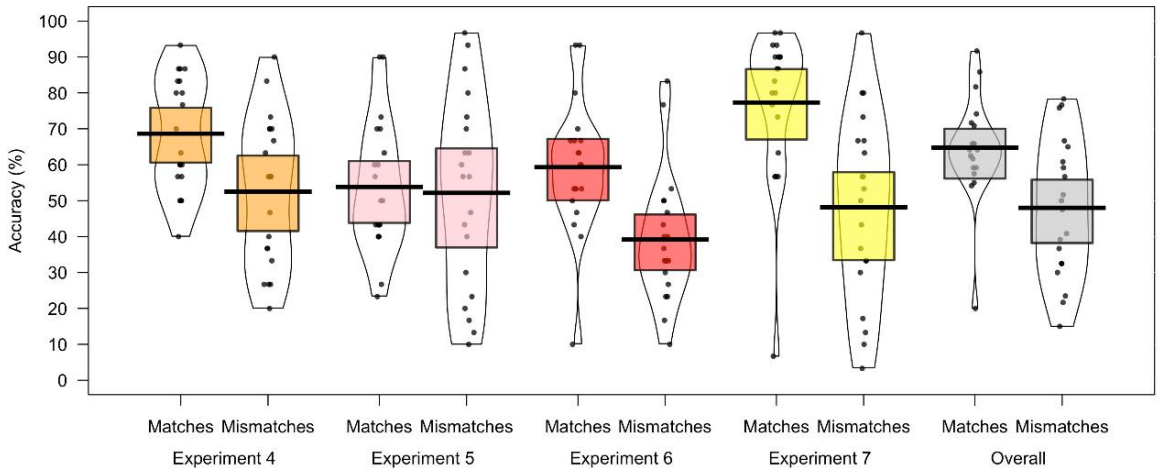

Figure 2.12. Percentage accuracy data by avatar item for Experiments 4 to 7. The mean performance of each avatar trial type is denoted by the black lines with the coloured boxes representing 95% confidence intervals. The black dots represents accuracy for individual face pairs. The width of each violin represents the expected probability density of performance.

This range in accuracy indicates that some items were consistently classified correctly, whereas other yielded consistently incorrect decisions. A reliability analysis was conducted across Experiments 4 to 7, with Cronbach's alpha showing accuracy for match items, $\alpha = .66$, to be more consistent than accuracy for mismatch items, $\alpha = .55$. However, despite the variation in item accuracy, strong positive correlations were obtained for by-item accuracy across Experiments 4 to 7 (see Table 2.1).

For match items, by-item accuracy correlated well for each progression towards face matching in VR. Accuracy when matching two avatar face portraits (Experiment 4) positively correlated with the accuracy of matching one of these avatar face images with an avatar body image (Experiment 5), $r = .499$, $p = .03$. When this avatar face-body matching was conducted in VR (Experiment 6), accuracy correlated with its still image counterpart (Experiment 5), $r = .515$, $p = .02$. Item accuracy in the original VR paradigm (Experiment 6) also correlated strongly with item accuracy when the VR paradigm was improved in Experiment 7, $r = .741$,

$p < .001$. However, all other correlations between experiments were non-significant, all rs ≤ .423, all ps ≥ .06.

Accuracy for many mismatch items was lower than for any of the match items across all experiments, but correlated strongly across all comparisons between Experiments 4 to 7, all rs ≥ .566, all ps < .009, except between the two VR experiments (Experiments 6 and 7), $r = .342$, $p = .14$. This discrepancy is attributed to the improvement gains possible from Experiment 6 to Experiment 7, which was much greater for some items compared to others.

Table 2.1. Mean accuracy and correlations between experiments across all avatar items

				Correlation coefficients (*r*)			
Trial Type	Experiment	Mean	SD	4	5	6	7
Overall	4	60.7	19.9	-			
	5	53.0	22.9	.552***	-		
	6	49.2	20.7	.539***	.484**	-	
	7	62.8	27.7	.627***	.553***	.647***	-
Match	4	68.7	15.8	-			
	5	53.8	18.4	.499*	-		
	6	59.3	18.5	.255	.515*	-	
	7	77.4	21.2	.394	.423	.741***	-
Mismatch	4	52.6	20.7	-			
	5	52.2	27.1	.639**	-		
	6	39.1	17.9	.566**	.571**	-	
	7	48.2	25.9	.613**	.752***	.342	-

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Overall, the finding that accuracy for items is highly consistent across experiments under the conditions investigated here provides a potential solution to the poor mean accuracy in the mismatch condition. To model the real world of passport control, match trials should occur with much greater frequency than mismatch trials in experiments on unfamiliar face matching

(see, e.g., Bindemann, Avetisyan, & Blackwell, 2010; Fysh & Bindemann, 2017b, 2018; Papesh & Goldinger, 2014; Susa, Michael, Dessenberger, & Meissner, 2019). One way to address the poor mean accuracy across mismatch items in VR here could therefore be to select the mismatches with the highest by-item accuracy for further experimentation. Ultimately, however, it is thought that this problem will be addressed also through future development of higher-quality avatars, which will enhance accuracy of avatar facial identification.

General Discussion

This chapter explored the feasibility of conducting face matching experiments in VR. This investigation is the first of its kind in this field and was conducted in three phases. The first phase investigated whether avatar faces can provide suitable replacements for face photographs, by asking participants to perform avatar-to-photograph identity matching. Accuracy was high when stimuli displayed avatar faces alongside the photograph from which these were derived (Experiment 1). This image-specific identity matching indicates that the avatars successfully captured their source face photograph. Matching accuracy also exceeded chance on mismatch trials, in which two different identities were shown (Experiments 1 and 2), and with different-image identity-matches, in which an avatar face was shown alongside a different source photograph of the same identity (Experiment 2). This indicates that the avatars captured not only the source image but also the identity of these targets. The final validation experiment in this first phase investigated whether accuracy when matching avatar-to-avatar would be consistent with the matching of pairs of photographs (Experiment 3). Despite avatar matching being a more difficult task than photograph matching, participant accuracy exceeded chance and correlated for the two image types. The experiments in this phase therefore demonstrate that the avatar stimuli can provide a suitable substrate to study such face identification processes in VR.

The second phase sought to validate the avatar stimuli further by correlating performance in avatar-to-avatar matching with two established tests of face-to-face matching (the GFMT, see Burton et al., 2010; and the KFMT, see Fysh & Bindemann, 2018). Avatar matching correlated consistently with these face tests, both when pairs of avatar faces were shown (Experiment 4) and when an avatar face was paired with a whole avatar body (Experiment 5). This indicates that matching of avatars and of real face photographs reflect similar cognitive processes.

In the final phase, avatar identification was examined with a VR airport environment, in which participants took up the role of passport officer at a control point. A first run of this paradigm proved difficult, with average accuracy for identity mismatch trials below chance level (Experiment 6). The application of higher-resolution VR equipment, and modifications to the experimental paradigm that allowed participants to view avatar faces more flexibly, improved accuracy (Experiment 7). However, accuracy on mismatch trials remained near chance. A by-item analysis was therefore performed to determine whether individual mismatch trials were classified consistently. This analysis revealed strong correlations across Experiments 4 to 7, indicating that by-item classification was robust across experiments. This by-item data revealed also that some mismatch trials were classified consistently with low but some also with high accuracy. Considering that mismatches should occur with much lower frequency than match trials when one seeks to mimic real-world conditions (see, e.g., Bindemann et al., 2010; Fysh & Bindemann, 2017b, 2018; Papesh & Goldinger, 2014; Susa et al., 2019), the by-item data could therefore provide a basis for selecting mismatch stimuli that give rise to high (or low) accuracy for further experimentation.

Overall, these data provide proof of principle for the use of VR for face-matching research. Whilst the generation of VR explored here does not yet meet real-world detail, realism, and identification accuracy, the rapid development of this technology provides a

promising outlook for future research. This opens up many avenues for face-matching research, by facilitating the study of new environment and social interaction factors that may be relevant in real-world operational settings. With regards to passport control, for example, it is possible that non-facial cues, such as body language, draw attention to potential impostors and could also support identification decisions (Rice, Phillips, Natu, et al., 2013; Rice, Phillips, & O'Toole, 2013). Similarly, environmental factors, such as the mere presence of passenger queues, might impair identification by exerting time pressure on passport officers (see, e.g., Bindemann et al., 2016; Fysh & Bindemann, 2017b; Wirth & Carbon, 2017). Crowd dynamics, such as animated body language throughout queues might also signal impatience to passport officers and exert further pressure. Crucially, such factors cannot be captured well by current laboratory paradigms and are practically impossible to study in real life owing to the importance of person identification at passport control. The current study demonstrates the feasibility of VR for studying and understanding such phenomena, which can only improve as the technology continues to develop.

It is noted that this study still represents a relatively simple approach for the implementation of such experiments. For example, the avatar faces were created by a rather simplistic process that was based on the superimposition of 2D photographs on existing avatar structures. In future, it is anticipated that the 3D scanning of faces and the rigging of this information into avatars, as well as further development of VR technology will result in person stimuli and environments that provide increasingly closer representations of reality. This should support experimentation by further enhancing identification of identity matches and mismatches.

This chapter has demonstrated the feasibility of VR as a new method for investigating face matching in real-world environments, for example passport control. In these settings staff are required to perform face-matching tasks accurately in order to maintain security. It is clear,

however, that these passport staff widely vary in their ability to perform this task (White, Dunn, et al., 2015; White, Kemp, Jenkins, Matheson, & Burton, 2014). At present there is no formal assessment of face-matching ability for prospective passport control staff, and current laboratory tests do not provide a close correspondence to the real-world task (Ramon, Bobak, & White, 2019). Chapter 3 explores using the VR paradigm developed here, together with a self-report measure and establish laboratory tests of face-processing ability, as an assessment tool for personnel selection by simulating the task recruits would be expected to perform.

Chapter 3:

Simulation of Personnel Selection for Passport Control

Introduction

Chapter 2 explored the feasibility of virtual reality (VR) as a solution to practical difficulties of examining face-matching tasks in the field, such as at passport control. The matching of avatar faces was shown to draw on similar cognitive processes as the matching of real faces, suggesting that when employed in a VR simulation of a real-world task performance should be comparable to what could be expected in reality. This chapter seeks to extend this by investigating whether such simulations can be a useful tool for evaluating face-matching ability and thus for selecting personnel for occupations in which such tasks are commonplace.

Person identification is a critical security measure at airports and borders. In its most common form, this process requires the facial comparison between a person and their purported photographic documentation by a human operator at passport control. A substantial body of psychological research indicates that this task is difficult (for a review, see Fysh & Bindemann, 2017a), even for experienced passport officers (White, Kemp, Jenkins, Matheson, & Burton, 2014; Wirth & Carbon, 2017). One potential approach to improving the accuracy of this process is training in facial identification, but the efficacy of this is questionable. Laboratory-based training approaches, such as face shape classification (Towler, White, & Kemp, 2014) and attention-direction to specific facial features (Megreya & Bindemann, 2018; Towler, White, & Kemp, 2017), appear to be unreliable strategies for generating improvement. Short professional training courses containing such components also do not produce notable accuracy gains in face identification (Towler et al., 2019).

A viable alternative to training may be the selection of observers with an aptitude for facial identification (Bobak, Dowsett, & Bate, 2016; Lander, Bruce, & Bindemann, 2018), as people vary substantially in their ability to identify faces (Bobak, Hancock, & Bate, 2016; Burton, White, & McNeill, 2010; Fysh & Bindemann, 2018). These individual differences appear to have a genetic basis (Wilmer et al., 2010; Zhu et al., 2010) and are stable across tests

(Bate et al., 2018; Fysh & Bindemann, 2018; Noyes, Hill, & O'Toole, 2018). Selecting observers with a high aptitude for facial identity comparison may therefore provide improvements in identification accuracy at passport control (Balsdon, Summersby, Kemp, & White, 2018). Such personnel selection bypasses the training problem and should result in immediate security gains.

This chapter investigates this possibility experimentally. In Experiment 8, facial identification ability is assessed first with three established laboratory tests of face perception. Passport control relies on a process in which memory demands are minimised (i.e. faces do not have to be remembered for later identification). As such, the three laboratory face tests selected also pose low memory requirements, instead focusing on visual discrimination and identity comparison. The Cambridge Face Perception Test (CFPT; Duchaine, Germine, & Nakayama, 2007) requires the sorting of a face set in terms of its similarity to a given target. The Glasgow Face Matching Test (GFMT; Burton et al., 2010) and the Kent Face Matching Test (KFMT; Fysh & Bindemann, 2018) require the visual comparison of paired faces to determine whether these depict the same person or different people. In addition, a self-report measure of face identification ability is included (20-item Prosopagnosia Index, PI20; Shah, Gaule, Sowden, Bird, & Cook, 2015) owing to the easier administration of such measures in applied settings than laboratory tests, as well as evidence that scores on the PI20 correlate with facial identity comparison ability (Shah, Sowden, Gaule, Catmur, & Bird, 2015). Individual performance on these tests is then used to predict a person's identification accuracy at passport control. This is simulated with an immersive VR paradigm, in which participants compare photo-identity cards to the faces of passengers in an airport, which was developed and validated in Chapter 2.

Experiment 8

In this experiment, based on these findings, it was expected that self-report scores of face processing ability on the PI20 would relate positively to performance on the three established laboratory tests (GFMT, KFMT and CFPT). More importantly, if self-reported ability on the PI20 and performance on the three face processing tasks relate to person identification at passport control, then individual scores on these tests should correlate positively with identification accuracy in the VR airport environment. Furthermore, if accuracy on these tests correlates with individual performance on the VR passport control task (VRPC), then it is also important to establish which combination of the self-report test and the established laboratory tests best predicts performance on this task.

Method

Participants

This study was pre-registered (https://osf.io/428zh). Sample size was determined using a power calculation for Pearson's *r* (see www.anzmtg.org/stats/PowerCalculator/). The level of power was set at the convention of .80 with alpha of .05 and calculated with effect sizes reported in previous comparisons of the laboratory tasks used in this experiment (Shah, Sowden, et al., 2015; Fysh & Bindemann, 2018). The sample size required to detect the smallest effect (i.e., the correlation between the KFMT and CFPT, $r = -.34$; Fysh & Bindemann, 2018) was selected. Thus, 66 Caucasian students (10 male, 56 female) from the University of Kent, with a mean age of 20.7 years ($SD = 4.6$ years), were recruited for course credit. All participants reported to have normal or corrected-to-normal vision. No persons with epilepsy or liable to motion sickness were recruited. Before immersion in VR, participants were also briefed about potential side effects, such as discomfort from wearing the headset and symptoms of motion sickness, and health and safety procedures.

On completion of the task, participants were also invited to return for a re-test to examine the reliability of the VRPC task. Twenty-one participants (with a mean age of 21.8 years, *SD* = 7.1 years) agreed to return at least two weeks later (*M* = 33 days, *SD* = 8 days) in exchange for a small fee.

Stimuli

In this experiment, participants completed a self-report measure of face processing ability followed by three laboratory tests and the VRPC task. Examples of these stimuli can be seen in Figure 3.1.

20-item Prosopagnosia Index: Participants first completed the PI20 (Shah, Gaule, et al., 2015), which was presented on a desktop computer using Qualtrics software. The PI20 is a self-report measure consisting of 20 statements about a person's face processing ability, which are rated on 5-point scales. Fifteen items, such as "My face recognition ability is worse than most people", are positively scored, with 5 points for "strongly agree" and 1 point for "strongly disagree". The remaining 5 items, for example "I find it easy to picture individual faces in my mind", are reverse coded. This self-report measure has been validated against a range of established face processing tests (Gray, Bird, & Cook, 2017; Shah, Sowden, et al., 2015), and in different cultures (Ventura, Livingston, & Shah, 2018). These studies typically reveal mild-to-moderate effect sizes, suggesting this self-report provides a general proximate for face recognition ability.

The GFMT: The short version of the GFMT (Burton et al., 2010) was employed. This test consists of 40 trials depicting pairs of faces. These face pairs comprise of 20 identity matches, in which two different images of a person's face are shown side-by-side, and 20

mismatches, in which photographs of different people are shown. In identity matches, the images were taken approximately 15 minutes apart. In all stimuli, the faces are cropped around the head and presented at a width of 350 pixels. On each trial of this test, participants are shown a face pair and are asked to decide whether this depicts the same person (i.e., an identity match) or two different people (an identity mismatch). Responses are normally self-paced with accuracy of response emphasised, as in this experiment. Participants were instructed to match for identity, to ensure they did not rely on simplistic image-matching strategies to complete the task (Burton, 2013), and were not informed about the ratio of match-to-mismatch trials in order to not bias their responses. The test was presented with PsychoPy software (Peirce, 2007) and responses were recorded via button presses on a standard computer keyboard. For further detail, see Burton et al. (2010).

The KFMT: The short version of the KFMT (Fysh & Bindemann, 2018) is similar in composition to the GFMT. It also consists of pairs of face photographs comprising of 20 matches and 20 mismatches, for which same- or different-identity judgements are required. In contrast to the GFMT, the face stimuli in each pair of the KFMT comprise of a highly controlled photograph, in which people are recorded with the same image-capture device, standardised lighting and a neutral expression, and a photograph that was not constrained by such factors. These two images for each person were also taken several months apart, providing a more challenging test of facial identity comparison than the GFMT. As with the GFMT, responses are self-paced with accuracy of response emphasised and participants were instructed to match for identity rather than image, having not been informed about the ratio of match-to-mismatch trials. This test was also presented with PsychoPy software in this experiment and responses were recorded via button presses on a standard computer keyboard. For further detail, see Fysh and Bindemann (2018).

The CFPT: The CFPT (Duchaine et al., 2007) is a pre-build experiment, which is run in Java and has been applied widely in psychology. In this task, participants are presented with six greyscale faces in frontal view which they are required to order by similarity to a greyscale target face in ¾ view using a computer mouse. The six faces are variations of a morph between a frontal view image of the target and another identity. The test comprises of 16 trials, with eight sets of stimuli presented once upright and once inverted. Participants are restricted to 60 seconds to complete each trial. For further detail, see Duchaine et al. (2007).

Virtual Reality Passport Control: In the VRPC task, participants were immersed in a virtual airport by wearing an HTC Vive headset. This headset has a resolution of 1080 x 1200 pixels per eye with a 110 degree field of view, and an image refresh rate of 90Hz. Participants were given two hand-held controllers to interact with the environment and took on the role of a passport control officer, standing in a booth area in an airport faced with a queue of people to process.

The person stimuli consisted of 100 animated 3D avatars, each paired with a 2D face portrait of a second avatar, which was embedded on a passport-style card. The 3D avatars were created by combining 2D photographs of real faces with an avatar from an existing database (see www.kent.ac.uk/psychology/downloads/avatars.pdf). Using graphics software (Artweaver 5), the internal features of a face photograph were mapped onto the features of the avatar's face area, with the edges smoothed and skin colour adjusted to blend the graphics. This process was repeated for each identity to produce a match pair, with a 2D face portrait captured from one avatar to create a passport image, which was sized to 438 x 563 pixels at a resolution of 150 ppi. A more detailed description of this avatar construction process is provided in Chapter 2.

For identity mismatch trials, an avatar was paired with a 2D face portrait of a similar-looking identity, matched for gender and approximate age. To provide a closer proximate to real-world conditions, mismatches occurred with lower frequency than matches (see, e.g., Fysh & Bindemann, 2018). Therefore, of the 100 stimulus pairings, 94 trials consisted of the same person (identity matches) while six trials were of two different people (identity mismatches).

In the VRPC task, each avatar approached the booth area in turn and their passport image appeared on a passport-style card. Participants were able to pick up this card with the controller in their right hand for close inspection. They then decided whether the identity of the person on the card was the same as or different to the avatar, by pressing corresponding buttons on the hand-held controllers (right hand for matches, left for mismatches). The avatar then walked away and the queue moved forward, resulting in presentation of the next avatar and its photo-ID card. Participants were instructed to match for identity rather than image, in this instance asked to imagine how they would vary to their own passport as an example. They were also informed how at passport control the majority of passengers would be a match to their passport and so their task was to detect the small number of mismatches if there were any to be found. This information regarding the likelihood of mismatches here was necessary because the participants could develop an expectation for a similar frequency of match and mismatch trials based on their experience of the matching previous tasks (the GFMT and KFMT), therefore may falsely report mismatches in order to even out their responses when uncertain. Participants continued to process the avatars until the queue was cleared. An illustration of the VRPC can be seen in Figure 3.1. Further detail and validation of this paradigm can be seen in Chapter 2.

Procedure

The experiment consisted of a within-subjects design with each participant completing all five tasks. The PI20 was presented first, followed by the GFMT, KFMT, and CFPT, which

were administered in a counterbalanced order across participants. The final task was the VRPC. On completion, participants were also invited to participate in a two-week follow-up (pre-registered, https://osf.io/cmrb9), which comprised of a repetition of the VRPC only to examine its test-retest reliability.

