Time Pressure and Human-Computer Interaction in Face Matching

Matthew C. Fysh

School of Psychology

University of Kent

A thesis submitted for the degree of Ph.D. in the Faculty of Social Sciences at the University of Kent

September 2017

Abstract

Research has consistently demonstrated that the matching of unfamiliar faces is remarkably error-prone. This raises concerns surrounding the reliability of this task in operational settings, such as passport control, to verify a person's identity. A large proportion of the research investigating face matching has done so whilst employing highly optimised same-day face photographs. Conversely, such ideal conditions are unlikely to arise in realistic contexts, thus making it difficult to estimate accuracy in these settings from current research. To attempt to address this limitation, the experiments in this thesis aimed to explore performance in forensic face matching under a range of conditions that were intended to more closely approximate those at passport control. This was achieved by developing a new test of face matching – the Kent Face Matching Test (KFMT) – in which to-be-matched stimuli were photographed months apart (Chapter 2). The more challenging conditions provided by the KFMT were then utilised throughout the subsequent experiments reported, to investigate the impact of time pressure on task performance (Chapter 3), as well as the reliability of human-computer interaction at passport control (Chapter 4). The results of these experiments indicate that person identification at passport control is substantially more challenging than is currently estimated by studies that employ highly optimised face-pair stimuli. This was particularly evident on identity mismatch trials, for which accuracy deteriorated consistently within sessions, due to a match response bias that emerged over time (Chapters 2 & 3). These results are discussed within the context of passport control, and suggestions are provided for future research to further reveal why errors might arise in this task.

Acknowledgements

I would like to thank my supervisor Dr. Markus Bindemann for three years of superb supervision and friendship, and whose patience, positive attitude, and unwavering dedication to the pursuit of the truth inspired me to be a better researcher. I am immeasurably grateful to have worked with Markus, who made all of this possible in the first instance.

In addition, I would like to thank the administrative staff in the School of Psychology, as well as the technical support team, who assisted in the implementation of some of the paradigms that were used in this work. This research was supported by the Graduate Teaching Assistant Scholarship, University of Kent.

I would also like to offer a special thanks to my friends and family who supported me over the last three years. In particular, I am grateful to my fellow PhD students from the office, who kept me laughing, and who made these past three years highly memorable and enjoyable.

Finally, I dedicate this thesis to my parents, who supported and believed in me every step of the way.

Declaration

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

_____

Matthew C. Fysh

Publications

Within this thesis, Chapters 1 and 2 (Experiments 1 & 2) are in press, and Chapter 3 (Experiments 3 & 4) has been published. Chapter 4 (Experiments 5-7) is currently under review for publication.

Chapter 1

Fysh, M. C., & Bindemann, M. (in press). Forensic face matching: A review. In M. Bindemann & A. M. Megreya (eds.), Face Processing: Systems, Disorders, and Cultural Differences. New York, NY: Nova.

Chapter 2 (Experiments 1 & 2)

Fysh, M. C., & Bindemann, M. (in press). The Kent Face Matching Test. The British Journal of Psychology. doi: 10.1111/bjop.12260

Chapter 3 (Experiments 3 & 4)

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

Chapter 4 (Experiments 5-7)

Fysh, M. C., & Bindemann, M. (under review). Human-computer interaction in face matching. Cognitive Science.

Table of Contents

Chapter 1

Forensic Face Matching: A Review

1.1. Introduction

At airports and national borders, passport officers routinely compare travellers to their passport photographs. A key purpose of this task is to confirm that the person depicted in the identity document matches its bearer. Travellers may attempt to evade detection at this stage by using a fraudulent passport into which their photograph has been inserted. However, with the development of sophisticated passports, such counterfeit identity documents are increasingly difficult to forge. An alternative method to avoid detection is for travellers to use the stolen or borrowed passport of another person that is of similar appearance. These identity mismatches, or impostors, are now a documented security concern (NCA, 2015; Stevens, 2011), and are on the increase (Bundesdruckerei, 2013; National Audit Office, 2007).

In Psychology, this problem has been studied with forensic face matching tasks, in which observers compare two concurrent faces to decide whether they depict one person (an identity match) or different individuals (a mismatch) (see Johnston & Bindemann, 2013). The purpose of this research is to estimate face-matching accuracy in applied settings, given that detection rates in these contexts are unknown due to factors such as inadequately documented arrivals (NCA, 2015). This research has consistently shown that face matching is highly error-prone, and raises concern about reliance on this task for security purposes.

Figure 1.1. Example pairs from the Glasgow Face Matching Test (GFMT) (Burton et al., 2010). The top pair depicts an identity match, whilst the bottom pair is an example of an identity mismatch.

In laboratory experiments, observers often compare highly optimised pairs of faces, which are presented under neutral lighting and with a similar expression and pose (for an example, see Figure 1.1). Crucially, identity matches in this task are constructed from two images that have been taken on the same day, but using different cameras. Even when matching faces under such favourable conditions, up to 20% errors arise (Burton, White, & McNeill, 2010). This level of performance is already considered problematic for operational contexts (Jenkins & Burton, 2008a; Robertson, Middleton, & Burton, 2015), but further errors occur when the task more closely approximates realistic settings. For example, when faces are matched for a sustained duration, accuracy on mismatch trials deteriorates enormously, to around 50% (Alenezi & Bindemann, 2013; Alenezi, Bindemann, Fysh, & Johnston, 2015). This

raises the possibility that the detection of impostors at passport control, where face matching over long periods is the norm, is extremely vulnerable to human error.

Separate studies have also shown that impostor detection accuracy can be reduced to near-chance level with realistic photo-ID documents (Bindemann & Sandford, 2011; Kemp, Towell, & Pike, 1997). This problem arises from the variation in appearance that people exhibit naturally over time, through changes in hairstyle, age, weight, or facial adiposity, for example (Jenkins, White, van Montfort, & Burton, 2011; Megreya, Sandford, & Burton, 2013). Research has also shown that embedding a face within a passport frame alongside information such as name and date of birth is sufficient to reduce accuracy on mismatch trials by around 8%, and promotes a bias to erroneously classify pairs as identity matches (McCaffery & Burton, 2016). Together, these findings suggest that identity verification from photographic documents such as passports, which are typically valid through a ten-year period and require the validation of important biographical details, is a particularly difficult task.