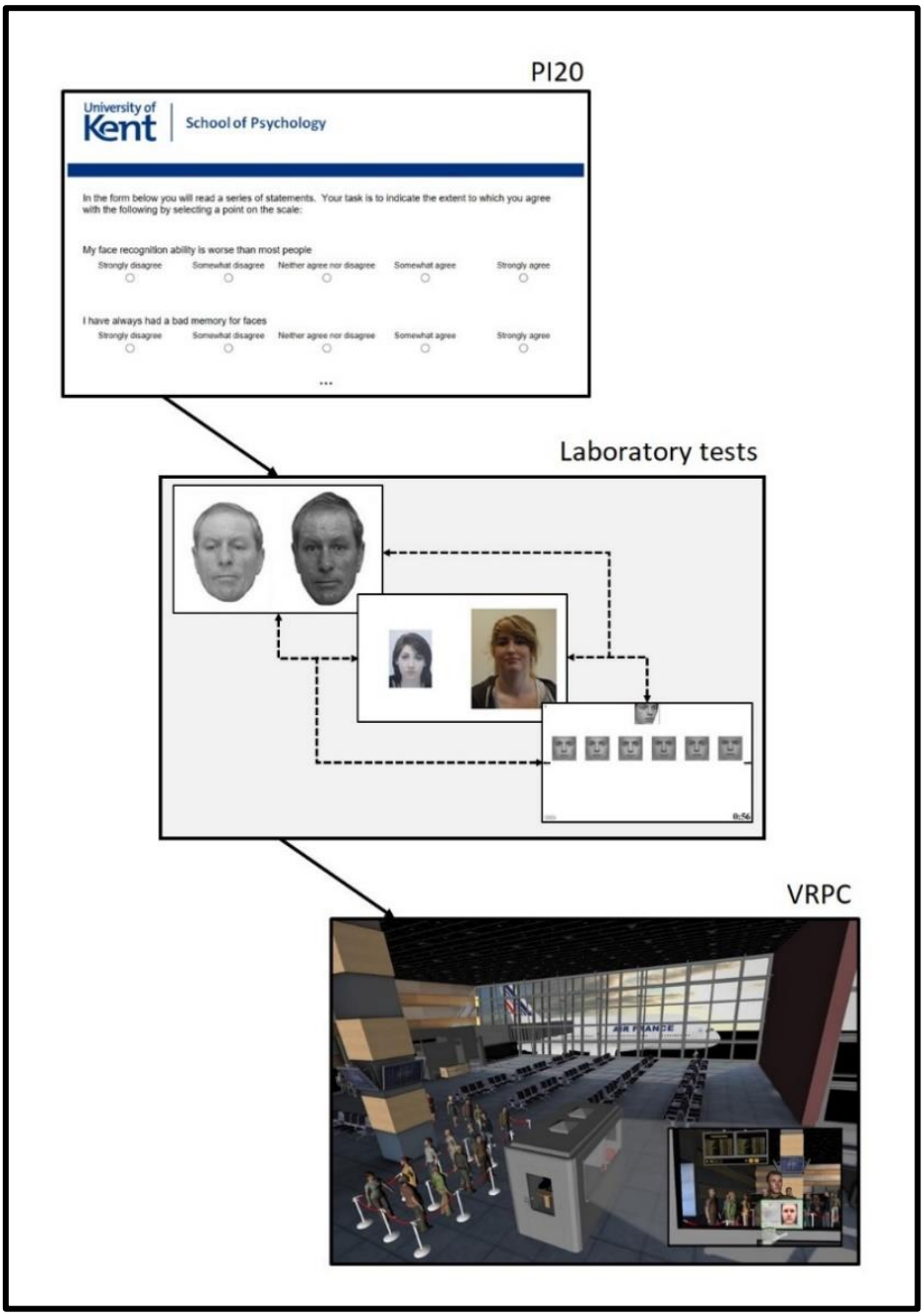

Figure 3.1. Stimuli examples for all tasks, displayed in order of presentation. The order of the GMFT, KFMT and CFPT tasks was fully counterbalanced within the laboratory tests block.

Results

Descriptive statistics

A self-report face recognition ability score was calculated from the PI20, by summing all responses (scoring 15 statements positively, and reverse coding five; as per Shah, Sowden, et al., 2015). For the GFMT and KFMT, the number of correct responses was converted to a percentage score (as per Burton et al., 2010; Fysh & Bindemann, 2018). This was calculated for both match and mismatch trials, and also as an overall accuracy score in order to provide a comparison to the PI20 and CFPT. The CFPT computed the total number of errors made, for example if one face was ranked three places out of order in its sequence then this was classified as three errors (as per Duchaine et al., 2007). The following analyses use the CFPT total error score for the upright face trials only. For the VRPC task, six critical match trials were identified for the analyses, which were accuracy matched to the six mismatch trials prior to the experiment based on data from Chapter 2 (mean accuracy matches = 68.1%, *SD* = 21.7; mismatches = 65.7%, *SD* = 22.0; $t(119) = 0.77$, $p = .44$, $d = 0.11$). As with the GFMT and KFMT scores, the number of correct responses given to the critical match and mismatch trials were converted to percentage scores, and an overall accuracy measure combining these trials was also calculated for comparison to the PI20 and CFPT. The cross-subject means for all of these measures are summarized in Table 3.1.

Table 3.1. Descriptive statistics for all measures. Total scores are presented for the PI20 and CFPT, whilst percentage accuracy is presented for the GFMT, KFMT, and VRPC tasks.

		Mean	*SD*	Min	Max
PI20	Score	42.9	11.3	26	84
GFMT	Match	80.4	17.4	25.0	100.0
	Mismatch	79.3	17.6	25.0	100.0
	Overall	79.8	12.8	52.5	100.0
KFMT	Match	69.8	13.7	40.0	100.0
	Mismatch	68.7	15.0	40.0	100.0
	Overall	69.2	7.9	47.5	85.0
CFPT	Upright Total Deviation Error	38.6	13.4	14	78
VRPC	Critical Match	91.7	12.1	50.0	100.0
	Mismatch	47.5	25.7	0.00	100.0
	Overall	69.6	11.6	50.0	91.7

Note. High face-processing skills are represented by low PI20 and CFPT scores and high GFMT, KFMT, and VRPC matching accuracy scores.

Correlation of Self-Report and Face-Processing Accuracy

To assess whether participants' self-report of face-processing ability related to actual performance, PI20 scores were correlated with accuracy scores on the laboratory tests. For GFMT and KFMT, this analysis was conducted with match, mismatch and overall accuracy scores (with alpha corrected to .017 [.05/3] for three correlations). The PI20 score correlated negatively with overall accuracy, r = -.386, p = .001, and match accuracy on the GFMT, r = -.332, p = .007, indicating that those who self-reported fewer problems with face recognition also performed better on this laboratory test. A trend in this direction was also observed for mismatch trials, r = -.235, p = .06.

Overall accuracy on the KFMT also correlated negatively with the PI20 score, r = -.342, p = .005, again indicating that observers exhibited some insight into their face recognition ability that translated into performance on this matching test. Similarly to the GFMT, a comparable correlation was observed for mismatch trials but did not survive correction for multiple comparisons, r = -.245, p = .05. No correlation was found between the PI20 score and KFMT match trials, r = -.127, p = .31.

Finally, a positive correlation was also observed between the PI20 score and the CFPT total errors, r = .433, p < .001, indicating that those who reported fewer problems with face recognition were also better in the visual discrimination of faces on this laboratory test.

In contrast to these laboratory face tests, no correlations were found between the PI20 score and accuracy in the VRPC task for identity matches, r = .048, p = .70, mismatches, r = .067, p = .59, nor overall accuracy, r = .099, p = .43.

Correlation of Face Tests

In order to assess whether performance on the laboratory face tests reflects similar underlying abilities, correlational analyses were conducted to compare performance on the GFMT and KFMT (with alpha corrected to .017 [.05/3] for the three correlations of trial type and overall accuracy), and the CFPT. For GFMT and KFMT, overall accuracy, r = .596, p < .001, and performance on match trials, r = .558, p < .001, and mismatch trials, r = .558, p < .001, were positively correlated, indicating consistency in participants' performance across these tests. Overall accuracy on the GFMT and KFMT also correlated negatively with the CFPT, r = -.425, p < .001 and r = -.405, p < .001, respectively. This indicates that observers who made fewer discrimination errors on the CFPT were also more accurate at face matching. Overall, this analysis converges with previous work to demonstrate correlation of individual performance across these three face tests (e.g., Fysh & Bindemann, 2018).

Correlation of Face Tests with VRPC

In order to assess whether accuracy on the laboratory face tests relates to person identification in more complex environments, performance on the GFMT and KFMT was correlated with the VRPC task for both match and mismatch trials (with alpha corrected to .017 [.05/3] for three comparisons between each trial type), and also with the CFPT total upright error score for overall performance (with alpha corrected to .008 [.05/6] for six comparisons). Overall accuracy on the VRPC task did not correlate with overall accuracy on the GMFT, r = .087, p = .49, KFMT, r = -.051, p = .68, nor CFPT, r = -.164, p = .19. Similarly, accuracy on match trials of the VRPC did not correlate with the match condition of the GFMT, r = -.033, p = .79. A trend in the expected direction was observed with match trials on the VRPC and the KFMT, r = .265, p = .03, but this did not survive correction for multiple comparisons. However, mismatch performance on the GFMT and KFMT both correlated positively with the mismatch trials of the VRPC task, r = .430, p < .001 and r = .374, p = .002, respectively. These mismatch correlations are illustrated in Figure 3.2.

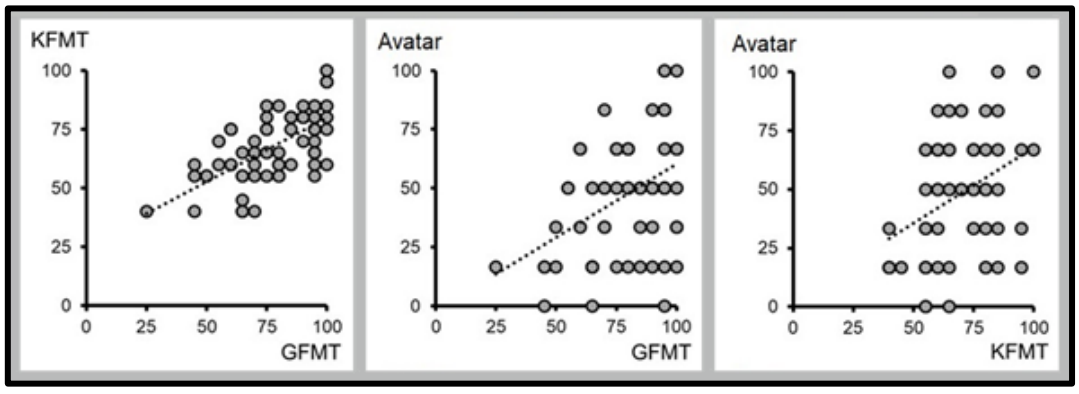

Figure 3.2. Correlations between GFMT, KFMT and VRPC accuracy on mismatch trials.

d-prime and criterion

The accuracy data for the GFMT, KFMT and VRPC tasks were also converted to *d'* and *criterion* to examine sensitivity and response bias. Correlational analyses between *d'* scores

(with alpha corrected to .017 [.05/3] for three comparisons) found GFMT (1.49) and KFMT (0.81) sensitivity correlated well, $r = .609$, $p < .001$, however VRPC sensitivity (1.17) did not correlate with either the GFMT, $r = .003$, $p = .98$, nor KFMT, $r = -.121$, $p = .33$. For *criterion* on the other hand, correlations (with alpha corrected to .017 [.05/3] for three comparisons) were found between the VRPC (-0.91) and both the GFMT (-0.02), $r = .385$, $p = .001$, and KFMT (-0.01), $r = .499$, $p < .001$, as well as between the two face-matching tests, $r = .633$, $p < .001$.

Regression for VRPC

A multi-linear regression was also conducted to assess whether individual performance across the self-report tests and the laboratory tests is predictive of person identification accuracy in complex environments. The PI20 score ($\beta = .21$, $p = .13$), overall accuracy on the GFMT ($\beta = .17$, $p = .30$) and KFMT ($\beta = -.18$, $p = .25$), and CFPT total error score on upright faces ($\beta = -.26$, $p = .08$) were not found to be significant predictors of overall accuracy on the VRPC task, with an overall model fit of $R^2 = 0.09$ ($F(4,61) = 1.47$, $p = .22$).

VRPC Re-Test

A follow-up test was conducted to establish the consistency of performance on the VRPC task over time. There was no correlation between the accuracy on match trials at Time 1 (M = 96.8%, SD = 6.71) compared to Time 2 (M = 92.1%, SD = 16.3), $r = .139$, $p = .55$, likely due to near-ceiling performance at both time intervals. Accuracy on mismatch trials, on the other hand, was more varied between participants (Time 1, M = 46.8%, SD = 25.1; Time 2, M = 43.7%, SD = 26.6) and correlated across time intervals, $r = .635$, $p = .002$. This was confirmed by a correlation of overall accuracy at Time 1 (M = 71.8%, SD = 11.9) and Time 2 (M = 67.9%, SD = 14.7), $r = .477$, $p = .03$.

Discussion

This experiment investigated whether self-reported face-recognition ability and laboratory tests of face discrimination and face matching can be used to determine an individual's suitability to perform face-matching tasks in passport control security settings, simulated here with a novel virtual reality paradigm (VRPC). Consistent with previous work (Shah, Sowden, et al., 2015), the self-report measure (PI20) exhibited correlations with overall performance on the GFMT, KFMT and CFPT, but not with person identification in the VRPC task. Similarly, accuracy also correlated across the laboratory tasks, replicating associations of the abilities to perform these discrimination and identification tests (e.g., Fysh & Bindemann, 2018). In contrast, match accuracy on the GFMT and KFMT, and overall accuracy on these tests and the CFPT, did not correlate with accuracy on the VRPC. The absence of such correlations were attributed to the near-ceiling performance for match trials on the VRPC, which will have constrained this analysis (see Table 3.1).

However, such associations were observed in the ability to detect identity mismatches, both for the GFMT and KFMT, with the VRPC task. This indicates that observers who were skilled at detecting identity mismatches on these established laboratory tests were also good at doing so in the more complex environment that the VRPC provides. These identity mismatch trials are likened to the documented security threat of impostors, who seek to avoid detection in airport security settings by utilising the valid identity documents of someone else that is of similar appearance (Bindemann, Fysh, Cross, & Watts, 2016; Meissner, Susa, & Ross, 2013; Susa, Michael, Dessenberger, & Meissner, 2019). The association of mismatch performance on laboratory tests of face matching with mismatch detection in the VRPC suggests that collectively these tasks could be used to select personnel for such roles to improve security.

Nevertheless, inspection of individual data also shows that this process is not precise (see Figure 3.2). The VRPC task is in the early stages of development and limited in complexity

and realism of the person stimuli. It is envisaged that VR will soon be capable of capturing the reality of passport control more effectively. Then, it could act as a means for personnel selection for such real-world settings in a way similar to how the face matching tests served as selection methods for the VRPC here. Just as the strongest correlations were observed between the most comparable tests here (i.e., GFMT and KFMT versus VRPC), stronger correlations between VRPC and real-life passport control might therefore emerge as this approach develops further. Considering that training methods have so far proven to be ineffective for improving face-matching accuracy (e.g., Towler et al., 2019), and the challenges which real-world settings may present are not incorporated into standardised laboratory tests (Ramon, Bobak, & White, 2019), it is also possible that a VR-based approach may provide a more viable alternative in future. Analogous simulations are already used routinely in other capacities for training and personnel selection in high stakes environments, such as flight simulators to train and test the abilities of pilots.

The findings from this chapter demonstrate the potential of this personnel selection process. This will continue to be refined with development of this VR application, opening up innovative new methods for recruitment and training for relevant security roles. The use of laboratory tasks as an assessment tool is currently limited by their correspondence to the real-world task (Ramon et al., 2019) and so improving on the realism of the VR task may help to bridge this gap. In addition, passport staff will encounter other factors which cannot be captured by photograph comparison paradigms but which can be simulated in VR, such as the social interaction with passengers. At passport control, the central challenge is to identify imposters seeking to evade detection, yet these individuals may betray their intentions through non-verbal cues. The final experimental chapter seeks to explore the influence of passenger body language on person identification, and whether this varies depending on observers' inherent face-matching ability.

Chapter 4:

Body Language Influences on Person Identification

Introduction

The previous chapter sought to apply the virtual reality (VR) paradigm to simulate the real-world task of passport control as a means to evaluate a person's suitability for operational deployment. In doing so, identity mismatches appeared infrequently during the task and so became challenging to detect. Importantly, for these trials performance correlated with established laboratory tests of face matching, demonstrating the potential for the VR simulation. However, in real-world tasks social interaction factors may also exert an influence on face matching, such as body language, which cannot be investigated with current laboratory paradigms. This will be explored using VR in this chapter.

International airports provide key entry points for people into other countries, with heightened security measures in recent years leading to greater interaction between passengers and security personnel (Trainer, 2017). Admission of entry relies critically on the routine identification of a large volume of passengers. This is typically achieved by identification from photographic documentation, by comparing the article image with its bearer. Extensive laboratory research has highlighted the difficulty of this task (for reviews, see Fysh & Bindemann, 2017a; Jenkins & Burton, 2008a, 2011; Robertson, Middleton, & Burton, 2015), even for trained and experienced security personnel (White, Dunn, Schmid, & Kemp, 2015; White, Kemp, Jenkins, Matheson, & Burton, 2014; White, Phillips, Hahn, Hill, & O'Toole, 2015; Wirth & Carbon, 2017). However, the presentation of such identity documents occurs in a context in which social interaction cues, such as body language, are also present.

The psychological literature demonstrates that body language can have substantial impact on interpersonal interaction and judgements (e.g., Burgoon, Guerrero, & Manusov, 2011; Knapp, Hall, & Horgan, 2013), but few studies have systematically examined the impact of such social interaction factors on the type of facial identification required at passport control. A heuristic technique employing factors such as body language may be pivotal in passport

control settings, for example, by seeking out those who appear to be behaving unusually. Consequently, substantial effort has been invested in real-world aviation settings in programmes that train staff to look for such nonverbal cues, for example the Screening of Passengers by Observation Techniques (SPOT) programme in the United States (see United States Government Accountability Office, 2010). The aim of the programme is to equip personnel for identifying persons seeking to evade detection or those who pose potential threats, but it is not clear whether this has enhanced security (United States Government Accountability Office, 2013). The current study therefore examined how monitoring of body language influences facial identification decisions in a security context, using a novel paradigm that simulates passport control with a virtual reality airport.

In psychology, person identification at passport control is widely studied through the task of unfamiliar face matching. In this task, a pair of facial images of unknown people are compared and classified either as an identity match (the same person) or an identity mismatch (two different people). One reason for studying classification of these identity mismatches is to simulate the real-world problem of impostors, who travel on legitimate identity documents of someone that is similar in facial appearance to avoid detection at passport control (Bindemann, Fysh, Cross, & Watts, 2016; Meissner, Susa, & Ross, 2013; Susa, Michael, Dessenberger, & Meissner, 2019). Typically, the faces for these tasks are displayed in isolation on plain backgrounds. This approach has been successful for advancing understanding of how a range of factors affect face matching, such as variation in a person's appearance (Bindemann & Sandford, 2011; Megreya, Sandford, & Burton, 2013; Ritchie & Burton, 2017), the addition of disguise (Henderson, Bruce, & Burton, 2001; Kramer & Ritchie, 2016; Wirth & Carbon, 2017), and individual differences in the ability of observers (e.g., Bindemann, Avetisyan, & Rakow, 2012; Bobak, Dowsett, & Bate, 2016; Bobak, Hancock, & Bate, 2016; Megreya &

Burton, 2006b). However, these simplistic approaches offer a limited proxy for understanding how additional factors, such as body language and motion, affect face matching.

It is already established that facial motion facilitates person identification, particularly when this is challenging (Butcher & Lander, 2017; Knight & Johnston, 1997; Lander, Bruce, & Hill, 2001; Lander, Christie, & Bruce, 1997; Lander & Chuang, 2005; O'Toole, Roark, & Abdi, 2002; Thornton & Kourtzi, 2002). However, although the face is the primary information source for person identification, body information also appears to have valuable input (Robbins & Coltheart, 2012), especially in the identity matching of unfamiliar people. When the face and body are presented in isolation, facial information is more diagnostic of identity; however, when both sources of information are available accuracy is enhanced (Rice, Phillips, & O'Toole, 2013). Furthermore, at increasing viewing distances this effect is amplified, shifting observers' reliance on identity information further towards the body (Hahn, O'Toole, & Phillips, 2016). This useful combination of both body and facial information has been demonstrated in identity sorting tasks, where intra-personal variability is easier to distinguish for whole persons than faces and bodies in isolation (Balas & Pearson, 2017). It would, however, seem that observers remain unaware of their reliance on body information when facial information in insufficient to provide a reliable identification, since self-reported feature usage is much lower for the body than for the face (Rice, Phillips, Natu, An, & O'Toole, 2013).