While much of this evidence has been collected through experimentation with student participants, some studies have also investigated the performance of passport officers. In one such study, passport renewal officers, who routinely compare new passport photographs with an expired image, incorrectly accepted 14% of mismatching identity pairs in a person-to-photo comparison task (White, Kemp, Jenkins, Matheson, & Burton, 2014). In a further experiment, officers performed similarly to students, with an average error rate of 20% when matching pairs of optimised face photographs. More recent studies have also investigated the performance of facial review staff; who verify the eligibility of new passport applications, and facial examiners; who perform specialist comparisons in circumstances where a suspected fraudulent application is detected (White, Dunn,

Schmid, & Kemp, 2015). This research found that the facial review staff performed comparably to student participants, and made 52% errors in this task. Facial examiners were considerably less error-prone than the other two groups, but still made 31% errors when comparing target faces.

These experimental findings are corroborated by government reports that up to 61% of fraudulent passport applications in the UK are annually missed by the Passport Office (HM Passport Office, 2016). Furthermore, between 2015 and 2016 Border Force UK detected 1,013 travellers that carried Fraudulently Obtained Genuine (FOG) passports, which comprise stolen or borrowed identity documents (UK Parliament, 2016). These records do not take into consideration any of the factors that are already known to impact accuracy in this task. It is therefore likely that these numbers greatly underestimate the true scale of the problem.

1.2. Forensic Face Matching: A Data Problem

One explanation for errors in face matching is that these arise from data limits, whereby the identity information within to-be-compared face stimuli might be too limited to make a definitive decision. High-quality face images that are matched for illumination, pose, and expression should present the fewest data limits, and result in an average performance of 80-90% accuracy (Burton et al., 2010). Accuracy deteriorates by 8-10% following a change in viewpoint between to-be-compared faces (Bruce et al., 1999; Estudillo & Bindemann, 2014), and up to 23% through differences in image quality (Bindemann, Attard, Leach, & Johnston, 2013; Henderson, Bruce, & Burton, 2001). Such factors demonstrate that the amount of information that is available across both images in a face pair is closely related to performance in this task. Below, these data-limiting factors are reviewed separately.

1.2.1. Illumination

Under natural viewing conditions, faces can be illuminated from several different directions, such as from above, the side, or the front. Changes in lighting direction affect the information that is visible in a face. For example, illuminating a person from one side creates shading on the opposing side, obscuring a significant proportion of their face.

In an early face-matching study, Hill and Bruce (1996) showed that fewer matching errors occurred when lighting direction was consistent for both faces in a pair, and top-lit faces were matched more accurately than bottom-lit faces. Variation in lighting direction within pairs increased false match decisions (Experiment 2). These findings are of practical relevance to applied contexts, which typically require the comparison of an evenly-lit passport photograph to the image of its bearer under ambient lighting conditions. This variation represents a substantial source of noise, and contributes to the difficulty of person identification (Jenkins et al., 2011).

One solution to this problem is the implementation of face averages, which are merged representations of multiple face photographs. These averages portray elements of a person's face that remain stable over time, whilst eliminating sources of noise such as changes in lighting (see Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2011), and appear to enhance person identification in face matching (Robertson, Kramer, & Burton, 2015; White, Burton, Jenkins, & Kemp, 2014). Alternatively, single face images could be pre-processed to reduce the variation associated with incongruent lighting. This approach was explored in a recent study, in which observers made up to 27% identification errors when sequentially viewing disparately-lit faces. These errors were reduced by 11% by raising the luminance of

shaded areas relative to the rest of the face, whilst preserving the contours and depth of the original image (Liu, Chen, Han, & Shan, 2013).

1.2.2. Viewpoint

Changes in viewpoint impair the recognition of newly learned faces (see, e.g., Hill & Bruce, 1996; Longmore, Liu, & Young, 2008) and face matching (see, e.g., Bindemann et al., 2013; Bruce et al., 1999; Estudillo & Bindemann, 2014; Hill & Bruce, 1996). Hill and Bruce (1996) found, for example, that observers were 13% more accurate at matching faces that were both presented in profile or three-quarter views than when viewpoint differed. A change in view appears to specifically impair accuracy on mismatch trials, by 10-15% (Bindemann et al., 2013; Estudillo & Bindemann, 2014). However, changes in viewpoint did not interact with reductions in image quality (Bindemann et al., 2013) or external-feature masking (Estudillo & Bindemann, 2014). This indicates that view changes are not exacerbated by additional factors.

One explanation for the effect of view on face-matching accuracy could be that observers cannot refer to the same internal features (i.e., the eyes, nose, and mouth) across to-be-compared face targets. These features are fixated frequently during face-matching tasks (Bobak, Parris, Gregory, Bennetts, & Bate, 2017; Özbek & Bindemann, 2011), but the proportion of fixations that land on the features are affected greatly by changes in view (Bindemann, Scheepers, & Burton, 2009). On the other hand, identification across view is robust with familiar face targets (Hill & Bruce, 1996; Jenkins et al., 2011), and some observers are also highly accurate in matching unfamiliar faces across views (Estudillo & Bindemann, 2014). This indicates that

sufficient visual information can exist within faces to identify unfamiliar faces across views, but some observers are better at utilising this information than others.

1.2.3. Image Degradation

Passports only provide small face images, and the resolution of these images can be reduced further during passport printing and lamination compared to the source photograph. In addition, three-dimensional holograms are typically applied to passport photographs, which appear to move when the passport is tilted. These manipulations degrade the visual quality of a passport image and create a mismatch to the presentation of its bearer. Matching poor-quality footage to high-quality targets reduces the detection of mismatches by 11% and also increases the false rejection of identity matches by 10% (Bruce, Henderson, Newman, & Burton, 2001). In fact, in some studies where observers compare low-quality images of targets extracted from CCTV footage to high-quality counterparts, performance only marginally exceeds chance levels, but improves considerably, by roughly 26%, when to-be-compared face images are both of high-quality (Henderson et al., 2001).

More recent research suggests that poor image quality might specifically reduce accuracy on match trials, thus leading observers to classify more stimuli as "impostors". For example, accuracy on match trials was reduced by 16% when comparing a blurred image to a high-quality counterpart, whereas performance on mismatch trials was comparable to when viewing two high-quality images (see Experiments 1 & 3; Strathie & McNeill, 2016). Similarly, accuracy deteriorates from 90% when matching high-resolution stimuli, to below 50% when observers match a high-quality face to a heavily-pixelated low-resolution target, such as that depicted in Figure 1.2 (Experiment 1; Bindemann et al., 2013). These findings suggest that

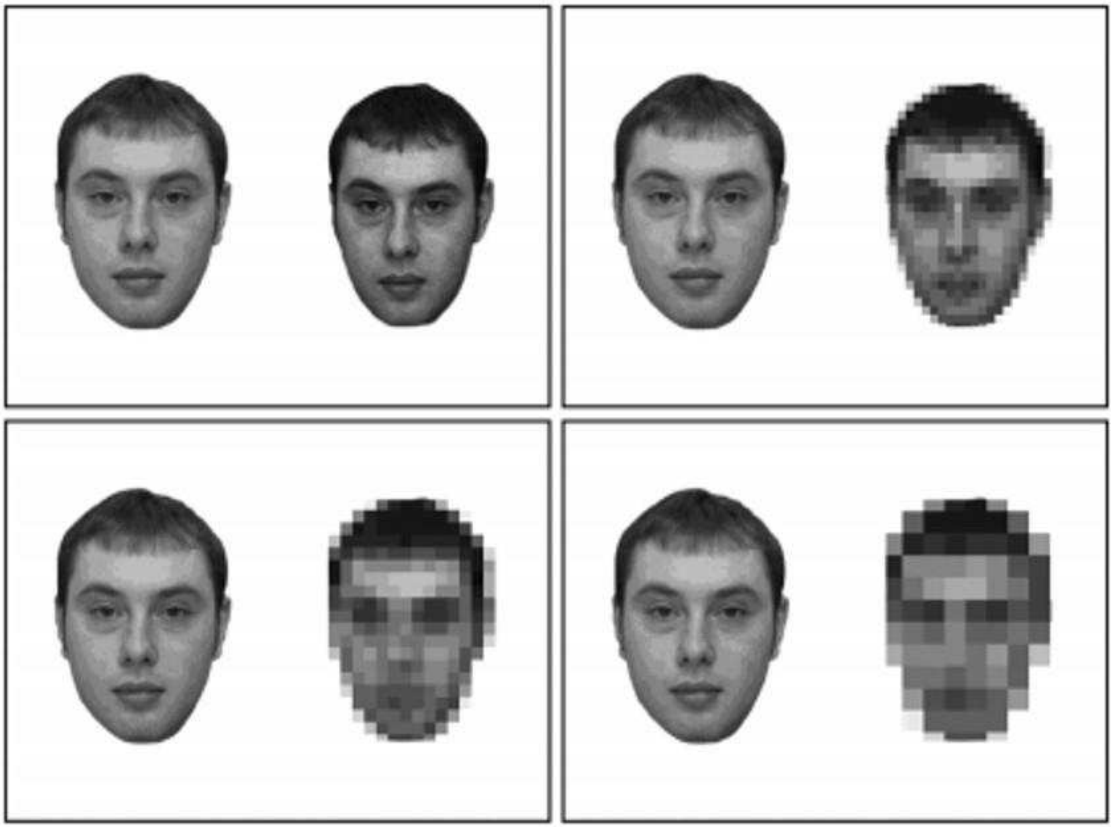

Figure 1.2. Examples of the pixelated stimuli used by Bindemann et al. (2013), in which both faces were presented in high-quality (top left), or one face was degraded to a resolution of 20 pixels (top right), 14 pixels (bottom left), or 8 pixels across (bottom right).

observers might adopt a bias to classify pairs of faces as depicting different identities when image quality is degraded. Perhaps counterintuitively, reducing the size of poor-quality faces can partially offset the detrimental effects of low image resolution (see Experiments 2 & 3; Bindemann et al., 2013), which points to potential solutions to this problem. Accurate identification from low-resolution images also remains possible for individuals with a high aptitude for matching faces (Robertson, Noyes, Dowsett, & Burton, 2016), or who are familiar with the target identities (Burton, Wilson, Cowan, & Bruce, 1999; Jenkins & Kerr, 2013). This indicates that degradation in image quality imposes data limits that reduce performance generally, but that the task can remain

Figure 1.3. Faces naturally vary over time, due to changes in, for example, age, hairstyle, and camera angle. A result of such variation is that no face casts the same image twice (see also, Jenkins & Burton, 2011).

solvable and high accuracy in some select individuals may remain relatively preserved.

1.2.4. Within-Target Variation

Faces undergo considerable variation over time (see Figure 1.3). This encompasses changes in hairstyle, weight, and facial paraphernalia such as glasses (Jenkins & Burton, 2011; Jenkins et al., 2011). A consequence of this is that the degree of similarity between a passport photograph and its owner decreases as the time interval between these lengthens. Studies have suggested that such variation increases the error-rate on identity match trials. For example, Megreya et al. (2013) found that

accuracy deteriorated from approximately 90% when comparing two images of the same person that were taken on the same day, to around 70% when the time interval between these images increased to several months (Experiment 2). These findings indicate that in operational contexts such as passport control, accuracy is compounded by the variation that arises between a person and their passport photograph over time.

Variation within targets also influences the extent to which a photograph represents a given identity. For example, Bindemann and Sandford (2011) found that when matching a target face to one of three different ID photographs, accuracy ranged from 46-67%. In addition, Jenkins et al. (2011) tasked observers with sorting 40 intermixed images of two unfamiliar identities into single-identity piles. The researchers found that observers most commonly detected nine identities out of the 40 images, and that none of the participants arrived at the correct solution. A further experiment also revealed that images meeting the requirements for passport photographs were rated as being less identity-representative than ambient images that did not meet these requirements (Experiment 3).

These findings reflect that the variability that arises within targets over time leads to a considerable number of errors in face matching. One solution to this problem is to present observers with multiple exemplars of a target from different contexts. An early demonstration of this was provided by Bindemann and Sandford (2011), who found that accuracy was at 60% when matching one of three ID photographs to a target, but improved to 85% when three ID photographs of the target were presented concurrently.

Later research exploring this strategy found that performance improved from approximately 80% when comparing many pairs of faces, to 90% when matching four concurrent images of a single identity to a probe image (White, Burton, et al., 2014;

Experiment 2). Further evidence suggests that accuracy improves incrementally as the number of single-identity exemplars increases, but that this improvement is identity-specific and does not generalise to novel targets (Dowsett, Sandford, & Burton, 2016). Together, these studies indicate that the problem of within-target variation in face matching can be partially mitigated by increasing the amount of target data available in these tasks.