This integration of facial and body information has also been evident in research examining the identification of people in motion; when static stimuli are observed facial cues are prioritised over body information, whilst for dynamic stimuli both are examined more evenly and identification accuracy improves (O'Toole et al., 2010). This effect persists when moving footage from video clips is compared with multiple static images (Simhi & Yovel, 2016), thus providing converging evidence that it is motion itself that enables information from

multiple cues, such as the face and body, to be combined to enable accurate person identification (Yovel & O'Toole, 2016).

Given the role of the body in person identification, using static stimuli to investigate non-facial cues for person identification may be insufficient to progress research in this field. The impact of the body on person identification has successfully been investigated, however the specific influence of body language, which is not indicative of identity *per se* but may reflect a hidden motivation, on this task in security settings requires further study. Individuals using fraudulent passports at airports may, for example, betray their intent to avoid detection by displaying common non-verbal cues of anxiety, such as restless fidgeting (Ekman & Friesen, 1969). With regard to face matching, the impact of such factors is difficult to study. Real-world interactions involving face matching have only been examined by a few studies (e.g., Kemp, Towell, & Pike, 1997; White, Kemp, Jenkins, Matheson, & Burton, 2014), yet such experiments face logistical challenges and the systematic control of variables such as non-verbal behaviour is difficult to maintain. As a result, additional measures, such as double-blind procedures, are taken to *prevent* intrusion of such variables. Equally, owing to the security-sensitive nature of this task in occupational field settings, such as at passport control, these factors cannot easily be manipulated.

In this study, a new methodology is applied in an attempt to overcome these limitations, by measuring the impact of body language on person identification at a VR airport. VR enables the simulation of complex and detailed environments but that can be strictly controlled for the purpose of experiments. This novel approach therefore allows for the study of factors that may impact real-world person identification, but that conventional laboratory experiments cannot easily address. This approach has been developed and validated through a stringent series of experiments in Chapter 2. These demonstrate that VR avatars can preserve identity information

from real faces, and that matching of pairs of such avatar faces also reflects similar cognitive processes to the matching of photographs of real faces.

Here, this approach was employed to investigate whether body language influences decision-making in a face-matching task. For this purpose, participants were immersed in a VR airport environment as passport control officers, who were required to make identification decisions for a queue of passengers in an arrivals hall. These passengers were equipped with an idle mode that creates small shifts in body posture when a person is stationary, to increase observers' sense of realism in VR. To manipulate body language, the majority of passengers were programmed to idle in the same manner. In a proportion of these passengers, however, the idle level was raised to simulate more restless body language. The question of main interest here was whether this alternate display of body language would be perceived as unusual in this context and would therefore affect face-matching decisions. Specifically, it was reasoned that the detection of a person with unusual body language might increase attention to such passengers, leading to enhanced scrutiny of their facial identity. Thus, the aim was to examine whether identity mismatches (i.e., the critical impostors) would be more likely to be detected when these were exhibiting unusual body language and, in turn, whether they would be more frequently missed when not.

Experiment 9

The aim of this experiment was to investigate whether body language influences person identification from the face in a matching task. At real-life passport control, officers would be positioned in front of a queue of passengers for which they are required to compare their faces to their passport photographs. The virtual reality airport of this study was designed to replicate this setup, with participants standing within a booth looking towards a queue of person avatars. Participants compared each of these three-dimensional (3D) avatars to a respective two-

dimensional (2D) face portrait, which was displayed on a passport-style ID card, to determine whether this presented an identity match or mismatch for its bearer.

The 3D avatars were equipped with body language that someone might exhibit naturally while waiting to be processed at passport control. Thus, they were programmed to look around and shift in their stance occasionally. For most avatars this animation was performed at 'idle' speed, which represented a normal level of animation. In a subset of avatars, however, these activity levels were increased to represent 'restless' and 'lively' waiting behaviours. It was then aimed to determine how these increases in body language affected classification of identity matches and mismatches. The detection of identity mismatches is a primary concern for person identification at passport control, but these cases also occur with less frequency than identity matches (Bindemann, Avetisyan, & Blackwell, 2010; Fysh & Bindemann, 2017b, 2018; Papesh & Goldinger, 2014; Susa et al., 2019). In these cases, unusual behaviour, such as body language that differs from the majority of passengers, might serve as a behavioural indicator of deceptive behaviour. Consequently, if observers are sensitive to unusual body language, then this may lead to an enhanced detection of identity mismatches by drawing attention to these specific cases. In turn, mismatches that do not exhibit unusual behaviour might be more frequently missed, and unusually-behaving matches might be more likely to be identified as mismatches instead.

Method

Participants

The participants consisted of 30 Caucasian students from the University of Kent (5 male, 25 female), with a mean age of 20.5 years (*SD* = 5.0 years). This sample size is comparable to studies using a range of face-matching paradigms (e.g., Bindemann, Attard, Leach, & Johnston 2013; Megreya & Burton, 2007; White, Rivolta, Burton, Al-Janabi, & Palermo, 2017). All

participants reported normal or corrected-to-normal vision and completed the experiment in exchange for course credit. As with all experiments in this study, owing to the use of virtual reality equipment, no persons with epilepsy or who were liable to motion sickness were recruited. Before immersion in the VR, participants were briefed about potential side effects of using VR, such as discomfort from wearing the headset and symptoms of motion sickness, and health and safety procedures.

Stimuli

During the experiment, participants were immersed in a VR passport control environment with an HTC Vive headset with a resolution of 1080 x 1200 pixels per eye. Two handheld controllers enabled participants to interact with the environment and respond to the stimuli. The passport control environment was constructed by positioning 3D objects within a pre-built 3D airport hall model (https://www.turbosquid.com/3d-models/airport-departures-lounge-3d-model/626226). This model was built in 3DS Max and used VRay for rendering. The completed passport control environment consisted of a booth area in which the participants were standing, equipped with a desk, chair and computer. These objects were added to improve the realism of the booth and so response button instructions could be overlaid on a virtual computer screen inside the passport control booth. This booth was situated inside the airport hall with other visual cues, such as departure boards and a waiting aeroplane, which were clearly visible to participants. The environment is illustrated in Figure 4.1.

The person stimuli consisted of 100 animated 3D avatars, each paired with a 2D face portrait of a second avatar, which was embedded on a passport-style card. The 3D avatars were created by combining 2D photographs of real faces with an avatar from an existing database (see www.kent.ac.uk/psychology/downloads/avatars.pdf). Using graphics software (Artweaver 5), the internal features of a face photograph were mapped onto the features of the

avatar's face area, with the edges smoothed and skin colour adjusted to blend the graphics. This process was repeated for each identity to produce a match pair, with a 2D face portrait captured from one avatar to create a passport image, which was sized to 438 x 563 pixels at a resolution of 150 ppi.

For identity mismatch trials, an avatar was paired with a 2D face portrait of a similar-looking identity, matched for gender and approximate age. To provide a closer proximate to real-world conditions, mismatches occurred with much lower frequency than matches (see, e.g., Bindemann et al., 2010; Fysh & Bindemann, 2017b, 2018; Papesh & Goldinger, 2014; Susa et al., 2019). Therefore, of the 100 stimulus pairings, 94 trials consisted of the same person (identity matches) while six trials were of two different people (identity mismatches).

Figure 4.1. The virtual reality airport environment. The inset demonstrates the forward-facing view from inside the passport control booth.

Procedure

The experiment was controlled using Vizard 5 software. In the VR airport environment, the 3D avatars approached from the back of the airport hall and proceeded to queue around rope barriers. One at a time, they walked towards the passport control booth, where the participants were positioned inside the VR environment, and waited to be processed. The corresponding 2D face portrait for each avatar passenger appeared on a passport-style photo card, which could be picked up and moved with the VR controllers. This enabled participants to hold the passport in any position necessary to facilitate an identity comparison, for example, close to the face of the animated avatar (see inset of Figure 4.1). Participants pressed the thumb-pad of the right controller to report an identity match or the thumb-pad of the left controller to report an identity mismatch. Participants were instructed to match for identity rather than image, in this instance asked to imagine how they would vary to their own passport as an example. They were also informed how at passport control the majority of passengers would be a match to their passport and so their task was to detect the small number of mismatches if there were any to be found. This information was provided to ensure participants did not have an expectation of a similar frequency of match and mismatch trials and therefore falsely report mismatches to even out their responses when uncertain. Once a response was given the avatar walked away and the photo on the card changed to the one corresponding for the next avatar in line as it approached the desk. Participants continued making these match or mismatch decisions until the whole queue had been processed.

Whilst queuing and standing at the desk the avatars shifted in their stance through the avatars' built-in animation "idle1" in Vizard, initiated at a random starting point in the cycle to prevent synchronised motion. Once at the desk area, the scale factor of this animation was adjusted so that selected avatars were moving with different levels of activity (i.e. completing the animation cycle in differing durations). Three activity levels were used. These

corresponded to an animation scale factor of 1 for the ‘idle’ condition, in which a cycle lasted 13.3 seconds before being repeated, and of 2 (6.7 seconds per cycle) and 3 (4.4 seconds per cycle) for the ‘restless’ and ‘lively’ conditions, respectively.

In the experiment, each participant completed 100 trials, comprising of 94 match trials and six mismatch trials. The 94 match trials were broken down further in 88 non-critical trials, all of which displayed ‘idle’ body language and were used to provide a task context, and six critical match trials, which were used as a direct comparison for the mismatches. These critical match trials were selected to match the mismatch trials for accuracy, based on data from Chapter 2 (mean accuracy matches = 68.1%, *SD* = 21.7; mean accuracy mismatches = 65.7%, *SD* = 22.0; $t(119) = 0.77$, $p = .44$, $d = 0.11$). Two of the critical matches and mismatches displayed ‘idle’, ‘restless’ or ‘lively’ body language. When the experimental program launched, the activity manipulations were randomly assigned to the critical match and mismatch trials and these trials were randomly distributed throughout the last 90 places in queue. The first 10 trials of the queue always consisted of 10 non-critical match trials to accustom participants to the idle activity level. Participants were informed at the beginning of the experiment that mismatch frequency would be low, but were not made aware of variation in body language. They were not given any time restrictions in which to complete the task to encourage accurate performance.

Following the VR task, participants completed a questionnaire to report any differences in animation that they might have noticed. The purpose of the questionnaire was to ascertain whether participants were sensitive to the activity-level manipulation (i.e., correctly perceived three levels of activity). Participants were first asked if they noticed anything unusual during the experiment, providing opportunity for them to report freely without being led to suspect differences in body language. Secondly, they were informed that some avatars may have been moving at different speeds (activity levels) and were asked to report how many they had

perceived throughout the experiment. Finally, they reported the relative speeds of the perceived number of activity levels using sliding scales to verify the response given in the previous question.

Results

Percentage accuracy

Overall accuracy for the 88 non-critical match trials was 92.3% (*SD* = 5.2). Inferential analysis was only applied to the data for the six critical match and six mismatch trials, which are displayed in Figure 4.2. In a first step of this analysis, participants were given a sensitivity score of how far their reported number of activity levels deviated from the actual number. For example, those who correctly reported three levels scored 0, whilst those who reported only one level scored -2. Twelve of the 30 participants correctly reported that there were three activity levels. This body language sensitivity score was used as a covariate in the inferential analysis. The critical data were then analysed using a 2 (trial type: match vs. mismatch) x 3 (activity level: idle vs. restless vs. lively) within-subjects ANCOVA[1]. This showed a main effect of trial type, $F(1,28) = 43.21$, $p < .001$, $\eta_p^2 = .61$, due to higher match than mismatch accuracy, but no main effect of activity level, $F(2,56) = 0.34$, $p = .72$, $\eta_p^2 = .01$, nor sensitivity, $F(1,28) = 1.45$, $p = .24$, $\eta_p^2 = .04$. Two-way interactions between any of the factors, all $Fs \leq 1.01$, all $ps \geq .37$, all $\eta_p^2 \leq .04$, and a three-way interaction were not found, $F(2,56) = 0.97$, $p = .39$, $\eta_p^2 = .03$.

[1] Owing to the small number of critical trials, and to ensure a normal distribution of these percentage accuracy data, an arcsine-square root transformation was also applied and the analyses repeated. In this and all subsequent experiments the same result was obtained, as shown in the Appendix.

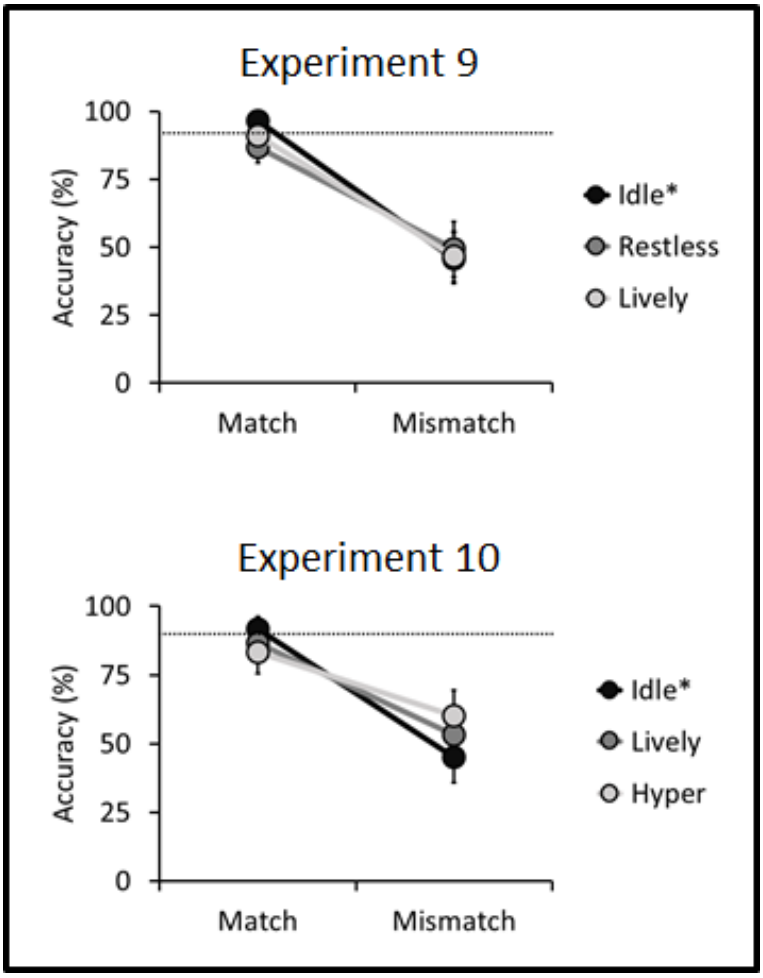

Figure 4.2. Mean accuracy scores of the critical match and mismatch trials for each activity level and trial type for Experiments 9 and 10. The error bars represent the standard errors of the means. The dashed line represents the mean accuracy for the non-critical match trials, the activity level of which is denoted by the asterisk (i.e., the majority activity level).

d-prime and criterion

The accuracy data were also converted to *d'* and *criterion* to examine sensitivity and response bias. One-factor ANOVAs did not show an effect of activity level for *d'* (idle = 0.40, restless = 0.37, lively = 0.35), $F(2,58) = 0.12$, $p = .88$, $\eta_p^2 < .01$, or *criterion* (idle = -0.35, restless = -0.21, lively = -0.29), $F(2,58) = 1.84$, $p = .17$, $\eta_p^2 = .06$. These results therefore converge with the analysis of the percentage accuracy data to show that body language did not affect face matching.

Discussion

This experiment manipulated the activity levels of avatars within a virtual passport control environment to examine whether body language influences face-matching decisions. Identity matches with idle body language presented the majority of trials and their classification was near ceiling. By comparison, the occurrence of identity mismatches was low, in 6% of trials, and these face pairs were only classified correctly in half of these trials. Importantly to the question of main interest, however, classification of matches and mismatches was not affected by variation in body activity levels.

This pattern of results was observed irrespective of whether participants reported awareness of the differences in body language. However, participants were not made aware of this manipulation prior to the experiment, and only 12 of the 30 participants were accurate in reporting that three different activity levels of body language were used. Thus, it is possible that the body language manipulation that was trialled here was too weak to be detected by observers and, therefore, to influence person identification.

Experiment 10

In the previous experiment, the activity level of body language did not influence person identification. However, most participants showed limited awareness of the body language manipulation, by failing to notice any differences in activity level. It is therefore possible that differences in body language were too subtle to elicit an effect. To investigate this possibility, the different body language activity levels were increased in Experiment 10, to exaggerate the perceptual differences between conditions. For this purpose, the ‘idle’ and ‘lively’ activity levels of Experiment 9 were retained but a new ‘hyper’ condition, in which the lively activity level was doubled in magnitude, was added.

Method

Participants

Thirty Caucasian students from the University of Kent (7 male, 23 female), with a mean age of 21.0 years (*SD* = 6.7 years), participated in exchange for course credit. All participants reported normal or corrected-to-normal vision, and none had participated in Experiment 9.

Stimuli and Procedure

The stimuli, procedure and task instructions were identical to the preceding experiment except for the scale factors of the animation. The variations in animation cycle duration were increased to enable the higher activity trials to appear more perceptually different to the idle activity trials (scale factor 1, animation cycle of 13.3 seconds). The lively activity trials were maintained (scale factor 3, 4.4 second cycle), but 'hyper' activity trials (scale factor 6, 2.2 second cycle) replaced the restless trials (scale factor 2, 6.7 second cycle).

Results

Percentage accuracy

The data for this experiment were analysed using the same method as Experiment 9. Overall accuracy for the 88 non-critical match trials was 90.4% (*SD* = 8.2). As with Experiment 9, inferential analysis was only applied to the data for the six critical match and six mismatch trials, the data for which can be seen in Figure 4.2. A sensitivity score was again calculated based on participants' questionnaire responses. Fifteen of the 30 participants correctly reported that there were three activity levels. A 2 (trial type: match vs. mismatch) x 3 (activity level: idle vs. lively vs. hyper) within-subjects ANCOVA with the covariate sensitivity showed a main effect of trial type, $F(1,28) = 26.53$, $p < .001$, $\eta_p^2 = .49$, due to match trials being classified more accurately than mismatch trials. However, there was no main effect of activity level,

$F(2,56) = 0.02$, $p = .98$, $\eta_p^2 < .01$, sensitivity, $F(1,28) = 1.02$, $p = .32$, $\eta_p^2 = .04$, no two-way interactions between any of the factors, all $Fs \leq 1.87$, all $ps \geq .16$, all $\eta_p^2 \leq .06$, and no three-way interaction, $F(2,56) = 0.36$, $p = .70$, $\eta_p^2 = .01$.

d-prime and criterion

The accuracy data were also converted to *d'* and *criterion* to examine sensitivity and response bias. As in Experiment 9, one-factor ANOVAs did not show an effect of activity level for *d'* (idle = 0.35, lively = 0.38, hyper = 0.41), $F(2,58) = 0.21$, $p = .82$, $\eta_p^2 = .01$, or *criterion* (idle = -0.31, lively = -0.22, hyper = -0.16), $F(2,58) = 2.75$, $p = .07$, $\eta_p^2 = .09$.

Discussion

This experiment replicates the results of Experiment 9 closely. Match accuracy was higher than mismatch accuracy and around half of the participants reported sensitivity to the three body activity levels. However, despite the increase in body language activity, this did not influence face-matching accuracy. This finding appears at odds with previous literature suggesting that body cues are processed unconsciously and affect facial identification (Rice, Phillips, Natu, et al., 2013), but this research relied on identity information from the body rather than body language *per se*. This contrast suggests that, if body *language* affects person identification at all, then this may require conscious monitoring of such behavioural cues in order to have an impact on face-matching decisions.

Experiment 11

In Experiment 9 and 10, matching avatar faces to their passport image was the primary task. Participants were not required to monitor body language closely, which may explain why this did not influence identification decisions. In Experiment 11, participants were therefore instructed directly to monitor variation in body language, with a view to aiding the detection of identity mismatches, to determine whether such explicit instruction is required to influence matching decisions.

Method

Participants

A further 30 Caucasian participants (8 male, 22 female), with a mean age of 19.3 years (*SD* = 1.3 years), were recruited from the University of Kent for course credit. All participants reported normal or corrected-to-normal vision, and none had participated in the previous experiments.

Stimuli and Procedure

For this experiment, participants' attention was directed towards the animation of the avatars. It was explained prior to the task that the avatars would be shifting in their stance whilst waiting, that this level of activity could vary, and that such differences in body language might be useful for detecting identity mismatches. Owing to the inclusion of this additional instruction, participants did not complete the questionnaire for reporting avatar animation from Experiment 9 and 10. All other aspects of the stimuli, procedure and task instructions remained identical. Thus, most avatars comprised of identity matches displaying idle behaviour, with a subset of six critical matches and six mismatches displaying idle, lively and hyper body language.

Results

Percentage accuracy

Overall accuracy for the 88 non-critical match trials was 92.8% (*SD* = 5.5). Cross-subject mean accuracy scores were calculated for the six critical match and six mismatch trials for each level of activity and are displayed in Figure 4.3. A 2 (trial type: match vs. mismatch) x 3 (activity level: idle vs. lively vs. hyper) within-subjects ANOVA of this data did not show a main effect of trial type, $F(1,29) = 0.66$, $p = .42$, $\eta_p^2 = .02$, or activity level, $F(2,58) = 0.08$, $p = .92$, $\eta_p^2 < .01$, but an interaction of these factors, $F(2,58) = 32.83$, $p < .001$, $\eta_p^2 = .53$.

Analysis of simple main effects revealed an effect of activity level for match trials, $F(2,28) = 21.40$, $p < .001$, $\eta_p^2 = .60$, with paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) showing that idle match trials were identified more accurately than both lively and hyper match trials, $t(29) = 5.81$, $p < .001$, $d = 1.34$ and $t(29) = 5.22$, $p < .001$, $d = 1.45$, respectively. Accuracy for lively and hyper match trials did not differ, $t(29) = 1.14$, $p = .26$, $d = 0.26$.

A simple main effect of activity level for mismatch trials was also found, $F(2,28) = 19.22$, $p < .001$, $\eta_p^2 = .58$. Paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) showed that both lively and hyper mismatch trials were identified more accurately than idle mismatch trials, $t(29) = 4.94$, $p < .001$, $d = 1.23$ and $t(29) = 6.18$, $p < .001$, $d = 1.41$, respectively. As with match trials, accuracy was comparable for lively and hyper mismatch trials, $t(29) = 1.00$, $p = .33$, $d = 0.26$.

Finally, a simple main effect of trial type was found within the idle activity level, $F(1,29) = 49.83$, $p < .001$, $\eta_p^2 = .63$, with match trials being performed more accurately than mismatch trials. By contrast, mismatch accuracy was higher than match accuracy within the lively, $F(1,29) = 8.83$, $p = .006$, $\eta_p^2 = .23$, and hyper activity level, $F(1,29) = 19.33$, $p < .001$, $\eta_p^2 = .40$.

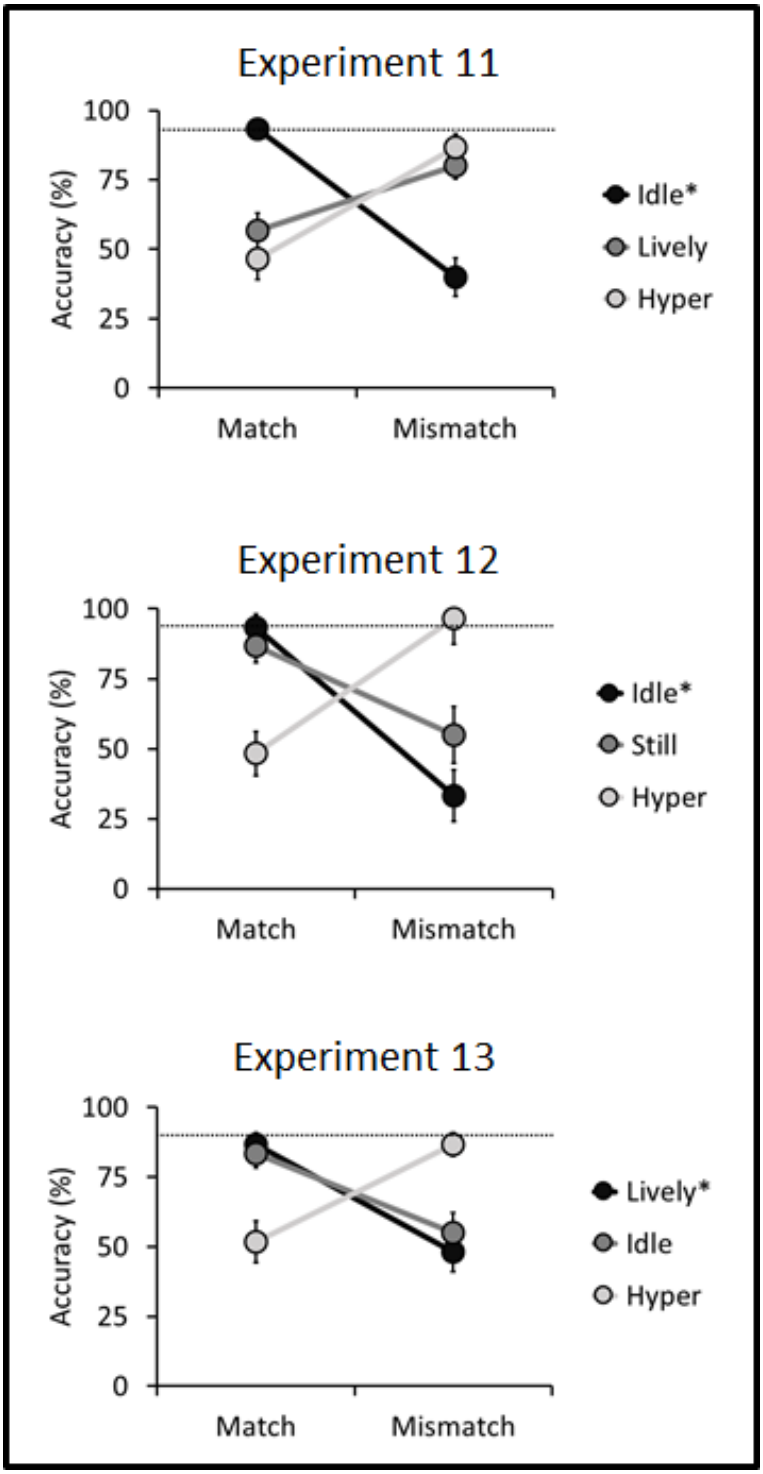

Figure 4.3. Mean accuracy scores of the critical match and mismatch trials for each activity level and trial type for Experiments 11, 12 and 13. The error bars represent the standard errors of the means. The dashed line represents the mean accuracy for the non-critical match trials, the activity level of which is denoted by the asterisk (i.e., the majority activity level).

d-prime and criterion

One-factor ANOVAs did not show an effect of activity level for *d'* (idle = 0.32, lively = 0.35, hyper = 0.30), $F(2,58) = 0.14$, $p = .87$, $\eta_p^2 = .01$, but for *criterion* (idle = -0.36, lively = 0.16, hyper = 0.28), $F(2,58) = 32.51$, $p < .001$, $\eta_p^2 = .53$. Paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) revealed a mismatch bias on hyper and lively trials compared to idle trials, $t(29) = 6.76$, $p < .001$, $d = 2.03$ and $t(29) = 6.86$, $p < .001$, $d = 1.80$, respectively. *Criterion* for hyper and lively trials did not differ, $t(29) = 1.52$, $p = .14$, $d = 0.37$.

Discussion

In this experiment, participants were informed that body language may assist detection of mismatches, to investigate whether this influences person identification when observers are explicitly instructed to monitor for such cues. In contrast to Experiments 9 and 10, body language exerted a clear effect on the classification of identity matches and mismatches. In close convergence with the preceding experiments, matches exhibiting idle body language were detected with near-perfect accuracy, whereas more than half of all mismatches with this body language were classified incorrectly. By contrast, however, identification of lively and hyper matches was greatly reduced, whereas mismatch classification was near ceiling in these body language conditions.

As most avatars exhibited idle body language, lively and hyper matches and mismatches occurred infrequently in this experiment. Considering the tendency to classify these lively and hyper avatars as mismatches, the findings of Experiment 11 therefore indicate that observers employed unusual body language as a heuristic to support mismatch decisions. This indicates that this non-facial information influences face identity matching when observers are explicitly monitoring for such cues. This converges with recent studies demonstrating that body cues can

assist in affirming or negating facial identification decisions (Balas & Pearson, 2017; Hahn, O'Toole, & Phillips, 2016; Rice, Phillips, & O'Toole, 2013), particularly when people are presented in motion (O'Toole et al., 2010; Simhi & Yovel, 2016). Experiment 11 extends these findings to face matching in a virtual passport control environment.

Experiment 12

The results of Experiment 11 indicate that avatars exhibiting unusual body language increased mismatch classifications. However, unusual body language was always characterised by raising activity levels from idle to lively and hyper. Therefore, the question arises of whether this effect is driven by an *increase* in normal body language, or reflects that lively and hyper avatars are behaving *differently* to the majority of idle avatars in the experiment. To investigate this issue, the majority of avatars again displayed idle body language in Experiment 12 and a subset exhibited hyper activity levels. However, the lively condition was replaced with 'still' avatars, which displayed no movement during identification. If the increase in mismatch decisions that was observed in Experiment 11 is driven by behaviour that is unusual from the norm, then this effect should be observed with both still and hyper avatars in Experiment 12. If, on the other hand, this effect relies on increased body language, then it should be observed only with hyper avatars.

Method

Participants

Thirty Caucasian participants (9 male, 21 female), with a mean age of 19.9 years (*SD* = 2.7 years), who had not participated in the previous experiments were recruited from the University of Kent. All reported normal or corrected-to-normal vision and were granted course credit or a small fee for their participation.

Stimuli and Procedure

The stimuli, procedure and task instructions were identical to the previous experiment except for changes to the animation cycles. Idle and hyper activity levels maintained scale factors of 1 and 6, but lively trials were replaced with a 'still' condition with a scale factor of zero. Thus, avatars in the still condition approached the passport control desk and then did not move at all during identification.

Results

Percentage accuracy

Overall accuracy for the 88 non-critical match trials was 94.0% (*SD* = 4.3). Cross-subject mean accuracy scores were calculated for the six critical match and six mismatch trials for each level of activity and are displayed in Figure 4.3. A 2 (trial type: match vs. mismatch) x 3 (activity level: still vs. idle vs. hyper) within-subjects ANOVA of this data did not show a main effect of activity level, $F(2,58) = 2.19$, $p = .12$, $\eta_p^2 = .07$, but a main effect of trial type, $F(1,29) = 8.44$, $p = .007$, $\eta_p^2 = .23$, and an interaction between factors, $F(2,58) = 48.92$, $p < .001$, $\eta_p^2 = .63$.

Analysis of simple main effects revealed an effect of activity level for match trials, $F(2,28) = 17.81$, $p < .001$, $\eta_p^2 = .56$, with paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) showing that idle and still match trials were performed more accurately than hyper match trials, $t(29) = 5.84$, $p < .001$, $d = 1.50$ and $t(29) = 5.14$, $p < .001$, $d = 1.16$, respectively. Accuracy for idle and still match trials did not differ, $t(29) = 1.07$, $p = .29$, $d = 0.30$.

A simple main effect of activity level within mismatch trials was also found, $F(2,28) = 43.77$, $p < .001$, $\eta_p^2 = .76$, with paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) showing higher accuracy for hyper mismatches than idle, $t(29) = 8.38$, $p <$

.001, d = 2.21, and still mismatches, $t(29) = 6.53$, $p < .001$, d = 1.54. In addition, accuracy on still mismatch trials was also higher than on idle mismatch trials, $t(29) = 2.54$, $p < .017$, d = 0.58.

Finally, simple main effects of trial type were found within the idle, $F(1,29) = 60.23$, $p < .001$, $\eta_p^2 = .68$, and still activity levels, $F(1,29) = 12.94$, $p = .001$, $\eta_p^2 = .31$, due to higher accuracy for match than mismatch trials. Within the hyper activity level, the reverse pattern was observed, with higher accuracy on mismatch than match trials, $F(1,29) = 35.40$, $p < .001$, $\eta_p^2 = .55$.

d-prime and criterion

One-factor ANOVAs did not show an effect of activity level for *d'* (still = 0.40, idle = 0.25, hyper = 0.43), $F(2,58) = 2.19$, $p = .12$, $\eta_p^2 = .07$, but for *criterion* (still = -0.21, idle = -0.40, hyper = 0.33), $F(2,58) = 48.92$, $p < .001$, $\eta_p^2 = .63$. Paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) revealed a greater mismatch bias on hyper trials compared to idle, $t(29) = 8.85$, $p < .001$, d = 2.44, and still trials, $t(29) = 7.74$, $p < .001$, d = 1.69. There was no difference in *criterion* between idle and still trials after correcting for multiple comparisons, $t(29) = 2.48$, $p = .019$, d = 0.60.

Discussion

This experiment replicates the key aspects of Experiment 11, by showing that match accuracy declines and mismatch accuracy is enhanced when avatars display unusually hyper body language. The current experiment examined in addition whether a similar effect is found when avatars do not display any body movement during identification. On match trials, accuracy for idle and still avatars converged. Thus, in a context in which the majority of avatars display idle body language, the absence of body language did not influence classification of

the identity matches. On mismatch trials, on the other hand, still avatars were classified correctly more often than idle avatars, but this effect was much smaller than for hyper avatars. Overall, these findings suggest that it is predominantly an *increase* in body language, rather than unusual body language, which affects face identification in the paradigm here. However, an alternative explanation is also possible, as the scale factors for still and idle avatars were more closely matched (at 0 and 1, respectively) than for hyper avatars (6). This opens the possibility that performance for still and idle avatars was more comparable due to the greater perceptual similarity of these activity levels, in comparison with hyper trials.

Experiment 13

In the previous experiment, still avatars exerted a much more limited influence on facial identification than hyper avatars, which suggests that it is an increase in body language, rather than unusual body language, which determines these effects. However, the activity levels of still and idle avatars were also more closely matched relative to the hyper condition. To investigate whether this can account for the results of Experiment 12, an experiment was conducted in which the non-critical avatars now exhibited lively instead of idle body language, with critical matches and mismatches exhibiting idle, lively and hyper activity levels. In contrast to the preceding experiments, the majority of avatars therefore exhibited lively body language, and hyper *as well as* idle body language represented the unusual body language conditions. In this design, idle and hyper body language are equidistant from the normative body language behaviour in terms of the scale-factor ratio. Thus, this experiment provides a better test for whether the effect of body language in the preceding experiments is due to some avatars exhibiting *increased* or *unusual* body language (i.e., increased or decreased activity levels from the norm).

Method

Participants

Thirty Caucasian participants (4 male, 26 female), with a mean age of 22.2 years (*SD* = 5.4 years), were recruited from the University of Kent and granted course credit or a small fee for their participation. None of these individuals had participated in the previous experiments. All reported normal or corrected-to-normal vision.

Stimuli and Procedure

The method, procedure and task instructions were identical to Experiment 11 with the exception that the non-critical match trials now displayed the lively activity level. As a result, most avatars displayed lively behaviour (at scale factor 3), with a small subset of avatars displaying idle (low activity; scale factor 1) or hyper (high activity; scale factor 6) behaviour.

Results

Percentage accuracy

Overall accuracy for the 88 non-critical match trials was 89.6% (*SD* = 10.1). The cross-subject mean accuracy scores were calculated for the six critical match and six mismatch trials for each level of activity and are displayed in Figure 4.3. A 2 (trial type: match vs. mismatch) x 3 (activity level: idle vs. lively vs. hyper) within-subjects ANOVA of this data did not show main effects of trial type, $F(1,29) = 2.89$, $p = .10$, $\eta_p^2 = .09$, or activity level, $F(2,58) = 0.07$, $p = .94$, $\eta_p^2 < .01$, but an interaction between these factors, $F(2,58) = 20.12$, $p < .001$, $\eta_p^2 = .41$.

For match trials, a simple main effect of activity level was found, $F(2,28) = 7.52$, $p = 002$, $\eta_p^2 = .35$, with paired-samples *t*-tests (with alpha corrected to .017 [.05/3] for three comparisons) showing that lively and idle matches were classified more accurately than hyper matches, $t(29) = 3.88$, $p < .001$, $d = 1.06$ and $t(29) = 3.60$, $p = .001$, $d = 0.91$, respectively. In

contrast, accuracy was comparable for lively and idle match trials, $t(29) = 0.63$, $p = .57$, $d = 0.13$.

A corresponding simple main effect of activity level was found for mismatch trials, $F(2,28) = 16.27$, $p < .001$, $\eta_p^2 = .54$, due to the more accurate classification of hyper than lively, $t(29) = 4.89$, $p < .001$, $d = 1.16$, and idle mismatches, $t(29) = 4.08$, $p < .001$, $d = 0.96$. Also as for identity matches, accuracy for lively and idle mismatches did not differ, $t(29) = 0.68$, $p = .50$, $d = 0.16$.

Finally, simple main effects of trial type within the lively, $F(1,29) = 17.41$, $p < .001$, $\eta_p^2 = .38$, and idle conditions were found, $F(1,29) = 7.90$, $p = .009$, $\eta_p^2 = .21$, due to higher accuracy for match than mismatch trials. For the hyper condition, on the other hand, the reverse pattern of superior mismatch accuracy was shown, $F(1,29) = 14.07$, $p < .001$, $\eta_p^2 = .33$.

d-prime and criterion

One-factor ANOVAs did not show an effect of activity level for *d'* (idle = 0.37, lively = 0.33, hyper = 0.37), $F(2,58) = 0.07$, $p = .94$, $\eta_p^2 < .01$, but for *criterion* (idle = -0.19, lively = -0.26, hyper = 0.24), $F(2,58) = 20.12$, $p < .001$, $\eta_p^2 = .41$, due to a greater mismatch bias on hyper compared to lively and idle trials, $t(29) = 6.03$, $p < .001$, $d = 1.45$ and $t(29) = 5.09$, $p < .001$, $d = 1.19$, respectively. *Criterion* for lively and idle trials did not differ, $t(29) = 0.77$, $p = .45$, $d = 0.19$.

Discussion

As in the two preceding experiments, accuracy for trials with the most common activity level, which was lively body language in this case, was high when these were identity matches and low for mismatches. This demonstrates, once again, that mismatches are frequently missed in this paradigm when additional cues from unusual body language are not available. As in the

preceding experiments also, this pattern was reversed dramatically when such unusual body language was present in the hyper condition. The primary aim of this experiment was to confirm whether these effects are only present when unusual body language is characterised by an increase in activity compared to the norm, or also when activity levels are attenuated on idle trials. Classification of idle matches and mismatches aligned with the most common lively condition. This provides converging evidence with Experiment 12 to indicate that the current effects are driven by increased expressive body language rather than body language that differs to the norm *per se*.

Experiment 14

The experiments reported so far demonstrate that body language strongly biases face-matching decisions. However, since people differ greatly in their ability to match the identities of faces (e.g., Bindemann et al., 2012; Burton, White, & McNeill, 2010; Fysh & Bindemann, 2018; White, Burton, Kemp & Jenkins, 2013; White, Kemp, Jenkins, Matheson, & Burton, 2014), it is possible that the body language effect is influenced by these individual differences. It is conceivable, for example, that observers with high face-matching ability rely on facial information more strongly and, in turn, exhibit an attenuated bias to the presence of body language cues. This final experiment investigates this possibility by comparing performance on the VR passport control task (VRPC) with two laboratory tests of face-matching ability, the Glasgow Face Matching Test (GFMT; Burton et al., 2010) and the Kent Face Matching Test (KFMT; Fysh & Bindemann, 2018). Both tests require identity comparisons of pairs of face photographs and reveal broad individual differences in matching ability. These tests also correlate with a range of facial discrimination and identification tasks (see, e.g., Fysh, 2018; Fysh & Bindemann, 2018; McCaffery, Robertson, Young, & Burton, 2018), and with critical

mismatch trials in Chapter 3, indicating that these are stable measures against which person identification in the VRPC task can be compared.

Method

Participants

One-hundred Caucasian students from the University of Kent (18 male, 82 female), with a mean age of 19.3 years (*SD* = 1.8 years), participated in this experiment in exchange for course credit. All participants reported normal or corrected-to-normal vision.

Stimuli and Procedure

For this experiment, the VRPC task was modified as following. The same proportion of 88 non-critical match trials, six critical match trials and six mismatch trials was used. However, only two activity level conditions were employed, comprising idle and lively body language, resulting in the presentation of three match and mismatch trials of each. In addition, the order of the 100 trials was fixed, with critical matches displayed on trial 18, 35, 61, 66, 87, and 92, and mismatches on trial 24, 28, 48, 71, 83, and 97. These 12 trials alternated in activity level condition, which was counterbalanced across participants. These changes were implemented to ensure that performance was more directly comparable across observers for analysis of individual differences. All remaining aspects of this task, including instructions given to participants, remained the same as in Experiments 11 to 13, with participants' attention directed towards the animation of the avatars.