1.2.5. Feature Masking

Unfamiliar face matching appears to be disproportionately dependent on external features, such as hairstyle and head outline (Bruce et al., 1999; Clutterbuck & Johnston, 2002; Estudillo & Bindemann, 2014; Henderson et al., 2001; Kemp, Caon, Howard, & Brook, 2016; Megreya & Bindemann, 2009). This is demonstrated in matching research where the external features of faces are obscured or removed. In an early study, for example, Bruce et al. (1999) removed the external features from one face image in a pair (see Figure 1.4, for an example). This manipulation reduced accuracy by 35%, whereas removing the internal features such as the eyes, nose, and mouth, reduced accuracy by only 11%. Reliance on external features in unfamiliar person identification was also demonstrated by Henderson et al. (2001), who found that accuracy was reduced from 64% when hair was visible, to 43% when targets' hair was covered. By contrast, removing the external features of difficult face-pair stimuli, such that observers may only extract identity-relevant information from internal features, can improve accuracy by 5% on difficult stimuli (Kemp et al., 2016). These findings suggest that observers rely heavily on external features when matching pairs of faces, but that such features may also be misleading.

This reliance on external features appears to vary across cultures. For example, in Middle-Eastern countries, headscarves are traditionally worn, which can obscure defining external features such as hair and head outline. As a consequence, facial identification in these settings relies to a greater extent on internal features than in some other countries, and facilitates an internal-feature advantage. This was shown by Megreya and Bindemann (2009), who found that Egyptian observers were more accurate at matching internal-feature faces than British observers, but also that British observers were more accurate when comparing external-feature faces (Experiment 4). However, this internal-feature advantage was absent in Egyptian children, indicating that viewing strategies in face processing continue to develop throughout adolescence (Experiment 5).

Figure 1.4. Example full-face (top), internal-feature (middle), and external-feature (bottom) stimuli. Images reproduced from Megreya and Bindemann (2009).

1.2.6. Live-to-Photo Matching

Person identification at passport control involves a comparison between a dynamic three-dimensional individual and a two-dimensional face photograph. This differs from many face-matching studies, which typically involve comparisons between image pairs. Some research has explored this discrepancy by comparing face photographs to video footage. For example, Bruce et al. (1999) found that accuracy improved from 68% when observers compared pairs of face images, to 79% when comparing a face image to video footage. However, in a later study this advantage was not replicated in student participants, but in a patient with prosopagnosia, whose accuracy improved from 31% when comparing static photographs, to 75% when one image was replaced with a video clip (see Experiment 3, Lander, Humphreys, & Bruce, 2004).

More recent research suggests that matching a live person to an image does not improve accuracy, but facilitates a response bias to classify pairs as identity matches (see Davis & Valentine, 2009; Megreya & Burton, 2008; see also, Kemp et al., 1997; White, Kemp, Jenkins, Matheson, et al., 2014). For example, 15% more errors occurred on mismatch trials when observers compared a live target, rather than a high-quality photograph, to one-week-old video footage (Experiment 3, Davis & Valentine, 2009). Further evidence for this bias is provided by Megreya and Burton (2008), who found that although overall accuracy was comparable between photo-to-photo and person-to-photo comparisons, mismatch errors increased by 7% in the latter condition, reflecting an increased tendency to identify pairs as matching when comparing a live individual to a photograph (Experiment 3).

1.3. Forensic Face Matching: A Resource Problem

The previous section reviewed factors that impose data limits on face matching. However, considerable evidence also suggests that performance in this task depends on resource limits, whereby errors occur because observers fail to correctly use the available information within stimuli. This is reflected in studies with minimal data limitations, where individual accuracy nonetheless ranges from 50-100% (see Bindemann, Avetisyan, & Rakow, 2012; Burton et al., 2010; Estudillo & Bindemann, 2014). Other studies show also that select observers are able to match faces with consistently high accuracy despite high data limitations (e.g., Robertson et al., 2016; White, Phillips, Hahn, Hill, & O'Toole, 2015). Below, factors that are relevant to resource limits in face matching are reviewed.

1.3.1. Individual Differences in Face Matching

Large performance differences arise between individuals in unfamiliar face matching (Bindemann, Avetisyan, et al., 2012; Bindemann, Brown, Koyas, & Russ, 2012; Burton et al., 2010; Estudillo & Bindemann, 2014; White, Kemp, Jenkins, & Matheson, et al., 2014). For example, when matching optimised face-pair stimuli with minimal data limitations, average performance is 80-90%, but individual accuracy ranges from 50-100% (Burton et al., 2010). Moreover, the accuracy of some individuals fluctuates by up to 20% when matching the same set of faces across consecutive days, whilst other individuals consistently achieve perfect performance (Bindemann, Avetisyan, et al., 2012). These findings reflect that sufficient data is portrayed within optimised face pairs, but that some observers fail to utilise this information effectively when making an identity judgement.

Such differences also emerge under more taxing conditions. For example, accuracy ranges from 55-100% when comparing optimised targets across different viewpoints (Estudillo & Bindemann, 2014), and from 25-100% when matching pixelated images of familiar faces (Robertson et al., 2016). Taken together, this research suggests that even under high data constraints, identification remains possible for some observers, suggesting that face-matching ability exists on a continuum.

Recent research has identified individuals who fall on the higher end of this continuum. For example, Bobak, Dowsett, and Bate (2016) found that a group of 'super-recognisers' (see Russell, Duchaine, & Nakayama, 2009) outperformed control subjects by 10% in an optimised matching task, and by 18% under more taxing conditions. In another study, super-recognisers scored 93% in an array matching task, outperforming one control group by 20% (Bobak, Hancock, & Bate, 2016).

Research has also investigated high-performing individuals who operate within applied settings. For example, Robertson et al. (2016) found that four metropolitan police super-recognisers scored 96% on an optimised face matching task, outperforming a control group of police trainees by 15%. Furthermore, the super-recognisers made only 7% errors when matching pixelated images of familiar faces, and outperformed student observers by 20%. In another study, specialist facial examiners outperformed groups of students and facial review staff by 21% (White, Dunn, et al., 2015), but still made 31% errors. Additionally, government-employed forensic experts were more accurate than student observers across constrained viewing conditions, and when stimuli were inverted (White, Phillips, et al., 2015). Overall, this research reflects that data limitations can be offset by individuals with high resource capacity for matching faces.

By contrast, low resource capacity for face matching reduces accuracy despite minimal data limitations. For example, individuals with developmental prosopagnosia perform consistently poorly in facial identification (see Dalrymple & Palermo, 2016; Duchaine & Nakayama, 2006) and face-matching tasks (White, Rivolta, Burton, Al-Janabi, & Palermo, 2017). These differences appear to be exacerbated further under more taxing conditions. This is reflected by increased errors on match trials (White et al., 2017), indicating a specific impairment for determining that two faces depict the same identity.