In addition to the VRPC task, the GFMT (Burton et al., 2010) and KFMT (Fysh & Bindemann, 2018) were included as additional tasks in this experiment. The GFMT face pairs consist of images of faces taken from a frontal view displaying a neutral expression. Both images in a face pair are taken with different cameras and, in the case of identity matches,

approximately 15 minutes apart. Each face image is cropped to show the head only and converted to greyscale with a resolution of 72 ppi. The dimensions of the faces range in width from 70 mm to 90 mm and in height from 85 mm to 125 mm, and are spaced between 40 mm and 55 mm apart on screen. This experiment employed 20 identity match and 20 mismatch trials from the GFMT (for more information, see Burton et al., 2010) and participants were not informed as to the ratio of match-to-mismatch trials.

The KFMT face pairs consist of an image from a student ID card, presented at a maximal size of 35 mm (w) x 47 mm (h), and a portrait photo, sized at 70 mm (w) x 82 mm (h) at a resolution of 72 ppi, spaced 75 mm apart. The student ID photos were taken at least three months prior to the face portraits and were not constrained by pose, facial expression, or image-capture device. The portrait photos depict the target's head and shoulders from a frontal view whilst bearing a neutral facial expression and were captured with a high-quality digital camera. In this experiment, 20 identity match and 20 mismatch trials from the KFMT were employed (for more information, see Fysh & Bindemann, 2018) and as with the GFMT participants were not informed as to the ratio of match-to-mismatch trials. Example stimuli for the two face-matching tests are shown in Figure 4.4. The GFMT and KFMT tasks were presented using PsychoPy (Peirce, 2007) and were completed after the VRPC task, with the order counterbalanced across participants.

Figure 4.4. Example stimuli of match (left) and mismatch (right) trials for the GFMT (top row) and KFMT (bottom row).

Results

GFMT and KFMT performance

The mean percentage accuracy on the GFMT and KFMT is illustrated in Figure 4.5. To establish that performance in the face-matching tests conformed with previous findings, a 2 (trial type: match vs. mismatch) x 2 (face-matching task: GFMT vs. KFMT) within-subjects ANOVA was conducted. Consistent with previous work (Fysh & Bindemann, 2018), this showed a main effect of test, $F(1,99) = 217.58$, $p < .001$, $\eta_p^2 = .69$, whereby the GFMT was performed more accurately than the KFMT. A main effect of trial type was also found, $F(1,99) = 90.32$, $p < .001$, $\eta_p^2 = .48$, due to higher match than mismatch accuracy. There was no interaction between these factors, $F(1,99) = 1.34$, $p = .25$, $\eta_p^2 = .01$.

Correspondingly, a paired-samples *t*-test revealed that *d'* was higher for the GFMT (1.44) than the KFMT (0.67), $t(99) = 14.50$, $p < .001$, $d = 1.38$. In addition, a second paired-samples *t*-test for *criterion* showed a greater bias to make match responses on the GFMT (-0.45) than the KFMT (-0.36), $t(99) = 2.13$, $p = .04$, $d = 0.19$.

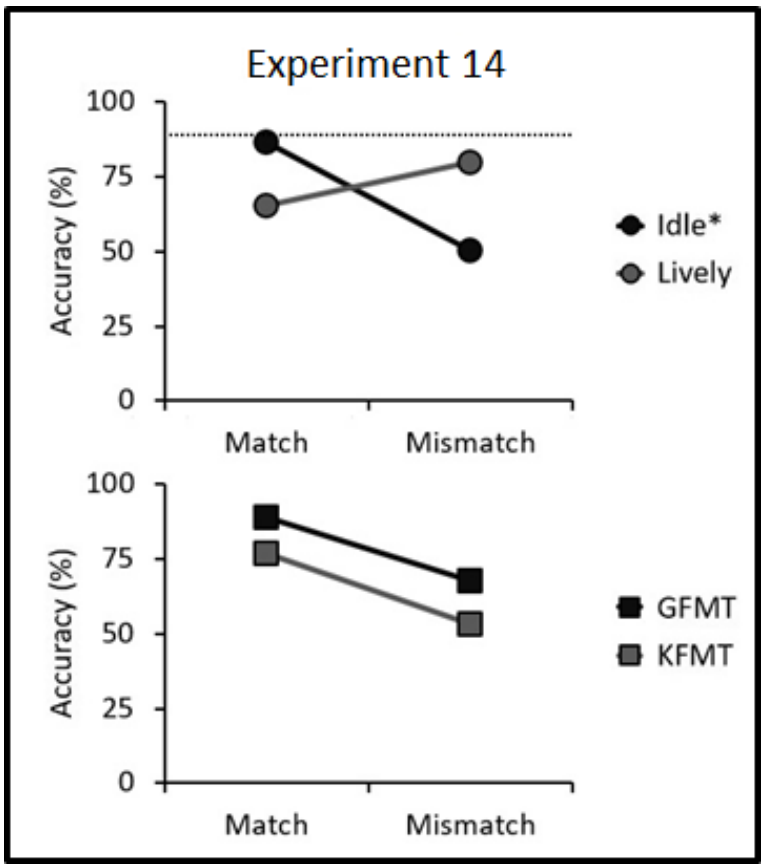

Figure 4.5. Mean accuracy scores of the critical match and mismatch trials for the idle and lively activity levels in the VRPC task (circles), and match and mismatch accuracy on the two face-matching tests (squares) for Experiment 14. The standard errors of the means are too small for the error bars to be visible. The dashed line represents the mean accuracy for the non-critical match trials on the VRPC task, the activity level of which is denoted by the asterisk (i.e., the majority activity level).

VRPC performance

A next step of the analysis sought to confirm the body language effect in VR that was observed in Experiments 11 to 13. Accuracy for the 88 non-critical match trials on the VRPC task was 89.0% (*SD* = 9.2). The mean percentage accuracy for the six critical match and six mismatch trials for the two activity levels is displayed in Figure 4.5. A 2 (trial type: match vs. mismatch) x 2 (activity level: idle vs. lively) within-subjects ANOVA of this data did not show a main effect of activity level, $F(1,99) = 3.24$, $p = .08$, $\eta_p^2 = .03$, but a main effect of trial type, $F(1,99) = 9.28$, $p = .003$, $\eta_p^2 = .09$, and an interaction between factors, $F(1,99) = 78.58$, $p < .001$, $\eta_p^2 = .44$.

Analysis of simple main effects revealed an effect of activity level for match trials, $F(1,99) = 35.25$, $p < .001$, $\eta_p^2 = .26$, since idle trials were classified more accurately than lively trials. A simple main effect of activity level was also found for mismatch trials, $F(1,99) = 61.37$, $p < .001$, $\eta_p^2 = .38$, but due to superior accuracy for lively trials. In addition, simple main effects of trial type were revealed within both the idle and the lively activity levels, $F(1,99) = 66.01$, $p < .001$, $\eta_p^2 = .40$ and $F(1,99) = 9.85$, $p = .002$, $\eta_p^2 = .09$, respectively. For the idle activity level, this was due to higher accuracy on match trials than mismatch trials, whilst the reverse pattern was observed for the lively activity level.

Finally, a paired-samples *t*-test showed *d'* did not differ for the idle (0.44) and lively (0.53) activity levels, $t(99) = 1.71$, $p = .09$, $d = 0.22$. However, *criterion* revealed a greater bias to make mismatch responses on lively (0.13) than idle trials (-0.30), $t(99) = 8.89$, $p < .001$, $d = 1.13$. Overall, these findings therefore converge with previous experiments to show that unusual, lively body language biases observers' responses, resulting in an increase in accuracy on mismatch trials and a decrease on match trials.

Correlational analyses between tasks

In the final step of this analysis, a series of Pearson correlations was carried out to assess how performance on the VRPC task relates to individual differences in face-matching performance. These correlations are illustrated in Figure 4.6. Consistent with previous research, accuracy on the GFMT and KFMT correlated both for match, $r = .529$, $p < .001$, and mismatch trials, $r = .718$, $p < .001$. Performance on these two face-matching tests was then compared to the VRPC task, both under idle and lively body language conditions. Accuracy on idle match trials correlated with GFMT, $r = .280$, $p = .005$, but not KFMT match accuracy, $r = .149$, $p = .14$. For mismatch trials, accuracy under idle body language conditions correlated with both tests, $r = .232$, $p = .02$ and, $r = .234$, $p = .02$, respectively.

In contrast to idle body language, accuracy on lively match trials did not correlate with GMFT, $r = .101$, $p = .32$, or KFMT match accuracy, $r = .141$, $p = .16$. To provide a measure of the impact of unusual body language that takes account of matching performance under normal conditions, VRPC task performance was also recalculated by subtracting accuracy on lively from idle match trials. Pearson correlations of this match score with the two face-matching tests confirmed there was no relationship between GFMT and KFMT match accuracy with the VRPC, $r = .081$, $p = .42$ and $r = -.036$, $p = .72$, respectively.

The findings extend to the mismatch conditions, for which accuracy on lively trials also did not correlate with GMFT, $r = .131$, $p = .19$, or KFMT accuracy, $r = .112$, $p = .27$. Again, this finding held when performance was recalculated by subtracting accuracy for idle from lively mismatches on the VRPC, which did not correlate with GFMT or KFMT mismatch accuracy, $r = -.112$, $p = .27$ and $r = -.127$, $p = .21$, respectively.

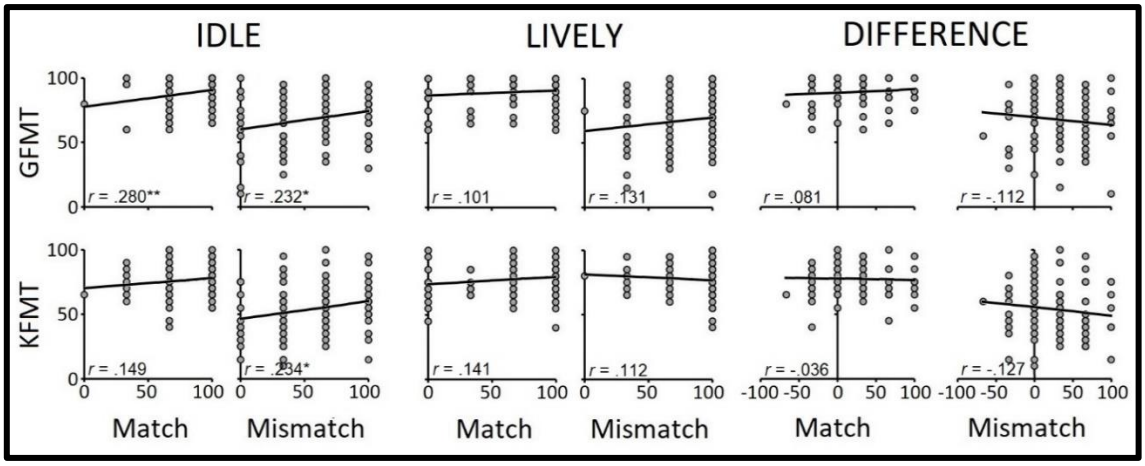

Figure 4.6. Correlations between accuracy on the idle and lively activity levels in the VRPC and the GFMT and KFMT tasks, for match and mismatch trials. The additional difference measure was used to further assess the impact of unusual body language on both trial types. Note: * $p < .05$, ** $p < .01$.

Discussion

This experiment sought to investigate whether the biasing effect of body language is attenuated by the ability to match faces. For this purpose, the VRPC task was modified to consist of frequently-occurring idle and infrequently-occurring lively activity level trials. Performance on this task was then compared with two established face matching tests, comprising of the GFMT (Burton et al., 2010) and KFMT (Fysh & Bindemann, 2018).

Overall, the VRPC task replicated the body language effect of the preceding experiments, by demonstrating that observers were more likely to classify face pairs as identity mismatches when these exhibited unusually lively body language, in comparison with the frequent idle body language trials. As expected from previous research, the GFMT also proved to be an easier face-matching task than the KFMT, and accuracy across both of these tests correlated well (Fysh & Bindemann, 2018). More importantly, individual performance on the face-matching tests also correlated with the VRPC task. This finding is consistent with previous validation of this VR paradigm in Chapter 2, in which similar correlations were reported

between the GFMT, KFMT and the matching of static avatar faces. However, such correlations were observed only for idle trials. For lively trials, on the other hand, no association between GFMT, KFMT and identification accuracy in virtual reality was found.

This combination of results indicates that the same cognitive processes are employed to complete the two face-matching tests *and* the VRPC when body language is normal. When unusually lively body language is displayed, on the other hand, this information dominates decision-making during identification to the extent that this process no longer associates with face identification ability. This suggests that, rather than drawing attention to facial identity information of passengers who require particular scrutiny during passport control, unusual body language actually undermines person identification in this environment by reducing reliance on the primary facial identity cues.

General Discussion

This chapter employed the novel VR paradigm to investigate whether body language influences person identification. In the experiments reported here, participants assumed the role of passport control officers in a VR airport and were required to process a queue of passengers by comparing their faces to passport-style photo cards. In this paradigm, identity mismatches, in which two different persons are shown, occurred infrequently to provide a closer approximate to the rare occurrence of these critical cases in real-life conditions (see, e.g., Bindemann et al., 2010; Fysh & Bindemann, 2017b, 2018; Papesh & Goldinger, 2014; Susa et al., 2019). The body language of these mismatches and a corresponding set of matches was manipulated, so that this either reflected the behaviour of the majority of passengers or was more or less active than this norm.

When participants were not informed in advance of this variation in body language in Experiments 9 and 10, this manipulation did not influence person identification at all. In

contrast, when participants were explicitly instructed to monitor variation in body language to aid detection of mismatches in Experiments 11 to 13, the detection of these cases increased. The magnitude of this effect was substantial, resulting in an increase of correct mismatch decisions of 47.1% across experiments. Thus, lively (Experiment 11) and hyper body language (Experiments 11 to 13) effectively doubled the number of mismatch detections.

The finding that body language only affects person identification with conscious monitoring deviates from previous work, which suggests that the body influences identification even when observers have limited awareness of this (Rice, Phillips, Natu, et al., 2013). In contrast to this previous research, however, body cues could not be used directly for identification in the current study, as only the avatars faces, but not the bodies, were shown on the photo-identity documents. Rather, in the current study the body language served to alert observers to the presence of potential identity mismatches, so that more attention could be given to the accurate facial identification of these infrequent trial types. This difference between studies may explain why body language needed to be actively monitored in the experiments here to exert an effect on the separable task of face matching.

Another reason as to why the body language effect hinged on explicit instruction to look for these cues may be that, without some prior warning, it may not be clear what unusual body language is and, therefore, that this is present. In the current study, the effect of body language was observed consistently when the activity level of this behaviour was higher than the norm. In contrast, when body language differed from the norm through reduced activity, this caused only a small (Experiment 12) or no increase (Experiment 13) in mismatch accuracy. This indicates that it was not unusual body language *per se* that influenced mismatch detection, but specifically an increase in behavioural activity. This effect is particularly striking in a comparison of Experiments 11 and 13. These experiments included the same body language activity levels (idle, lively and hyper), but differed in terms of which of these levels was

assigned as the norm (idle in Experiment 11, lively in Experiment 13). When unusual body language was defined as increased activity from this norm in Experiment 11, both of the unusual body language conditions affected face matching. By contrast, only increased but not decreased activity from the norm affected face matching in Experiment 13.

This finding converges with other work that shows that observers expect increased head and body movements and gaze aversion to indicate deception (Akehurst, Köhnken, Vrij, & Bull, 1996; Bogaard, Meijer, Vrij, & Merckelbach, 2016; Hartwig & Granhag, 2015; Strömwall & Granhag, 2003). Such research typically uses methods in which behaviour cues displayed whilst telling the truth and lying are recorded from video footage and compared, with observers asked to rate each cue in terms of how deceptive they perceive those behaviours to be (see a meta-analysis by DePaulo et al., 2003). A classic paradigm for deception research by DePaulo and Rosenthal (1979) involves subjects being recorded giving descriptions of people they like and dislike, providing one truthful and one deceitful account of each. These videos are watched by other subjects who provide a rating of how likely it is they are telling the truth. These ratings are compared with the actual behaviour and the videos examined for consistent cues that were present when deception was detected.

There exists a wider belief that deceptive people are more nervous than truth-tellers, so it is common to interpret increased body language as signs of deception (Vrij, 2008a). These beliefs appear to be stable across professions and laypersons (Akehurst et al., 1996; Bogaard et al., 2016; Vrij, Akehurst, & Knight, 2006), but in reality no such associations seem to exist (Bogaard et al., 2016; DePaulo et al., 2003; Hartwig & Granhag, 2015; Strömwall & Granhag, 2003; Strömwell, Granhag, & Hartwig 2004; Vrij et al., 2006). For example, in a phenomenon referred to as the *Othello error* (Ekman, 1985/2001), people can also appear nervous when they are expressing the truth, through fear of not being believed or simply from being accused (Vrij, Granhag, & Porter, 2010). In turn, there is evidence that deceptive behaviour may *actually* be

characterised by reduced body movements (Akehurst et al., 1996; Vrij, 2008a; Vrij & Mann, 2001). As a demonstration of this, observers with these stereotypical beliefs of increased body movements and gaze aversion as deceptive behavioural cues were found to be less successful at detecting the lies told in a recorded police interview by a later-convicted murder suspect (Vrij & Mann, 2001). In this context, it is poignant that reduced body language did not affect identification here. If this finding were to generalize to passport control in real-world settings, then this would imply that, while unusually hyperactive body language influences person identification, *truly* deceptive body language may not.

Another important characteristic of the body language effect emerged in the current experiments. Whenever body language exerted an influence on person identification, by increasing the number of correct mismatch decisions, this was consistently met by a corresponding decrease in match accuracy. Thus, body language did not *improve* the accuracy of person identification here, but *biased* this towards the detection of mismatches. In contrast to the typical body language conditions, for which identification accuracy correlated with established tests of face matching (i.e. GFMT - Burton et al., 2010; KFMT – Fysh & Bindemann, 2018), the biasing body language effect was not affected by individual face-matching ability (Experiment 14). This suggests that, rather than serving to focus observers' face-matching resources more strongly on suspicious cases, unusual body language diverted attention away from facial information. This finding resonates with the observation that an overemphasis on detecting deception through nonverbal behaviour can result in the adoption of inaccurate stereotypical cues (Vrij et al., 2010). A similar mechanism may underlie the effects observed here, whereby task instructions highlighting unusual body language as a potential behavioural indicator of identity mismatches resulted in a strong inclination to classify match trials with unusual body language also as identity mismatches (see Vrij, 2008b).

Put differently, these findings indicate that the available facial information to reach identity decisions here was surpassed by non-identity-specific body language cues.

This present research was motivated in part by the limited understanding of the impact of social interaction factors, such as body language, on facial identification at passport control. Some heuristic techniques to detect identity impostors, and other threats, at passport control have focused on the detection of unusual behaviour, with large-scale programmes in existence to train staff to look for such nonverbal cues (e.g., SPOT, United States Government Accountability Office, 2010). While it is not clear whether this has enhanced aviation security, the current experiments suggest that body language might only affect person identification at passport control when observers are explicitly monitoring variation in such behaviour. Under those circumstances, increases in activity, rather than behaviour that is generally different to the norm, might drive these effects. However, the current experiments indicate also that such body language is utilised strongly in a stereotypical fashion, regardless of one's face-matching ability, that biases rather than improves the accuracy of person identification. Such powerful biases may be useful in these occupational settings under particular circumstances, for example, where mismatch detection is prioritized irrespective of its cost such as the false classification of matches, or where mismatch detection is important but (not all) staff are necessarily capable of doing so (see, e.g., White, Kemp, Jenkins, Matheson, & Burton, 2014). In terms of enhancing actual accuracy, however, the current findings support the notion that behaviour detection activities provide an inadequate means to improve aviation security in real-world settings (United States Government Accountability Office, 2013).

Of course, such claims are made most tentatively. The current study presents a novel experimental approach to study face matching in more complex settings by utilising *virtual* reality. This is a novel and highly exploratory approach in this field that requires much further development, for example, to enhance the person avatars and realism of the VR environment

(see Chapter 2). In the current context, these issues are compounded by the fact that there appear to be no body language cues that uniquely associate with deceptive behaviour (DePaulo et al., 2003; Vrij, 2008a). This issue was circumvented through prior instruction of what constitutes unusual body language. As a consequence, however, the current effects may simply reflect how unusual body language was implemented. Manipulation of the activity level of the built-in idle modes of avatars provides a very limited proxy for the variety of behaviours that people can exhibit naturally outside of the laboratory. Given the novelty of this approach in the study of face matching, a wider range of behaviours would have been difficult to implement. At the least, however, the manipulation of idle modes meets the operational demands of behaviour indicative of deception, of being unusual *somehow* (United States Government Accountability Office, 2010).

Overall, this chapter has successfully demonstrated the potential application of the VRPC paradigm to investigate the influence of body language on face matching in security settings. With further development to improve the realism of the VR paradigm, more sophisticated means of manipulating avatar behaviour may be achieved. This would enable the examination of how more complex social interactions with passengers, such as engaging in conversation, could impact face-matching performance. Further prospective explorations will be discussed in the final chapter, which provides a summary of this thesis and subsequent conclusions.