These differences between individuals are reflected in viewing strategies, with super-recognisers spending longer than developmental prosopagnosics and control subjects when fixating internal regions such as the eyes and nose in free-viewing tasks (Bobak et al., 2017). However, performance is also impacted by differences between individuals such as race (Megreya, White, & Burton, 2011; Meissner, Susa, & Ross, 2013), age (Megreya & Bindemann, 2015), emotional state (Attwood, Penton-Voak, Burton, & Munafò, 2013), sex, (Megreya, Bindemann, & Havard, 2011) and personality (Megreya & Bindemann, 2013). These factors are reviewed separately below.

Face matching and race. The Own-Race Bias or Cross-Race Effect has been widely demonstrated in face recognition research, whereby person identification is more reliable when a target belongs to one's own race (see, e.g., Meissner & Brigham, 2001; Chiroro, Tredoux, Radaelli, & Meissner, 2008). This effect has also been demonstrated in face matching. For example, Megreya, White, et al. (2011) found that British observers made 14% more errors when matching other-race faces than when matching own-race faces. Similarly, Egyptian observers made more errors when matching British faces than Egyptian faces. This converges with a later study, where

face-matching accuracy deteriorated from 83% when viewing own-race faces, to 72% when faces belonged to a different race (Meissner et al., 2013). The researchers also found that errors increased further when to-be-matched faces were photographed many months apart (Experiment 2), and when targets were partially disguised with a cap and sunglasses (Experiment 3).

Face matching and age. Research has also suggested that age influences face identification ability (see, e.g., Dolzycka, Herzmann, Sommer, & Wilhelm, 2014). In one study, Megreya and Bindemann (2009) found that Egyptian adults were more reliant on the internal features of faces compared to the external features. However, this dependency was reversed in Egyptian children, indicating that viewing strategies in face matching continue to develop past adolescence. More recent research indicates that face-matching performance improves into adulthood, but deteriorates thereafter. For example, Megreya and Bindemann (2015) found that 19-year-olds were 12% more accurate than 65-year-olds on match trials, and also outperformed 7-year-olds by 10% (Experiment 2), reflecting the continued development of face processing ability throughout adolescence, as well as an age-related deterioration in this task.

Face matching and personality. Person identification may also be influenced by individual differences in personality traits, such as extraversion (Lander & Poyarekar, 2015) or empathy (Bate, Parris, Haslam, & Kay, 2010). In one study, for example, female observers who scored lower in emotional stability made a greater number of mismatch errors (Megreya & Bindemann, 2013). Conversely, another study found that induced anxiety resulted in 6% more errors on identity match trials (Attwood et al., 2013). Together, these findings reflect that face-matching accuracy is negatively related with anxiety, and converge with studies showing that anxiety also

reduces face recognition accuracy (Deffenbacher, Bornstein, Penrod, & McGorty, 2004).

Face matching and sex. Recent studies have suggested that sex differences between observers are also influential in face matching. For example, Megreya and Bindemann (2013) found a negative relationship between anxiety and face-matching accuracy, but only for female observers. Other studies show that female observers are 4% more accurate than male observers when matching female faces (Megreya, Bindemann, et al., 2011). Conversely, male observers performed comparably when matching female and male face pairs. These findings suggest that an own-gender bias exists in female, but not male observers when matching faces. However, these differences are relatively small, when considered within the context of other individual differences such as cross-race effects or age differences, which have been found to account for up to 11% and 18% increases in error rates, respectively (see Megreya & Bindemann, 2015; Meissner et al., 2013).

1.3.2. Mismatch Frequency

A key objective for passport officers is the detection of impostors (Stevens, 2011). These identity mismatches are rare in operational settings (UK Parliament, 2016; HM Passport Office, 2016). In laboratory settings featuring an equal number of match and mismatch trials, observers typically mistake 20% of mismatch pairs for identity matches (see Burton et al., 2010). Perhaps counterintuitively, some research indicates that low mismatch prevalence does not reduce face-matching accuracy. For example, when a single mismatch was featured among 50 trials, accuracy was 93%, but decreased to around 88% when match and mismatch trials occurred equally often (Bindemann et al., 2010). Reversing the order of these conditions resulted in

comparable accuracy between equal and low mismatch prevalence, with 96% and 92% accuracy, respectively (Experiment 2).

By contrast, more recent studies indicate that mismatch errors increase when these trials are infrequent. For example, Moore and Johnston (2013) found that performance deteriorated from 83% under equal mismatch prevalence to 57% when mismatches were rare. More recently, Papesh and Goldinger (2014) investigated low mismatch prevalence using a more challenging range of stimuli. In this study, accuracy on mismatch trials deteriorated from 77% with equal match and mismatch frequency, to 48% when mismatches were rare.

These findings indicate that under more challenging conditions, detection of infrequent mismatches is more error-prone. However, due to some important differences between studies, it is difficult to determine the true extent to which low mismatch prevalence impacts face matching accuracy. For example, Papesh and Goldinger (2014) did not inform observers that mismatches would be occurring infrequently, but applied feedback for errors throughout the task. Conversely, Bindemann et al. (2010) informed observers that mismatches would be rare, but also employed a considerably easier range of stimuli. These differences require further exploration to fully understand face-matching performance under low mismatch prevalence.

1.3.3. Response Bias

Studies have shown that observers sometimes develop a response bias in face matching. Alenezi and Bindemann (2013) found, for example, that in an extended face-matching task, observers developed a response bias to erroneously classify pairs as identity matches (see Figure 1.5). This resulted in a deterioration of 31% for

mismatch identification accuracy over 1000 successive trials. This bias appears to be alleviated when feedback is provided on a trial-by-trial basis (Alenezi & Bindemann, 2013), but is impervious to regular rest breaks and changes in environment (Alenezi et al., 2015), and appears to be compounded by time pressure (Bindemann, Fysh,