Chapter 5:

Summary, Conclusions and Future Research

5.1 Summary and Conclusions

This thesis explored the potential for using virtual reality (VR) as an experimental method for investigating person identification in security settings. This often involves comparing a person's face to their purported photographic documentation, such as a passport, and determining if the identity is a match. In the first chapter, the challenging nature of this face-matching task was summarised alongside the contributing factors which heighten task difficulty, for example within-face variability (e.g., Jenkins, White, Van Montfort, & Burton, 2011; Megreya, Sandford, & Burton, 2013). Face matching has typically been investigated by using photo-to-photo comparisons, which has been a useful methodology for investigating the impact of stimuli characteristics and individual differences on task performance. However, such methods do not capture the complexities of real-world settings in which such face-matching tasks are routinely performed, for instance passport control. Owing to the security-critical nature of the task, few field studies have been conducted (e.g., White, Kemp, Jenkins, Matheson, & Burton, 2014), which present their own logistical challenges; social interaction variables are difficult to standardise in such experiments and as such additional measures are taken to prevent the intrusion of these factors. VR provides a potential solution to such problems by reducing the trade-off between experimental control and ecological validity (Blascovich et al., 2002; Bombari, Schmid Mast, Canadas, & Bachmann, 2015; de la Rosa & Breidt, 2018; Loomis, Blascovich, & Beall, 1999). The aim of this thesis was to create a successful simulation of a passport control task in VR which could be used as a methodological tool for face-matching research, and investigate the influence of social interaction cues such as body language on this task.

In Chapter 2, the suitability of avatars as a substitute for real faces was assessed across three phases. These avatars were created by mapping the internal facial features from 2D photographs onto existing 3D avatars, with two avatars created per identity to create a matching

stimulus pair; identity mismatch pairs were formed based on general visual similarity of the avatars. Although the 3D structure of the avatar faces could not be adapted to incorporate the facial shape information from the original photographs, the 2D textural information was expected to be more diagnostic for identity (see, e.g., Calder, Burton, Miller, Young, & Akamatsu, 2001; Hancock, Burton, & Bruce, 1996; Itz, Golle, Luttmann, Schweinberger, & Kaufmann, 2017), thus this method was thought to capture the original identities sufficiently. For all experiments in this chapter, an equal number of match and mismatch trials were presented in all tasks.

The first phase of experimentation assessed the quality of these avatars by comparing 2D avatar face portraits with the photographs from which they were derived. Experiment 1 consisted of a face-matching task with three stimulus types, pairing an avatar image with its original source image (same-image identity-match), with a different face photograph of the same person (different-image identity-match), or with a face photograph of a different person (identity-mismatch). Matching accuracy of the same-image identity-match trials was near-ceiling level, indicating image-specific identity information of the source images was well captured by the avatars. However, this stimulus type is not typically included in face-matching experiments (e.g., Fysh & Bindemann, 2018) to ensure that this task is not solved by using simple image-matching strategies (see, e.g., Burton, 2013; Jenkins & Burton, 2011). Furthermore, the inclusion in Experiment 1 may have resulted in the poor accuracy levels observed for the different-image identity-match trials due to the inevitably greater similarity of the same-image trials. As such, for Experiment 2 this condition was removed and subsequently accuracy for both identity-match and identity-mismatch trials was significantly above chance, providing initial evidence that avatars may provide a suitable substrate to study face-identification processes. Experiment 3 concluded this initial avatar validation phase by comparing avatar-to-avatar face matching with photograph matching of the source images.

Despite the avatar matching task being more difficult, possibly owing to the identity mismatch pairings being based on avatar similarity rather than photograph similarity, accuracy remained above chance level and correlated well with the photograph matching task, suggesting that the same underlying cognitive processes are reflected by both stimulus types.

This was further investigated in phase 2, through comparing the matching of the avatar stimuli with established face-matching tests, the widely-used GFMT (Burton, White, & McNeill, 2010) which represents a best-case scenario by providing highly-controlled same-day face photographs, and the newer, more challenging KFMT (Fysh & Bindemann, 2018) consisting of a controlled face portrait and an unconstrained image taken at least three months previously. These two face matching tests correlate well, and so the aim of the second phase in Chapter 2 was to examine if such correlations existed between these tests and the matching of avatar pairs consisting of two face portraits (Experiment 4) and also when avatar face portraits were paired with a whole body image (Experiment 5) as it would be seen in VR. In both of these experiments, matching on the GFMT, KFMT and avatar tasks consistently correlated, providing further evidence that facial identification across these stimulus types is based on similar cognitive processes.

For both the established face matching tests (GFMT and KFMT) and the avatar task, the same type of stimuli are presented, i.e. two different facial images simultaneously. However, the construction process for the VR stimuli by combining 2D face photographs with 3D avatars results in avatars which are arguably no longer real faces. Although Experiment 1 demonstrated this process retained much of the information from the original face photograph, it was clear from Experiment 3 that variance between instances of the same identity were lost through this process as the face photographs were matched with better accuracy than avatar images. This could result in the solving of the avatar task being driven by image similarity judgements rather than being processed like real faces would be. Despite these potential challenges for the stimuli,

avatar matching accuracy correlated with the matching of the original face photographs and there were also reliable correlations between face-matching performance on the GFMT and KFMT. This suggests that individuals have similar ability for matching the real face images for identity as the avatars, and therefore that similar cognitive mechanisms are being used when processing the different facial stimuli.

Whilst such processes have as yet not been measured directly, one can assume that several are required to resolve on an identity decision, although some speculation is necessary. For example, attention must initially be directed towards the stimuli of interest. It may be that both facial images are processed simultaneously, however it is also possible for attention to be directed towards one face at a time whilst storing a representation of the other in working memory. Comparisons must subsequently be made between the two faces, thus evaluative processes are utilised to either discern that despite some variance between the images they depict the same person (identity match), or that there are sufficient differences to outweigh the similarities, thus suggesting different people (identity mismatch). Individuals' ability to successfully detect matches and mismatches are dissociable, therefore successfully being able to detect one does not necessarily mean one can reliably detect the other (e.g., Megreya & Burton, 2007). This may be caused by different weight being given to these two evaluative judgements by different observers, for example some individuals may preferentially adopt the strategy of seeking out differences which suggest a mismatch over finding sufficient similarities for a match decision. This is likely to result in different accuracy scores to others taking the reverse approach, particularly when confronted with challenging stimuli (e.g., highly-similar mismatches). However, given the reliable correlations in accuracy found across the three stimuli types (GFMT, KFMT, and avatar faces) in this phase of experimentation, there appears to be sufficient evidence to suggest observers are consistent in their strategy across these three face matching tasks, thus using similar cognitive processes for task completion.

The final phase of Chapter 2 assessed whether the avatars provide sufficient information for person identification during immersion in a VR passport control environment. The VR airport was constructed so that visual cues as to the nature of the environment were clearly visible to the observer, such as departure boards and a waiting aeroplane within view of the passport control desk area. During the first iteration of the VR paradigm (Experiment 6), an avatar would approach and their respective passport photo appeared on a screen within the VR desk area. Participants were immersed in the VR with an Oculus Rift DK2 headset and operated the task using a computer mouse, recording their identity match or mismatch decisions via button presses. Once a response was recorded, the avatar would depart and the next in line would proceed to the desk area to be processed, with the participant tasked to clear the whole queue. This avatar matching task was found to be increasingly difficult when performed in VR compared to the static image tasks of Experiments 4 and 5. As such, a second iteration of the VR paradigm was attempted to optimise performance; in Experiment 7, the Oculus Rift DK2 headset was replaced with an HTC Vive, which provided an improved screen resolution and was equipped with handheld controlled to enable participants to better interact with the environment. A final modification consisted of the re-recording of avatar face portraits to produce a more natural face shape as it had been noted the original avatar face stimuli were narrow in appearance. These new face portraits were then inset into a passport-style card so that participants could bring these closer to their own face or that of the avatar to facilitate comparison. These modifications successfully improved accuracy for both match (from 59% to 77%) and mismatch (from 39% to 48%) trials. However, given the importance of accurate identification of mismatches in real-world security settings, such accuracy levels for mismatch trials would be a limiting factor for research.

On the other hand, it was clear that by-item difference existed for this trial type across Experiments 4 to 7, and subsequent analysis revealed that mismatch items ranged in accuracy

from as little as 3% to near-ceiling level at 97% in Experiment 7. When modelling the real world of passport control, match trials should occur with much greater frequency than mismatch trials (see, e.g., Bindemann, Avetisyan, & Blackwell, 2010; Fysh & Bindemann, 2017b, 2018; Papesh & Goldinger, 2014; Susa, Michael, Dessenberger, & Meissner, 2019), and so a solution would be to select mismatch items for VR experimentation with sufficiently high accuracy. Following on from these initial VR experiments, in Chapters 3 and 4 rather than tasks consisting of equal number of match and mismatch trials as used in Chapter 2, the VR tasks proceeded to have 100 trials consisting of six mismatch trials and 94 match trials. Of these, six were identified as critical match trials for the analyses which were accuracy matched to the six mismatch trials based on the data from Experiments 4 to 7.

Across the three phases, Chapter 2 successfully validated the avatars as a suitable substrate for facial identification research. The avatars created here captured sufficient diagnostic information for identity such that avatar face matching reflected the same cognitive processes as the matching of facial photographs. Whilst the VR task provided additional challenges, by appropriately selecting the mismatches which are to appear infrequently during the passport control task, a suitable simulation of the real-world task can be created. Chapter 3 sought to investigate whether such simulations could become a useful assessment tool for personnel selection in security settings. It has become clear that selecting observers with an aptitude for face matching may provide a viable alternative to training (Bobak, Dowsett, & Bate, 2016; Lander, Bruce, & Bindemann, 2018), since this has not been effective for improving task performance (e.g., Towler et al., 2019).

Experiment 8 compared performance on the VR passport control task (VRPC) with performance on the GFMT and KFMT (which had previously been correlated with static avatar stimuli in Experiments 4 and 5), the CFPT (Duchaine, Germine, & Nakayama, 2007) and a self-report test of face-processing ability, the PI20 (Shah, Gaule, Sowden, Bird, & Cook, 2015).

This was included owing to the ease of administration of self-report measures in applied settings and there was evidence to suggest that scores on the PI20 correlate with facial identity comparison ability (Shah, Sowden, Gaule, Catmur, & Bird, 2015). This remained consistent in Experiment 8, with PI20 scores correlating with accuracy on the three face-perception tasks, which in turn correlated with one another, but no such relationship was found with the VRPC task. Considering match and mismatch trials separately, match accuracy on the VRPC task was consistently near ceiling level, thus the absence of correlations with the more diversely performed GFMT and KFMT match trials may be attributed to this. Regarding mismatch trials, performance on the VRPC task correlated well with the GFMT and KFMT accuracy, indicating that those observers capable of detecting identity mismatches on these established laboratory tests also performed well on the more difficult VRPC task. At passport control, it is reasonable to assume that those seeking to evade detection would attempt to make the task as difficult as possible by using valid documentation of someone of similar face appearance to themselves (Bindemann, Fysh, Cross, & Watts, 2016; Meissner, Susa, & Ross, 2013; Susa et al., 2019). A primary objective for passport control staff is to detect these difficult and infrequently occurring imposters, therefore the association of mismatch performance on laboratory tests of face matching with mismatch detection in the challenging VRPC suggests that such tasks could be a useful assessment for those undertaking roles in these security settings.

Whilst accurate discrimination is clearly necessary for successfully completing this task, arguably there is a greater importance in security settings for the identification of mismatches at the potential cost of misclassifying identity matches. Those attempting to pass through security checkpoints with documentation belonging to another person can be assumed to have criminal intentions, therefore the detection of these individuals is essential for maintaining security. This could be taken into consideration when selecting personnel for such environments, for example those with a predisposed bias to report mismatches may be more

suitable for the role that those who overlook differences in favour of similarities, and thus overclassify matches. Of course, in real-world settings this trade-off needs careful management, as to be overly cautious would result in lengthy delays should large volumes of passengers have their documentation questioned. Experiment 8 has demonstrated not only that accuracy correlates between laboratory face-matching tests and the simulated passport control task, but also individual response bias as measured by *criterion*. Therefore, should response bias be taken into account for personnel selection tasks, the VRPC would be a reliable assessment tool in conjunction with the laboratory tests.

Further development of this VRPC task is required to capture the real-world more effectively in order to provide a detailed simulation of the passport control environment, yet this is promising given the analogous simulations already in frequent use for the training and selection of personnel in high stakes professions. For example, pilots are routinely assessed using flight simulators in which they can be tested with extreme manipulations to the environment, such as unpredictable weather conditions and malfunctions to their controls. Since VR allows for these controlled manipulations, the assessment of personnel for passport control may also incorporate variables which could influence face matching that cannot currently be evaluated with typical laboratory tests, for example social interaction factors. The display of unusual body language compared to the norm in security environments may be perceived as an indicator of deceptive behaviour. Body information has been demonstrated to facilitate person identification, particularly when facial cues are insufficient to determine identity alone (e.g., Hahn, O'Toole, & Phillips, 2016; Rice, Phillips, & O'Toole, 2013), although observers may not be aware of their reliance on such information (Rice, Phillips, Natu, An, & O'Toole, 2013). The influence of body cues which are not indicative of identity, but which could reflect a hidden motivation, have yet to be investigated in relation to face matching and so the impact of body language on such decisions was investigated in Chapter 4.

In this final experimental chapter, the body language of critical match and mismatch trials was manipulated so that selected avatars behaved differently to the norm whilst being processed. The aim was to examine whether this alternate display of body language would be perceived as unusual in the passport control context and therefore whether identity mismatches would be more likely to be detected when exhibiting unusual body language and, in turn, more frequently missed when not. In the VRPC task, the avatars were equipped with an idle animation mode which creates small shifts in body posture when a person is stationary, to increase observers' sense of realism in VR. For the majority of these avatars, this animation was set to the same level to establish a normal behaviour whilst in selected cases the idle level was raised to simulate restless body language. For the experiments in Chapter 4, there were typically three body language conditions evenly distributed across the six critical match and six mismatch trials of the VRPC task. The "normal" body language condition per experiment was established through its assignment to the remaining 88 non-critical match trials.

In Experiments 9 and 10, the two unusual body language conditions were of higher activity levels to the norm and participants were given no instruction to attend to body language. Participants demonstrated limited awareness of the body language manipulations, even when the differences between conditions was exaggerated from Experiment 9 to Experiment 10. In both experiments, match accuracy was higher than mismatch accuracy, yet body language had no influence, which suggested conscious monitoring of such cues may be required to influence person identification decisions. Therefore, from Experiment 11, participants were explicated instructed to monitor body language during the task. This had a profound effect on face matching such that observers employed unusual body language as a heuristic to support mismatch decisions; in the presence of unusual body language, detection of mismatches increased substantially at the expense of false classification of matches, yet were frequently missed when such cues were not available. Experiments 12 and 13 extended these

findings by establishing that this biasing effect was driven by increased expressive body language rather than body language that differs to the norm *per se*.

These findings converge with general beliefs concerning deceptive behaviour, whereby the expression of nervousness through increased body movements is widely perceived to be a cue of deception (e.g., Hartwig & Granhag, 2015; Strömwall & Granhag, 2003) when in fact no such associations exist (e.g., Bogaard, Meijer, Vrij, & Merckelbach, 2016; DePaulo et al., 2003) and truly deceptive behaviour may actually be characterised by reduced body movements (Akehurst, Köhnken, Vrij, & Bull, 1996; Vrij, 2008a; Vrij & Mann, 2001). Chapter 4 concluded with an examination of whether the biasing effect of body language is attenuated by individual differences in the ability to match faces. The VRPC task with body language manipulations was compared with accuracy on the GFMT and KFMT in Experiment 14. Correlations for accuracy were found across all three tasks, but only for the "normal" body language condition; no association between GFMT, KFMT and identification accuracy was present under the unusual body language condition, suggesting this biasing effect is robust across individual face-matching ability. It is noted that the experiments in this chapter used simple manipulations of avatar activity level through changing animation speed in order to provide an arbitrary indicator of unusual body language relative to a norm. This research could progress further with the examination of specific behaviours, such as impatience, and their respective effect on face matching.

These findings highlight the effect of response bias on face-matching, which can be defined as the tendency to give a particular response towards stimuli. As demonstrated above, observers tend to suspect unusually active avatars in the VRPC task as mismatches when directed to attend to body language. Whilst this seems to improve accuracy for mismatch detection, this differs from improving discrimination between trial types. This response bias consequently leads to an increase in false challenges towards matches should their activity

levels exceed the norm, and therefore observers' general ability to discriminate between match and mismatch trials does not improve.

By performing an ongoing face-matching task, observers may adapt their strategies towards the task, particularly when receiving feedback, to focus on those details relevant to the task (White, Kemp, Jenkins, & Burton, 2014). In this instance, knowing in advance that mismatches would not be prevalent and being asked to attend to body language led to the association of unusual body language being indicative of mismatches causing a response bias. It is important to note, however, that participants only gave a mismatch-biased response towards those which displayed more expressive body language compared to the norm, as opposed to any unusual behaviour such as reduced movement. This suggests instructions to attend to body language resulted in participants preconceiving what behaviours would be suspicious in the passport control scenario and biasing their responses accordingly. Papesh, Heisick and Warner (2018) also demonstrate how face matching performance can be biased as participants learn task-relevant details through feedback. During their investigations, participants were initially unaware of trial type frequency, however as they received trial-by-trial feedback their response bias shifted in accordance to the perceived likelihood of mismatch prevalence. For example, when presented with infrequent mismatches participants were more biased to match responses, thus increasing their miss rate and decreasing the number of false mismatch classification. The reverse pattern could be seen when mismatches were highly prevalent, whilst equal numbers of miss and false alarms were given when match and mismatch frequency was equal. Although this prevalence effect caused a shift in response bias (measured by *criterion*), it had no effect on participants discrimination abilities as measured by *d'*. This further demonstrates the dissociation between response bias and discrimination for face matching, with the specifics of the task affecting performance.

This thesis has demonstrated the application of VR for researching person identification in real-world security settings, and at a theoretical level can provide insights into the cognitive processes that govern face matching. A primary challenge of research in this field is being able to accurately measure face-processing ability. Individuals not only vary in their performance between one another (e.g., Burton et al., 2010; Duchaine & Nakayama, 2006) but also across different face-matching tasks (e.g., McCaffery, Robertson, Young, & Burton, 2018) and over time (Bindemann, Avetisyan, & Rakow, 2012). Whilst tasks may produce an overall accuracy score to reflect how successfully individuals completed the specific task demands, their inherent face skills cannot be quantified, since the tasks themselves cannot determine the cognitive processes used. Multiple face-processing tasks correlate in terms of accuracy, for example the GFMT with the KFMT (Fysh & Bindemann, 2018), PI20 (Shah, Sowden, et al., 2015), and a real-world passport task (Balsdon, Summersby, Kemp, & White, 2018), as also demonstrated throughout this thesis between the GFMT, KFMT and the VRPC tasks.

However, the moderate correlations obtained are not substantial enough to suggest the same perceptual mechanisms are used for all face tasks and performance is clearly affected by task-specific demands; McCaffery and colleagues (2018) estimate that approximately 25% of the variance observed across face-processing tasks can be accounted for by a general face-perception factor. This is further demonstrated by super-recognisers as a group outperforming controls yet individuals are not superior on all tasks (Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Bobak, Dowsett, & Bate, 2016). Furthermore, how laboratory face-matching ability translates to real-world ability remains to be seen. The research in this thesis has demonstrated that the correlations between laboratory tasks are much higher than their respective correlations with the simulated real-world task. Matching controlled photographs of faces in isolation is not reflective of what would typically be performed outside of the laboratory; distractors in visual scenes are often present alongside variable environmental conditions such as lighting and noise.

The VRPC is a more representative task of what would be performed in security settings than the photographic comparisons conducted for the GFMT and KFMT. Whilst the two face-matching tests consistently produce moderate-to-strong correlations, suggesting they capture face-processing ability under controlled conditions well, without improved correspondence to a real-world task, such as the VRPC, it is difficult to establish how face matching translates across these settings from laboratory tasks alone.