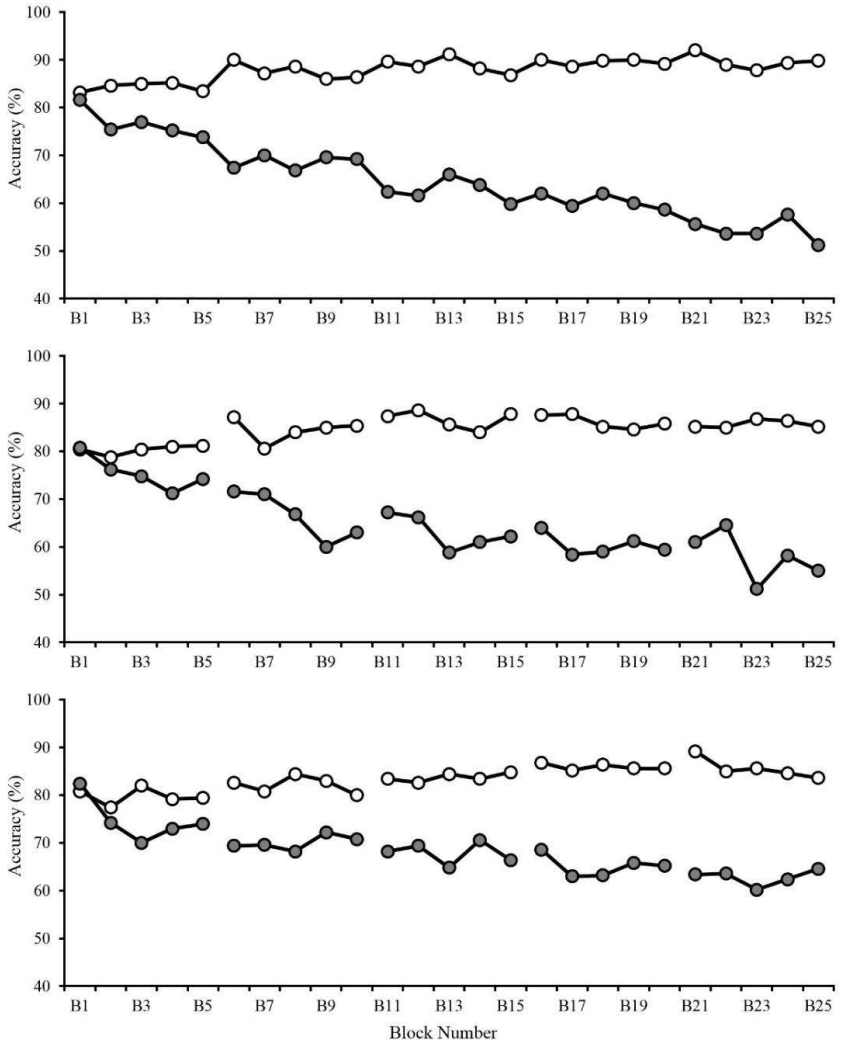

Figure 1.5. When matching faces for a prolonged duration, a response bias emerges that results in a profound deterioration in accuracy on mismatch trials. Open markers denote match trials, grey markers denote mismatch trials. The top graph was reproduced from Experiment 6 of Alenezi & Bindemann (2013), and the middle and bottom graphs were reproduced from Experiments 1 and 2 of Alenezi et al. (2015), respectively.

Cross, & Watts, 2016). Other research has indicated that a match bias also arises when matching live individuals to face photographs (see Davis & Valentine, 2009; Megreya & Burton, 2008), as well as when a target face is viewed in the context of a passport frame (McCaffery & Burton, 2016), and when mismatches are rare (Papesh & Goldinger, 2014). In addition, intranasal inhalations of oxytocin trigger a match bias in array-matching tasks, but do not facilitate higher accuracy (Bate et al., 2015).

By contrast, Moore and Johnston (2013) found that a bias to classify faces as identity mismatches emerged when observers were motivated to perform above average (Moore & Johnston, 2013). In addition, Strathie and McNeill (2016) observed a mismatch response bias when participants viewed poor-quality images (see also, Bindemann et al., 2013), whilst other research indicates that a mismatch bias arises when viewing durations of faces are highly constrained (Özbek & Bindemann, 2011).

1.3.4. Time Pressure

In laboratory settings, face-matching tasks are typically completed without time constraints, to measure best-possible accuracy. By contrast, passport officers must process queues of travellers within strict time targets that are frequently breached (Home Affairs Committee, 2012; ICI, 2014, 2015; Toynbee, 2016), suggesting that time pressure is regularly experienced in these settings. A recent study showed that under increasing time pressure, mismatch accuracy deteriorated from 89% when ten seconds were available per trial, to 79% when time pressure was two seconds (Bindemann et al., 2016). In addition, performance improved when time pressure receded, suggesting that this factor may contribute to errors in operational contexts.

Accuracy is also reduced when the viewing duration of stimuli is restricted. Research suggests that a minimum presentation time of 1-2 seconds is sufficient to

process and match some face images (O'Toole et al., 2007; Özbek & Bindemann, 2011), but that performance deteriorates under shorter durations, from 87% when viewing times are unlimited, to 80% when pairs are presented for only a second (Bindemann et al., 2010; Özbek & Bindemann, 2011).

Later research indicates that under more taxing conditions, observers require longer than two seconds to process stimuli (e.g., O'Toole et al., 2012; White, Phillips, et al., 2015). For example, government-employed forensic experts were more accurate than student control subjects when viewing difficult face pairs for two seconds, but also outperformed these observers by a greater margin when stimuli could be viewed for up to 30 seconds (White, Phillips, et al., 2015). These findings suggest that the allocation of cognitive resources in face matching can exceed a two-second threshold when task conditions are more challenging.

1.4. Forensic Face Matching: Solutions

The previous sections identified factors that constrain the amount of data in to-be-matched faces, as well as that limit resource capacity for performing this task. Understanding these limitations has been useful in studying some potential solutions to the problem of face matching. For example, data limitations posed by a single target face can be partially offset by the provision of multiple target exemplars (Bindemann & Sandford, 2011; White, Burton, Jenkins, & Kemp, 2014). In addition, aggregating the responses of multiple observers on each trial also improves performance, and may present a solution to overcoming resource limitations in face matching (Dowsett & Burton, 2015; White, Burton, Kemp, & Jenkins, 2013; White, Phillips, et al., 2015). In this section, factors that have been shown to improve performance in face matching are reviewed.

1.4.1. Feedback

Some studies have shown that the administration of feedback in face-matching tasks can benefit performance. For example, Alenezi and Bindemann (2013) found that feedback did not improve overall face-matching performance per se, but rather, prevented the onset of a match response bias. Thus, the administration of feedback in this study arrested the decline in mismatch accuracy that occurred when feedback was not provided (Experiment 1). Moreover, this positive effect of feedback was also observed even under additional task demands, such as across changes in view (Experiment 2), and when external facial features were occluded (Experiment 3). Perhaps importantly, the researchers found that feedback only benefitted performance when this was provided on a trial-by-trial basis, rather than cumulatively at the end of each block of trials (Experiment 5), indicating that observers utilised this information to adjust their criteria when matching unfamiliar faces. Further, these experimental findings reflect that permitting observers to monitor their own performance across face-matching tasks might help to reduce identification errors.