Performance on single tasks is evidently not demonstrative of actual face-processing ability and multiple tests must therefore be employed to effectively evaluate individual aptitude. Assessment task demands arguably should reflect realistic conditions in order to draw valid conclusions pertaining to real-world ability. Using multiple tasks with highly controlled paradigms may indeed produce a purer measure of laboratory face-matching ability, yet how this corresponds to real-world ability needs further examination. This leads to the question of *how* face matching is solved in real-world settings. Ascertaining how an individual's laboratory-based face matching reflects their ability to perform such tasks in the field effectively (Ramon, Bobak, & White, 2019), and whether the same cognitive mechanisms are used, remains challenging. There is currently limited data to correlate face-matching performance across these settings (Lander, Bruce, & Bindemann, 2018). VR provides a closer proximate to the real-world task whilst maintaining experimental control, and thus with further development may enable an ecologically valid assessment of face-matching ability in conjunction with laboratory tasks. Considered analysis of the methods used by observers to complete the VRPC task may therefore provide greater insight into the perceptual processes which may be utilised in real-world environments.

The strategies individuals use to match faces for identity in laboratory tasks, whilst difficult to capture, may provide a useful indictor of ability. For example, the criterion for discerning when two faces are sufficiently similar in appearance to be classified as a match (or

indeed dissimilar enough to be a mismatch) may vary across individuals with different levels of success and thus reflect different face-processing abilities. In addition, this may vary according to task demands; in a passport control setting with the majority of instances known to be matches the threshold for mismatch classification may be raised compared to tasks consisting of an even distribution of match and mismatch trials (see Papesh & Goldinger, 2014). These conditions are important to consider when establishing a theoretical understanding of face matching, as different cognitive processes may be required to overcome task idiosyncrasies. In real-world tasks, further variables may impact on task performance and alter the strategies used by observers. Given that faces capture attention and the presence of multiple faces impairs matching accuracy (Bindemann, Burton, & Jenkins, 2005; Bindemann, Sandford, Gillatt, Avetisyan, & Megreya, 2012; Megreya & Burton, 2006b), at passport control the mere presence of queues may provide sufficient distraction to the individual being processed by drawing the attention of the passport officer.

Research incorporating eye-tracking paradigms would provide insight into the strategies used by observers when matching faces for identity. There is evidence to suggest that focusing on central facial areas promotes better recognition; super-recognisers tend to fixate longer on these features, particularly the nose region, than controls when scanning visual scenes containing people (Bobak, Parris, Gregory, Bennetts, & Bate, 2017). When under time pressure, as would be expected at passport control, few fixations can be made and so developing effective strategies would be important to perform matching tasks well. Özbek and Bindemann (2011) found that match trials of face pairs could be accurately matched with three fixations, one to each face and an additional fixation. For mismatch trials, on the other hand, further fixations were required to achieve optimal accuracy. Evaluations of face matching strategies via examination of eye movements may prove more effective than self-reported descriptors, since individuals appear to have somewhat limited insight into their own face-

matching ability (e.g., Bobak, Pampoulov, & Bate, 2016; Palermo et al., 2017; c.f., Ventura, Livingston, & Shah, 2018). Observers also seem unaware of *how* they perform tasks, for instance when facial information is degraded they often unconsciously resort to body cues to inform their identification decisions (Rice, Phillips, Natu, et al., 2013). Eye-tracking can now be incorporated in VR paradigms using equipment such as the Tobii Vive Pro Eye. In addition to being able to track observers eye movements in a live environment, this would also allow for increased interaction with avatars, which can be programed to respond to eye movements (e.g., to maintain or avoid eye contact), providing further routes to investigate the influences of social interaction on face matching.

5.2 Future Research

Overall, this thesis has provided a proof of concept that VR has a future in face-matching research. VR can provide a useful alternative to field studies where logistical barriers, such as the security-critical nature of passport control, prevent detailed examination of the impact real-world factors. Whilst the research in this thesis has demonstrated avatars are processed in the same manner as photographic facial stimuli, further developments are required for the avatars to become closer representations of reality. The creation and visual realism of avatars is one of the central challenges facing VR (Bombari et al., 2015; Loomis et al., 1999; Pan & Hamilton, 2018) and a relatively simplistic approach was adopted for this research, by superimposing 2D facial photographs onto existing avatar structures. In future, it is anticipated that the realism of avatars can be enhanced by rigging 3D structural information of real faces into avatars through facial scans. This method would initially be costly and time-consuming, a typical challenge of VR experimentation (de la Rosa & Breidt, 2018), hence was not selected for avatar construction in this research; as VR is completely new to face matching, evidence of its efficacy was

required before employing expensive resources. A demonstration of the detail which can be captured through facial scanning is presented in Figure 5.1. This may provide a long-term solution to building a database of visually realistic person stimuli for VR experimentation. This principle can also be applied to the creation of realistic environments. Scanning locations using 360° photography could create detailed replicas of real-life settings and provide a convincing immersive experience.

Figure 5.1. Comparison of avatar construction methods. The left image is an avatar created from the centre face photograph using the method described in Chapter 2. The photograph was taken at the time of a facial scan, the outcome of which after initial rendering is shown in the right image.

A further avenue for improvement is the development of realistic avatar movement. In the VRPC task created for this research, avatars were programed to make small postural shifts whilst queuing to appear more natural than standing rigid. The same animation sequence was applied to all avatars but with random initial starting points to prevent synchronised movement. An alternative would be to incorporate real body movements into the avatars which is recorded using motion capture technology. This would also allow for further experimentation examining body language influences on face matching. Actors could be employed to exhibit specific

behaviours, for example impatience, to be recorded and then rigged into avatar movement. Subsequent experiments may examine the specific impact of these behaviours on task performance; encountering impatient passengers is a realistic source of time pressure for passport officers, which can be expected to be detrimental to face matching performance (see, e.g., Bindemann et al. 2016), and could be simulated in VR. Furthermore, by employing eye-tracking methods and programing avatars to respond to human movements, more sophisticated behaviours can be manipulated. In future it may become possible for avatars to be programed to engage in conversations, thus extending the body language paradigm from examining non-verbal cues of deception to include verbal cues, for instance lying about their reason for travelling.

Ultimately, it is expected VR will become an important research tool for investigating face perception in complex and realistic environments. The input of professionals, such as passport control staff, and increasing collaboration between researchers and developers will accelerate advancement in this field. A representative simulation of passport control will require the incorporation of many more factors such as variation in passenger ages and ethnicities, verbal interaction with passengers, and the completion of concurrent tasks. This thesis has provided a foundation onto which this can be built; it is hoped that it will spark new avenues for experimentation in security settings in which face matching takes a primary role.

References

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. *Applied Cognitive Psychology*, *10*(6), 461-471. doi:10.1002/(SICI)1099-0720(199612)10:6<461::AID-ACP413>3.0.CO;2-2

Alenezi, H. M., & Bindemann, M. (2013). The effect of feedback on face-matching accuracy. *Applied Cognitive Psychology*, *27*(6), 735-753. doi:10.1002/acp.2968

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ, 3,* e1184. doi:10.7717/peerj.1184

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *The Quarterly Journal of Experimental Psychology*, *68*(10), 2041–2050. doi:10.1080/17470218.2014.1003949

Armann, R. G. M., Jenkins, R., & Burton, A. M. (2016). A familiarity disadvantage for remembering specific images of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(4), 571-580. doi:10.1037/xhp0000174

Bailenson, J. N., Beall, A. C., Blascovich, J., & Rex, C. (2004). Examining virtual busts: Are photogrammetrically generated head models effective for person identification? *Presence: Teleoperators and Virtual Environments*, *13*(4), 416-427. doi:10.1162/1054746041944858

Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, *29*(7), 819-833. doi:10.1177/0146167203029007002

Bailenson, J. N., Blascovich, J., Beall, A. C., & Noveck, B. (2006). Courtroom applications of virtual environments, immersive virtual environments, and collaborative virtual environments. *Law & Policy*, *28*(2), 249-270. doi:10.1111/j.1467-9930.2006.00226.x

Bailenson, J. N., Davies, A., Blascovich, J., Beall, A. C., McCall, C., & Guadagno, R. E. (2008). The effects of witness viewpoint distance, angle, and choice on eyewitness accuracy in police lineups conducted in immersive virtual environments. *Presence: Teleoperators and Virtual Environments*, *17*(3), 242-255. doi:10.1162/pres.17.3.242

Balas, B., & Pearson, H. (2017). Intra- and extra-personal variability in person recognition. *Visual Cognition*, *25*(4-6), 456-469. doi:10.1080/13506285.2016.1274809

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, 3:25. doi:10.1186/s41235-018-0114-7

Bate, S., Adams, A., Bennetts, R., & Line, H. (2017). Developmental prosopagnosia with concurrent topographical difficulties: A case report and virtual reality training programme. *Neuropsychological Rehabilitation*, *29*(8), 1290-1312. doi:10.1080/09602011.2017.1409640

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., ... & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, 3:22. doi:10.1186/s41235-018-0116-5

Bennetts, R. J., Kim, J., Burke, D., Brooks, K. R., Lucey, S., Saragih, J., & Robbins, R. A. (2013). The movement advantage in famous and unfamiliar faces: A comparison of point-light displays and shape-normalised avatar stimuli. *Perception*, *42*(9), 950-970. doi:10.1068/p7446

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar-face matching. *Applied Cognitive Psychology*, *27*(6), 707–717. doi:10.1002/acp.2970

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. doi:10.1037/a0021893

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*(3), 277–291. doi:10.1037/a0029635

Bindemann, M., Burton, A. M., & Jenkins, R. (2005). Capacity limits for face processing. *Cognition*, *98*(2), 177–197. doi:10.1016/j.cognition.2004.11.004

Bindemann, M., Burton, A. M., Leuthold, H., & Schweinberger, S. R. (2008). Brain potential correlates of face recognition: Geometric distortions and the N250r brain response to stimulus repetitions. *Psychophysiology*, *45*(4), 535–544. doi:10.1111/j.1469-8986.2008.00663.x

Bindemann, M., Fysh, M., Cross, K., & Watts, R. (2016). Matching faces against the clock. *i-Perception*, *7*(5), 1–18. doi:10.1177/2041669516672219

Bindemann, M., Fysh, M. C., Sage, S. S. K., Douglas, K., & Tummon, H. M. (2017). Person identification from aerial footage by a remote-controlled drone. *Scientific Reports*, *7*(1), 13629. doi:10.1038/s41598-017-14026-3

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception*, *40*(5), 625–627. doi:10.1068/p7008

Bindemann, M., Sandford, A., Gillatt, K., Avetisyan, M., & Megreya, A. M. (2012). Recognising faces seen alone or with others: Why are two heads worse than one? *Perception*, *41*(4), 415–435. doi:10.1068/p6922

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social

psychology. *Psychological Inquiry*, *13*(2), 103-124. doi:10.1207/S15327965PLI1302_01

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition abilities. *Cortex*, *82*, 48-62. doi:10.1016/j.cortex.2016.05.003

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLoS ONE*, *11*(2), e0148148. doi:10.1371/journal.pone.0148148

Bobak, A. K., Hancock, P. J. B., & Bate, S. (2016). Super-recognisers in Action: Evidence from Face-matching and Face Memory Tasks. *Applied Cognitive Psychology*, *30*(1), 81–91. doi:10.1002/acp.3170

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology*, 7:1378. doi:10.3389/fpsyg.2016.01378

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *The Quarterly Journal of Experimental Psychology*, *70*(2), 201-217. doi:10.1080/17470218.2016.1161059

Bogaard, G., Meijer, E. H., Vrij, A., & Merckelbach, H. (2016). Strong, but wrong: Lay people's and police officers' beliefs about verbal and nonverbal cues to deception. *PLoS ONE*, *11*(6), e0156615. doi:10.1371/journal.pone.0156615

Bolt, D. (2015). *Inspection of Border Force operations at Heathrow Airport, June – October 2014*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/547614/Inspection-of-Border-Force-Heathrow-15.07.2015.pdf

Bolt, D. (2016). *An inspection of Border Force operations at Manchester Airport, June – October 2015*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549240/ICIBI_inspection_Border_Force_operations_Manchester_Airport_April_2016.pdf

Bolt, D. (2018). *An inspection of Border Force operations at Stansted Airport*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/695287/An_Inspection_of_Border_Force_Operations_at_Stansted_Airport.pdf

Bombari, D., Schmid Mast, M., Canadas, E., & Bachmann, M. (2015). Studying social interactions through immersive virtual environment technology: Virtues, pitfalls, and future challenges. *Frontiers in Psychology*, 6:869. doi:10.3389/fpsyg.2015.00869

Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual Cognition*, *10*(5), 527-536. doi:10.1080/13506280244000168

Bruce, V. (1994). Stability from variation: the case of face recognition. The M.D. Vernon memorial lecture. *The Quarterly Journal of Experimental Psychology*, *47*(1), 5–28. doi:10.1080/14640749408401141

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: Applied*, *5*(4), 339–360. doi:10.1037/1076-898X.5.4.339

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. doi:10.1037/1076-898X.7.3.207

Burgoon, J. K., Guerrero, L. K., & Manusov, V. (2011) Nonverbal signals. In M. L. Knapp & J. A. Daly (Eds.), *The SAGE handbook of interpersonal communication* (4th ed., pp. 239-281). London: SAGE Publications Inc.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. doi:10.1080/17470218.2013.800125

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*(3), 256–284. doi:10.1016/j.cogpsych.2005.06.003

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943–958. doi:10.1111/j.2044-8295.2011.02039.x

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223. doi:10.1111/cogs.12231

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. doi:10.3758/BRM.42.1.286

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243–248. doi:10.1111/1467-9280.00144

Butcher, N., & Lander, K. (2017). Exploring the motion advantage: Evaluating the contribution of familiarity and differences in facial motion. *The Quarterly Journal of Experimental Psychology*, *70*(5), 919-929. doi:10.1080/17470218.2016.1138974

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208. doi:10.1016/S0042-6989(01)00002-5

Cha, M., Han, S., Lee, J., & Choi, B. (2012). A virtual reality based fire training simulator integrated with fire dynamics data. *Fire Safety Journal, 50*, 12-24. doi:10.1016/j.firesaf.2012.01.004

Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition*, *26*(4), 780-790. doi:10.3758/BF03211397

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985-994. doi:10.1068/p3335

Cutler, B. L., Daugherty, B., Babu, S., Hodges, L., & Van Wallendael, L. (2009). Creating blind photoarrays using virtual human technology: A feasibility test. *Police Quarterly*, *12*(3), 289-300. doi:10.1177/1098611109339892

Daugherty, B., Babu, S., Cutler, B., & Hodges, L. (2007, November). Officer Garcia: a virtual human for mediating eyewitness identification. In *Proceedings of the 2007 ACM symposium on Virtual reality software and technology* (pp. 117-120). Retrieved from https://www.cs.clemson.edu/vegroup/pubs/2007/2007_gracia.pdf

Daugherty, B., Van Wallendael, L., Babu, S., Cutler, B., & Hodges, L. F. (2008). Virtual human versus human administration of photographic lineups. *IEEE Computer Graphics and Applications*, *28*(6), 65-75. doi:10.1109/MCG.2008.125

Davies, G., & Thasen, S. (2000). Closed-circuit television: How effective an identification aid? *British Journal of Psychology*, *91*(3), 411-426. doi:10.1348/000712600161907

Davis, J. P., & Valentine, T. (2009). CCTV on Trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*(4), 482–505. doi:10.1002/acp

de la Rosa, S., & Breidt, M. (2018). Virtual reality: A new track in psychological research. *British Journal of Psychology, 109*(3), 427-430. doi:10.1111/bjop.12302

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118. doi:10.1037/0033-2909.129.1.74

DePaulo, B. M., & Rosenthal, R. (1979). Telling lies. *Journal of Personality and Social Psychology*, *37*(10), 1713-1722. Retrieved from https://pdfs.semanticscholar.org/2469/f05fdeb3295c253951b96a96617dd59c3c37.pdf

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*(3), 433-445. doi:10.1111/bjop.12103

Dowsett, A. J., Sandford, A., & Burton, A. M. (2016). Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. *The Quarterly Journal of Experimental Psychology, 69*(1), 1-10. doi:10.1080/17470218.2015.1017513

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*(4), 419-430. doi:10.1080/02643290701380491

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576-585. doi:10.1016/j.neuropsychologia.2005.07.001

Ekman, P. (2001). *Telling lies: Clues to deceit in the marketplace, politics and marriage*. New York: Norton. (Original work published 1985)

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, *32*(1), 88-106. doi:10.1080/00332747.1969.11023575

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. doi:10.1068/p080431

Estudillo, A. J., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *i-Perception*, *5*(7), 589–601. doi:10.1068/i0669

Fletcher, K. I., Butavicius, M. A., & Lee, M. D. (2008). Attention to internal face features in unfamiliar face matching. *British Journal of Psychology*, *99*(3), 379-394. doi:10.1348/000712607X235872

Fox, J., Arena, D., & Bailenson, J. N. (2009). Virtual reality: A survival guide for the social scientist. *Journal of Media Psychology*, *21*(3), 95-113. doi:10.1027/1864-1105.21.3.95

Frowd, C., Bruce, V., McIntyre, A., & Hancock, P. (2007). The relative importance of external and internal features of facial composites. *British Journal of Psychology*, *98*(1), 61-77. doi:10.1348/000712606X104481

Fysh, M. C. (2018). Individual differences in the detection, matching and memory of faces. *Cognitive Research: Principles and Implications*, *3*:20, 1-12. doi:10.1186/s41235-018-0111-x

Fysh, M. C., & Bindemann, M. (2017a). Forensic face matching: A review. In M. Bindemann & A. M. Megreya (Eds.), *Face processing: Systems, Disorders and Cultural Differences* (pp. 1–20). New York: Nova Science Publishers, Inc

Fysh, M. C., & Bindemann, M. (2017b). Effects of time pressure and time passage on face-matching accuracy. *Royal Society Open Science*, *4*(6), 170249. doi:10.1098/rsos.170249

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of* Psychology, *109*(2), 219-231. doi:10.1111/bjop.12260

Gray, K. L. H., Bird, G., & Cook, R. (2017). Robust associations between the 20-item prosopagnosia index and the Cambridge Face Memory Test in the general population. *Royal Society Open Science*, 4:160923. doi:10.1098/rsos.160923

Hahn, C. A., O'Toole, A. J., & Phillips, P. J. (2016). Dissecting the time course of person recognition in natural viewing environments. *British Journal of Psychology*, *107*(1), 117-134. doi:10.1111/bjop.12125

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337. doi:10.1016/S1364-6613(00)01519-9

Hartwig, M., & Granhag, P. A. (2015). Exploring the nature and origin of beliefs about deception: Implicit and explicit knowledge among lay people and presumed experts. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting deception: Current challenges and cognitive approaches* (pp. 125-154). Chichester, UK: Wiley

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

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, *15*(4), 445–464. doi:10.1002/acp.718

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. Journal of Experimental Psychology. *Human Perception and Performance*, *22*(4), 986–1004. doi:10.1037/0096-1523.22.4.986

Hoepman, J.-H., Hubbers, E., Jacobs, B., Oostdijk, M., & Wichers Schreur, R. (2006). Crossing borders: Security and privacy issues of the European e-Passport. In H. Yoshiura, K. Sakurai, K. Rannenberg, Y. Murayama, & S. Kawamura (Eds.), *Advances in Information and Computer Security* (pp. 152–167). doi:10.1007/11908739

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception, 31*(10), 1221–1240. doi:10.1068/p3252

Itz, M. L., Golle, J., Luttmann, S., Schweinberger, S. R., & Kaufmann, J. M. (2017). Dominance of texture over shape in facial identity processing is modulated by individual abilities. *British Journal of Psychology*, *108*(2), 369–396. doi:10.1111/bjop.12199

Jenkins, R., & Burton, A. M. (2008a). Limitations in facial identification: The evidence. *Justice of the Peace*, *172*(January), 4–6. Retrieved from http://www.visimetrics.com/docs/technical/Limitations in Facial Recognition Article.pdf

Jenkins, R., & Burton, A. M. (2008b). 100% accuracy in automatic face recognition. *Science*, *319*(5862), 435. doi:10.1126/science.1149656

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

Jenkins, R., Burton, A. M., & White, D. (2006, April). Face recognition from unconstrained images: Progress with prototypes. A*utomatic Face and Gesture Recognition, 2006. FGR06. 7th international conference on*, 25-30. Retrieved from https://www.researchgate.net/profile/David_White25/publication/4232814_Face_Recognition_from_Unconstrained_Images_Progress_with_Prototypes/links/02e7e537ac470c7db7000000/Face-Recognition-from-Unconstrained-Images-Progress-with-Prototypes.pdf

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. doi:10.1016/j.cognition.2011.08.001

Johnston, R. A., & Bindemann, M. (2013). Introduction to forensic face matching. *Applied Cognitive Psychology*, *27*(6), 697–699. doi:10.1002/acp.2963

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*(5), 577–596. doi:10.1080/09658210902976969

Kemp, R. I., Caon, A., Howard, M., & Brooks, K. R. (2016). Improving unfamiliar face matching by masking the external facial features. *Applied Cognitive Psychology*, *30*(4), 622-627. doi:10.1002/acp.3239

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*(3), 211-222. doi:10.1002/(SICI)1099-0720

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Boston, MA: Wadsworth.

Knight, B., & Johnston, A. (1997). The role of movement in face recognition. *Visual Cognition*, *4*(3), 265-273. doi:10.1080/713756764

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2013). Effects of aging on face identification and holistic face processing. *Vision Research, 88*, 38-46. doi:10.1016/j.visres.2013.06.003

Kramer, R. S. S., Jenkins, R., Young, A. W., & Burton, A. M. (2016). Natural variability is essential to learning new faces. *Visual Cognition*, *25*(4-6), 470-476. doi:10.1080/13506285.2016.1242522

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising Superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology*, *30*(6), 841-845. doi:10.1002/acp.3261

Kramer, R. S. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision, 15*(4), 1-9. doi:10.1167/15.4.1

Lander, K., & Bruce, V. (2003). The role of motion in learning new faces. *Visual Cognition*, *10*(8), 897-912. doi:10.1080/13506280344000149

Lander, K., Bruce, V., & Bindemann, M. (2018). Use-inspired basic research on individual differences in face identification: Implications for criminal investigation and security. *Cognitive Research: Principles and Implications*, 3:26. doi:10.1186/s41235-018-0115-6

Lander, K., Bruce, V., & Hill, H. (2001). Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology, 15*(1), 101-116. doi:10.1002/1099-0720(200101/02)15:1<101::AID-ACP697>3.0.CO;2-7

Lander, K., & Butcher, N. (2015). Independence of face identity and expression processing: Exploring the role of motion. *Frontiers in Psychology*, 6:255. doi:10.3389/fpsyg.2015.00255/full

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition*, *27*(6), 974-985. doi:10.3758/BF03201228

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*(3), 429-442. doi:10.1080/13506280444000382

Lander, K., Chuang, L., & Wickham, L. (2006). Recognizing face identity from natural and morphed smiles. *Quarterly Journal of Experimental Psychology*, *59*(5), 801-808. doi:10.1080/17470210600576136

Lander, K., & Poyarekar, S. (2015). Famous face recognition, face matching, and extraversion. *The Quarterly Journal of Experimental Psychology, 68*(9), 1769-1776. doi:10.1080/17470218.2014.988737

Lee, M. D., Vast, R. L., & Butavicius, M. A. (2006). Face matching under time pressure and task demands. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society, Vancouver, Canada* (pp. 1675-1680). Retrieved from https://pdfs.semanticscholar.org/c6f0/53bc5dbdcd89cba842251feaa4bb8b91378b.pdf

Lee, W. J., Wilkinson, C., Memon, A., & Houston, K. A. (2009). Matching unfamiliar faces from poor quality Closed-Circuit Television (CCTV) footage: An evaluation of the effect of training on facial identification ability. *AXIS, 1*(1), 19-28. Retrieved from https://ojs.lifesci.dundee.ac.uk/index.php/Axis/article/viewArticle/21

Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers, 31*(4), 557-564. doi:10.3758/BF03200735

McCaffery, J. M., & Burton, A. M. (2016). Passport checks: interactions between matching faces and biographical details. *Applied Cognitive Psychology, 30*(6), 925-933. doi:10.1002/acp.3281

McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, 3:21. doi:10.1186/s41235-018-0112-9

Megreya, A. M., & Bindemann, M. (2009). Revisiting the processing of internal and external features of unfamiliar faces: the headscarf effect. *Perception, 38*(12), 1831-1848. doi:10.1068/p6385

Megreya, A. M., & Bindemann, M. (2013). Individual differences in personality and face identification. *Journal of Cognitive Psychology, 25*(1), 30-37. doi:10.1080/20445911.2012.739153

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS ONE, 13*(3), e0193455. doi:10.1037/journal.pone.0193455

Megreya, A. M., Bindemann, M., & Havard, C. (2011). Sex differences in unfamiliar face identification: Evidence from matching tasks. *Acta Psychologica, 137*(1), 83–89. doi:10.1016/j.actpsy.2011.03.003

Megreya, A. M., Bindemann, M., Havard, C., & Burton, A. M. (2012). Identity-lineup location influences target selection: Evidence from eye movements. *Journal of Police and Criminal Psychology, 27*(2), 167-178. doi:10.1007/s11896-011-9098-7

Megreya, A. M., & Burton, A. M. (2006a). Unfamiliar faces are not faces: evidence from a matching task. *Memory and Cognition, 34*(4), 865–876. doi:10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2006b). Recognising faces seen alone or with others: when two heads are worse than one. *Applied Cognitive Psychology, 20*(7), 957–972. doi:10.1002/acp.1243

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69*(7), 1175–1184. doi:10.3758/BF03193954

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. doi:10.1037/a0013464

Megreya, A. M., Memon, A., & Havard, C. (2012). The headscarf effect: Direct evidence from the eyewitness identification paradigm. *Applied Cognitive Psychology*, *26*(2), 308-315. doi:10.1002/acp.1826

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*(6), 700–706. doi:10.1002/acp.2965

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *The Quarterly Journal of Experimental Psychology, 64*(8), 1473-1483. doi:10.1080/17470218.2011.575228

Meissner, C. A., Susa, K. J., & Ross, A. B. (2013). Can I see your passport please? Perceptual discrimination of own- and other-race faces. *Visual Cognition, 21*(9-10), 1287-1305. doi:10.1080/13506285.2013.832451

Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception, 44*(11), 1332-1341. doi:10.1177/0301006615599902

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577-581. doi:10.1037/xhp0000049

Nachson, I., & Shechory, M. (2002). Effect of inversion on the recognition of external and internal facial features. *Acta Psychologica, 109*(3), 227-238. doi:10.1016/S0001-6918(01)00058-0

Nelson, R. J. (2013). Is virtual reality exposure therapy effective for service members and veterans experiencing combat-related PTSD? *Traumatology, 19*(3), 171-178. doi:10.1177/1534765612459891

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128*(1), 56-63. doi:10.1016/j.cognition.2013.03.006

Norell, K., Läthén, K. B., Bergström, P., Rice, A., Natu, V., & O'Toole, A. (2015). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences, 60*(2), 331-340. doi:10.1111/1556-4029.12660

Noyes, E., Hill, M. Q., & O'Toole, A. J. (2018). Face recognition ability does not predict person identification performance: using individual data in the interpretation of group results. *Cognitive Research: Principles and Implications*, 3:23. doi:10.1186/s41235-018-0117-4

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition, 165,* 97-104. doi:10.1016/j.cognition.2017.05.012

Noyes, E., & Jenkins, R. (2019). Deliberate disguise in face identification. *Journal of Experimental Psychology: Applied*, *25*(2), 280-290. doi:10.1037/xap0000213

O'Toole, A. J., Phillips, P. J., Weimer, S., Roark, D. A., Ayyad, J., Barwick, R., & Dunlop, J. (2010). Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach. *Vision Research*, *51*(1), 74-83. doi:10.1016/j.visres.2010.09.035

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. doi:10.1016/S1364-6613(02)01908-3

O'Toole, A. J., Vetter, T., & Blanz, V. (1999). Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: An application of three-dimensional morphing. *Vision Research*, *39*(18), 3145–3155. doi:10.1016/S0042-6989(99)00034-6

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research, 51*(19), 2145-2155. doi:10.1016/j.visres.2011.08.009

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., ... & Al-Janabi, S. (2017). Do people have insight into their face recognition abilities? *The Quarterly Journal of Experimental Psychology*, *70*(2), 218-233. doi:10.1080/17470218.2016.1161058

Pan, X., & Hamilton, A. F. de C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395-417. doi:10.1111/bjop.12290

Papesh, M. H. (2018). Photo ID verification remains challenging despite years of practice. *Cognitive Research: Principles and Implications*, 3:19. doi:10.1186/s41235-018-0110-y

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception & Psychophysics*, *76*(5), 1335–1349. doi:10.3758/s13414-014-0630-6

Papesh, M. H., Heisick, L. L., & Warner, K., A. (2018). The persistent low-prevalence effect in unfamiliar face-matching: The roles of feedback and criterion shifting. *Journal of Experimental Psychology: Applied*, *24*(3), 416-430. doi:10.1037/xap0000156

Parsons, T. D., & Rizzo, A. A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, *39*(3), 250-261. doi:10.1016/j.jbtep.2007.07.007

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8-13. doi:10.1016/j,jneumeth.2006.11.017

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., … O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171-6176. doi:10.1073/pnas.1721355115

Pike, G. E., Kemp, R. I., Towell, N. A., & Phillips, K. C. (1997). Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition*, *4*(4), 409-438. doi:10.1080/713756769

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, *110*(3), 461-479. doi:10.1111/bjop.12368

Ramon, M., Miellet, S., Dzieciol, A. M., Konrad, B. N., Dresler, M., & Caldara, R. (2016). Super-memorizers are not super-recognizers. *PLoS ONE*, *11*(3), e0150972. doi:10.1371/journal.pone.0150972

Rice, A., Phillips, P. J., Natu, V., An, X., & O'Toole, A. J. (2013). Unaware person recognition from the body when face identification fails. *Psychological Science*, *24*(11), 2235–2243. doi:10.1177/0956797613492986

Rice, A., Phillips, P. J., & O'Toole, A. (2013). The role of the face and body in unfamiliar person identification. *Applied Cognitive Psychology*, *27*(6), 761–768. doi:10.1002/acp.2969

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*(5), 897-905. doi:10.1080/17470218.2015.1136656

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition*, *141*, 161–169. doi:10.1016/j.cognition.2015.05.002

Robbins, R. A., & Coltheart, M. (2012). The effects of inversion and familiarity on face versus body cues to person recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1098-1104. doi:10.1037/a0028584

Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2015). Face averages enhance user recognition for smartphone security. *PLoS ONE*, *10*(3), e0119460. doi:10.1371/journal.pone.0119460

Robertson, D. J., Middleton, R., & Burton, A. M. (2015). From policing to passport control. The limitations of photo ID. *Keesing Journal of Documents and Identity*, *46*(February), 3–8. Retrieved from

https://www.researchgate.net/publication/305429407_From_policing_to_passport_control_The_limitations_of_photo_ID

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by Metropolitan Police super-recognisers. *PLoS ONE, 11*(2), e0150036. doi:10.1371/journal.pone.0150036

Rosenberg, R. S., Baughman, S. L., & Bailenson, J. N. (2013). Virtual superheroes: Using superpowers in virtual reality to encourage prosocial behavior. *PloS ONE, 8*(1), e55003. doi:10.1371/journal.pone.0055003

Royer, J., Blais, C., Barnabé-Lortie, V., Carré, M., Leclerc, J., & Fiset, D. (2016). Efficient visual information for unfamiliar face matching despite viewpoint variations: It's not in the eyes! *Vision Research, 123,* 33-40. doi:10.1016/j.visres.2016.04.004

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

Sauerland, M., Sagana, A., Siegmann, K., Heiligers, D., Merckelbach, H., & Jenkins, R. (2016). These two are different. Yes, they're the same: Choice blindness for facial identity. *Consciousness and Cognition, 40,* 93–104. doi:10.1016/j.concog.2016.01.003

Schouten, B., & Jacobs, B. (2009). Biometrics and their use in e-passports. *Image and Vision Computing, 27*(3), 305–312. doi:10.1016/j.imavis.2008.05.008

Segovia, K. Y., Bailenson, J. N., & Leonetti, C. (2012). Virtual human identification line-ups. In C. Wilkinson & C. Rynn (Eds.) *Craniofacial Identification* (pp.101-114). New York, NY: Cambridge University Press.

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. *Annals of Surgery,*

236(4), 458-464. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1422600/

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): a self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, *2*(6), 140343. doi:10.1098/rsos.140343

Shah, P., Sowden, S., Gaule, A., Catmur, C., & Bird, G. (2015). The 20 item prosopagnosia index (PI20): relationship with the Glasgow face-matching test. *Royal Society Open Science*, *2*(11), 150305. doi:10.1098/rsos.150305

Simhi, N., & Yovel, G. (2016). The contribution of the body and motion to whole person recognition. *Vision Research*, *122*, 12-20. doi:10.1016/j.visres.2016.02.003

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., ... & Sanchez-Vives, M. V. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PloS ONE*, 1(1), e39. doi:10.1371/journal.pone.0000039

Strathie, A., & McNeill, A. (2016). Facial wipes don't wash: Facial image comparison by video superimposition reduces the accuracy of face matching decisions. *Applied Cognitive Psychology*, *30*(4), 504–513. doi:10.1002/acp.3218

Strömwall, L. A., Granhag, P. A., & Hartwig, M. (2004). Practitioners' beliefs about deception. In P. A. Granhag & L. A. Strömwall (Eds.), *Deception detection in forensic contents* (pp. 229-250). Cambridge, UK: Cambridge University Press

Strömwall, L. A., & Granhag, P. A. (2003). How to detect deception? Arresting the beliefs of police officers, prosecutors and judges. *Psychology, Crime and Law*, *9*(1), 19-36. doi:10.1080/10683160308138

Susa, K. J., Michael, S. W., Dessenberger, S. J., & Meissner, C. A. (2019). Imposter identification in low prevalence environments. *Legal and Criminological Psychology*, *24*(1), 179-193. doi:10.1111/lcrp.12138

Thornton, I. M., & Kourtzi, Z. (2002). A matching advantage for dynamic human faces. *Perception*, *31*(1), 113-132. doi:10.1068/p3300

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PLoS ONE*, *14*(2), e0211037. doi:10.1371/journal.pone.0211037

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception*, *43*(2-3), 214-218. doi:10.1068/p7676

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*(1), 47-58. doi:10.1037/xap0000108

Trainer, P. (2017, March 21). How airport security has changed since 9/11, *Skyscanner*. Retrieved from https://www.skyscanner.com/tips-and-inspiration/guide-to-airport-security-since-9-11

United States Government Accountability Office (2010). *Efforts to Validate TSA's Passenger Screening Behavior Detection Program Underway, but Opportunities Exist to Strengthen Validation and Address Operational Challenges* (GAO Publication No. GAO-10-763). Retrieved from https://www.gao.gov/assets/310/304510.pdf

United States Government Accountability Office (2013). *TSA Should Limit Future Funding for Behavior Detection Activities* (GAO Publication No. GAO-14-159). Retrieved from https://www.gao.gov/assets/660/658923.pdf

Ventura, P., Livingston, L. A., & Shah, P. (2018). Adults have moderate-to-good insight into their face recognition ability: Further validation of the 20-item Prosopagnosia Index in a Portuguese sample. *Quarterly Journal of Experimental Psychology*, *71*(12), 2677-2679. doi:10.1177/1747021818765652

Vine, J. (2012). *Inspection of border control operations at Terminal 3, Heathrow Airport, August – November 2011*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/546255/Inspection-of-Border-Control-Operations-at-Terminal-3-Heathrow-Airport_2012.pdf

Vine, J. (2014). *An inspection of Border Force operations at Stansted Airport, May – August 2013*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/546959/An-Inspection-of-Border-Force-Operations-at-Stansted-Airport_Jan_2014.pdf

Vrij, A. (2008a). *Detecting lies and deceit: Pitfalls and opportunities*. (2nd ed.). Chichester, UK: Wiley

Vrij, A. (2008b). Nonverbal dominance versus verbal accuracy in lie detection: A plea to change police practice. *Criminal Justice and Behavior*, *35*(10), 1323-1336. doi:10.1177/0093854808321530

Vrij, A., Akehurst, L., & Knight, S. (2006). Police officers', social workers', teachers' and the general public's beliefs about deception in children, adolescents and adults. *Legal and Criminological Psychology*, *11*(2), 297-312. doi:10.1348/135532505X60816

Vrij, A. Granhag, P. A., & Porter, S. (2010). Pitfall and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, *11*(3), 89-121. doi:10.1177/1529100610390861

Vrij, A., & Mann, S. (2001). Telling and detecting lies in a high-stake situation: The case of a convicted murderer. *Applied Cognitive Psychology*, *15*(2), 187-203. doi:10.1002/1099-0720(200103/04)15:2<187::AID-ACP696>3.0.CO;2-A

Wang, Y., Thomas, J., Weissgerber, S. C., Kazemini, S., Ul-Haq, I., Quadflieg, S. (2015). The headscarf effect revisited: Further evidence for a culture-based internal face processing advantage. *Perception*, *44*(3), 328-336. doi.10.1068/p7940

Want, S. C., Pascalis, O., Coleman, M., & Blades, M. (2003). Recognizing people from the inner or outer parts of their faces: Developmental data concerning 'unfamiliar' faces. *British Journal of Developmental Psychology*, *21*(1), 125-135. doi: 10.1348/026151003321164663

White, D., Burton, A. L., & Kemp, R. I. (2016). Not looking yourself: The cost of self-selecting photographs for identity verification. *British Journal of Psychology*, *107*(2), 359–373. doi:10.1111/bjop.12141

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied*, *20*(2), 166-173. doi:10.1037/xap0000009

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology*, *27*(6), 769-777. doi:10.1002/acp.2971

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE*, *10*(10), e0139827. doi:10.1371/journal.pone.0139827

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, *21*(1), 100-106. doi:10.3758/s13423-013-0475-3

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

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1814), 20151292. doi:10.1098/rspb.2015.1292

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017). Face matching impairment in developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology*, *70*(2), 287–297. doi:10.1080/17470218.2016.1173076

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

Wilson, C. J., & Soranzo, A. (2015). The use of virtual reality in psychology: A case study in visual perception. *Computational and Mathematical Methods in Medicine*, 2015: 151702. doi:10.1155/2015/151702

Wirth, B. E., & Carbon, C.-C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied*, *23*(2), 138–157. doi:10.1037/xap0000114

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14*(6), 737-764. doi:10.1068/p140737

Yovel, G., & O'Toole, A. J. (2016). Recognising people in motion. *Trends in Cognitive Sciences*, *20*(5), 385-395. doi:10.1016/j.tics.2016.02.005

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., ... & Liu, J. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology*, *20*(2), 137-142. doi:10.1016/j.cub.2009.11.067

Appendix

A summary of the statistical comparisons of the VRPC task accuracy data, both as percentage accuracy and following arcsine square-root transformation, for all experiments in Chapter 4.

		Percentage accuracy		Arcsine square-root	
Experiment 9	Trial type	$F = 43.21$	$p < .001$	$F = 43.23$	$p < .001$
ANCOVA	Activity level	$F = 0.34$	$p = .72$	$F = 0.34$	$p = .72$
	Sensitivity	$F = 1.45$	$p = .24$	$F = 1.45$	$p = .24$
	Trial type x Activity level	$F = 1.01$	$p = .37$	$F = 1.01$	$p = .37$
	Trial type x Sensitivity	$F = 0.51$	$p = .48$	$F = 0.51$	$p = .48$
	Activity level x Sensitivity	$F = 0.79$	$p = .46$	$F = 0.79$	$p = .46$
	Trial type x Activity level x Sensitivity	$F = 0.97$	$p = .39$	$F = 0.97$	$p = .39$
Experiment 10	Trial type	$F = 26.53$	$p < .001$	$F = 26.54$	$p < .001$
ANCOVA	Activity level	$F = 0.02$	$p = .98$	$F = 0.02$	$p = .98$
	Sensitivity	$F = 1.02$	$p = .32$	$F = 1.02$	$p = .32$
	Trial type x Activity level	$F = 1.87$	$p = .16$	$F = 1.87$	$p = .16$
	Trial type x Sensitivity	$F = 0.15$	$p = .70$	$F = 0.15$	$p = .70$
	Activity level x Sensitivity	$F = 0.92$	$p = .40$	$F = 0.92$	$p = .40$
	Trial type x Activity level x Sensitivity	$F = 0.36$	$p = .70$	$F = 0.36$	$p = .70$
Experiment 11	Trial type	$F = 0.67$	$p = .42$	$F = 0.66$	$p = .42$
ANOVA	Activity level	$F = 0.08$	$p = .92$	$F = 0.08$	$p = .92$
	Trial type x Activity level	$F = 32.83$	$p < .001$	$F = 32.83$	$p < .001$
Experiment 12	Trial type	$F = 8.44$	$p = .007$	$F = 8.43$	$p = .007$
ANOVA	Activity level	$F = 2.19$	$p = .12$	$F = 2.18$	$p = .12$
	Trial type x Activity level	$F = 48.92$	$p < .001$	$F = 48.93$	$p < .001$
Experiment 13	Trial type	$F = 2.89$	$p = .10$	$F = 2.88$	$p = .10$
ANOVA	Activity level	$F = 0.07$	$p = .94$	$F = 0.07$	$p = .94$
	Trial type x Activity level	$F = 20.12$	$p < .001$	$F = 20.12$	$p < .001$
Experiment 14	Trial type	$F = 9.28$	$p = .003$	$F = 7.97$	$p = .006$
ANOVA	Activity level	$F = 3.24$	$p = .08$	$F = 2.02$	$p = .15$
	Trial type x Activity level	$F = 78.58$	$p < .001$	$F = 79.15$	$p < .001$