More recent research has shown that feedback can directly improve face-matching performance. For example, White, Kemp, Jenkins, and Burton (2014) found that accuracy in an optimised matching task improved from 82% to 92% when feedback was administered after every trial. In addition, the researchers found that for low-aptitude observers, these performance benefits were sustained from an optimised face-matching task to a more difficult task in which stimuli portrayed high within-person variation (Experiment 2). However, these gains were not observed in high-aptitude observers. This finding raises the possibility that observers who possess low

resource capacity for comparing faces, the administration of feedback might present a useful training paradigm.

1.4.2. Task Motivation

Research has suggested that providing observers with a performance incentive might improve face-matching accuracy. This was investigated by Moore and Johnston (2013), who found that observers who were incentivised with food-based rewards to perform above "average accuracy" were 9% more accurate than non-motivated controls. The researchers also found in an additional experiment that when the number of mismatches was reduced to two out of 32 trials, motivated observers were 29% more accurate on mismatch trials than non-motivated subjects (Experiment 2).

In a later study, Bobak, Dowsett, et al. (2016) found that under optimised conditions, monetary incentives did not promote accuracy in one group compared to non-motivated controls or super-recognisers. Conversely, in a more difficult test, incentivised observers were numerically, but not significantly, more accurate, and outperformed non-motivated controls by 5%. However, these observers made 13% more errors than super-recognisers. These findings reflect that task-based motivation might be of only limited benefit to face matching, and cannot supplant other factors that contribute to the high resource capacity possessed by super-recognisers.

Considered together with other work (e.g., Moore & Johnston, 2013), as well as research investigating the accuracy of passport officers, who possess a clear incentive to outperform student controls (see White, Kemp, Jenkins, Matheson, et al., 2014), but instead make a comparable number of errors, the full effects of motivation on face-matching performance are currently unclear. Additional studies should further

Figure 1.6. Stimuli used by Bindemann and Sandford (2011). Identification rates for ID1, ID2, and ID3 were 67%, 46%, and 58%, respectively. Viewing all three images concurrently improved performance to 85%.

investigate the extent to which rewards can incentivise observers to attain higher accuracy in this task.

1.4.3. Multiple Exemplars

Viewing faces belonging to the same person facilitates face learning (Ritchie & Burton, 2017) and improves performance in face-matching tasks (see Bindemann & Sandford, 2011; Dowsett et al., 2016; White, Burton, et al., 2014). This was shown in one study by Bindemann and Sandford (2011), who found that accuracy ranged from 46% to 67% when observers matched one of three ID photographs images to a face lineup. However, allowing observers to view all three photographs simultaneously improved performance to 85%, suggesting that observers utilised the additional data to reach a more accurate identification (see Figure 1.6).

This finding has been extended in subsequent studies. For example, White, Burton, et al. (2014) found that that performance improved between matching pairs of faces, and matching one face to an array of two concurrent photographs depicting a single identity (Experiment 2). However, this benefit did not continue with image arrays containing three or four photographs, reflecting that observers might only be

able to extract limited information from these multiple images. In a further experiment, the researchers found that on unfamiliar match trials only, comparing a target face to a face created by averaging together 12 identity photographs was less accurate than when comparing a target face to an array of four images. Performance on mismatch trials was comparable between these conditions (Experiment 3). These findings suggest that four face photographs yield similar identity data to two face photographs, but more identity data than an average face image that consists of 12 aggregated photographs.

More recently, Dowsett et al. (2016) provided observers with up to six face photographs of a single identity, who were then required to sort through a deck of 30 face images to locate the corresponding identity. Accuracy in this task improved in conjunction with the number of target images that were provided, from chance level when observers could refer to only one face-photograph when sorting through the deck, to around 70% when using six concurrent photographs. Contrary to the results of White, Burton, et al. (2014), these findings reflect that matching performance can benefit from additional photographs. However, these discrepancies might be explained by the extent to which concurrent face photographs of a single identity vary in relation to one another. For example, Ritchie and Burton (2017) found that viewing identity arrays that portrayed targets across highly-variable ambient conditions promoted accuracy on match trials, compared to when the arrays depicted the target across similar conditions (Experiment 2). However, mismatch accuracy was similar between high- and low-variability arrays.

These findings converge with an additional study, in which observers viewed image pairs of a single identity, and were then required to classify a subsequent image as belonging to the same person or as a different identity (Menon, White, & Kemp,

2015a). The researchers found that viewing high-variability image pairs improved performance over low-variability pairs, but also facilitated a match response bias, suggesting that too much variability can reduce observers' tolerance for between-identity variability. Together, these studies reflect that identification accuracy can be enhanced through viewing multiple target exemplars, but that this improvement is contingent on the extent to which images vary, and that too much variability can reduce accuracy on mismatch trials.

1.4.4. Face Averaging

An alternative to providing observers with multiple images of the same target is to aggregate photographs of a person together to form an average face image (see, Burton et al., 2005; Jenkins & Burton, 2011; White, Burton, et al., 2014). These images represent the stable aspects of a person's face over time, whilst disregarding extraneous sources of variance such as changes in illumination and expression (Jenkins & Burton, 2011). Such averages have been shown to aid identity recognition compared to when observers view only a face photograph from a single instance (Burton et al., 2005).

More recently, research has shown also that face averages improve face matching, when these are compared with a photograph that depicts a single photographic instance of an identity match or mismatch. In one study, for example, matching average face images to non-average face images was more reliable than when observers compared two non-average images (White, Burton, et al., 2014). However, this advantage was found only on match trials, and not on mismatch trials (Experiment 1). Moreover, this was less advantageous than when observers viewed four concurrent face photographs, suggesting that the quantity of data portrayed by

image averages might not exceed that which is portrayed across concurrent target images.

1.4.5. Response Aggregation

An alternative solution to improving performance in face matching is the aggregation of multiple responses across observers (see, e.g., White et al., 2013; White, Phillips, et al., 2015). For example, White et al. (2013) found that aggregating the responses of four observers on a trial-by-trial basis resulted in superior performance to individuals and pairs of observers. Moreover, groups of just 16 observers achieved near-perfect performance on match and mismatch trials. This converges with other work showing that pairs of observers outperform individuals, and that the benefits of working with a high-aptitude observer are sustained in a subsequent matching task that is performed alone (Dowsett & Burton, 2015).

Recently, research has also shown that near-perfect accuracy for difficult face-pair stimuli can be achieved by aggregating the responses of forensic examiners, who demonstrate superior performance to students and control subjects (White, Phillips, et al., 2015). Moreover, this research found that the accuracy of one forensic examiner was equivalent to four student observers. These findings reflect that, instead of reducing target data limitations through multiple photographs, aggregating the responses of multiple individuals might overcome the resource limitations that are present in solo observers.

1.5. Conclusion

This chapter outlined some of the key factors that impact human performance in forensic face matching tasks, with a specific focus on operational contexts such as

passport control. The current literature demonstrates that performance deteriorates generally under high data limitations, such as when image quality is poor (Bindemann et al., 2013; Henderson et al., 2001), but improves when additional data are available, such as when multiple exemplars are provided (see Bindemann & Sandford, 2011; White, Burton, et al., 2014). High data limitations can also be offset by high cognitive resources, within observers, for face matching (see Estudillo & Bindemann, 2014; Robertson et al., 2016; White, Phillips, et al., 2015). Conversely, low data limitations do not appear to offset low cognitive resources for this task (see Burton et al., 2010; White et al., 2017). In addition, the depletion of face-matching resources due to factors such as increased cognitive load (see McCaffery & Burton, 2016) or prolonged task duration (see Alenezi & Bindemann, 2013; Alenezi et al., 2015), is catastrophic for accuracy.

One encouraging observation from the available literature is that the face matching problem appears to be solvable, even though currently, a definitive solution remains unclear. Some factors already reduce errors considerably in this task, such as the administration of feedback on a trial-by-trial basis (see, Alenezi & Bindemann, 2013; White, Burton, et al., 2014), as well as the provision of multiple images of the same person (Bindemann & Sandford, 2011; White, Burton, et al., 2014). Moreover, this solution appears to be driven by individual differences in the ability to perform this task, which can overcome considerable data limitations in stimuli (see, e.g., Robertson et al., 2016; White, Phillips, et al., 2015). To further understand the cognitive factors that underpin face matching in realistic settings, it is important to investigate these individual differences increasingly in the context of practically relevant factors, such as time pressure, and within-target variation. This approach will serve to provide further information about the mechanisms driving performance in

face matching, which will subsequently reveal strategies for minimising errors in practical settings.

1.6. Structure of this Thesis

The purpose of this thesis is to investigate the reliability of forensic face matching using methods that more closely approximate applied settings such as passport control. The first experimental chapter describes the development of a new resource for studying face matching; the "Kent Face Matching Test" (KFMT), and provides normative data for this test. To establish the utility of this test, performance in the short version of the KFMT is compared against the Glasgow Face Matching Test (GFMT; Burton et al., 2010), which is an already established resource in face-matching research (Experiment 1). Additionally, performance is also explored in a longer version of the KFMT, to understand how accuracy varies in this task over an extended timeframe, as well as in relation to the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006), and the Cambridge Face Perception Test (CFPT; Duchaine, Germine, & Nakayama, 2007), which provide a benchmark measure of face memory and face perception, respectively (Experiment 2).

Chapter 3 explores the impact of time pressure on face-matching accuracy. This is achieved via a paradigm that flexibly administers time pressure across blocks of trials, and which permits observers to allocate their decision time on a trial-by-trial basis, provided that an entire block is completed within the required timeframe. Observers completed 200 face matching trials, under time pressure that ranged from ten to two seconds (Experiment 3), or eight to two seconds (Experiment 4). Both experiments found that matching accuracy was reduced by time pressure, but also by time passage, whereby a match response bias emerged over the task regardless of

whether time pressure was increasing or receding (see also, Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann, Fysh, et al., 2016).

The final empirical chapter explores human-computer interaction in face matching. Observers matched pairs of faces which were labelled onscreen as "same", "different", or "unresolved". The majority of these labels provided consistent trial information, however a small number were also inconsistent, in that they provided the incorrect identity judgement. With this information, observers were instructed to provide the final identification decision for each trial. Performance was severely reduced by inconsistent trial labels (Experiments 5 & 6). Moreover, responses on inconsistently-labelled mismatch trials were influenced to a greater extent by the trial labels than the facial information in stimuli, when given a compelling reason to trust these labels (Experiment 7). These findings are discussed in the final chapter, and suggestions for additional studies are provided that might further estimate face-matching performance in applied settings.

Chapter 2

The Kent Face Matching Test

Introduction

The previous chapter provided an overview of face-matching research, and identified a range of factors that impact performance in this task, such as time pressure (Bindemann, Fysh, et al., 2016), time passage (Alenezi & Bindemann, 2013; Alenezi et al., 2015), and changes in viewpoint (Estudillo & Bindemann, 2014). A key resource for this research has been the Glasgow Face Matching Test (GFMT; Burton et al., 2010). In this test, observers match high-quality, frontal-oriented pairs of faces that are evenly lit and bear neutral expressions. Crucially, identity matches also comprise same-day photographs taken only minutes apart, but with different image-capture devices, to provide optimised conditions to measure best-possible accuracy. Despite such favourable conditions, observers typically record 10-20% errors in this task. This level of performance is already considered problematic for operational settings (Jenkins & Burton, 2008a; Robertson, Middleton, et al., 2015), but shows also that observers find this task challenging even under ideal conditions.

The GFMT has already featured in over 30 face-matching studies to investigate how performance is impacted by factors such as time pressure (Bindemann, Fysh, et al., 2016), mismatch prevalence (Bindemann et al., 2010), sleep deprivation (Beattie, Walsh, McLaren, Biello, & White, 2016), image quality (Bindemann et al., 2013; Strathie & McNeill, 2016), and performance-related feedback (Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014). Moreover, this task has not only been administered to students, but also to non-students (Bobak, Dowsett, et al., 2016;

White, et al., 2017), forensic experts (White, Phillips, et al., 2015), police officers (Davis, Lander, Evans, & Jansari, 2016; Robertson et al., 2016), and passport officers (White, Kemp, Jenkins, Matheson, et al., 2014).

Despite its clear value for psychological research on person identification, the optimised conditions of the GFMT limit the utility of this test under some conditions. For example, one recent study found that working in pairs improved face-matching accuracy for low-performing, but not high-performing observers when comparing faces from the GFMT. However, when a more challenging stimulus set was employed, an advantage for working in pairs also emerged in high-performing observers (Dowsett & Burton, 2015). These findings suggest that the optimised conditions provided by the GFMT might obscure some effects that are better identified under more challenging conditions.

This chapter introduces the Kent Face Matching Test (KFMT), which aims to provide such conditions by encapsulating a more applied aspect of face matching. It is currently understood, for example, that face matching is more difficult when to-be-matched stimuli are taken months apart (see, e.g., Megreya et al., 2013) or comprise realistic photo-ID images (see, e.g., Bindemann & Sandford, 2011; Kemp et al., 1997; McCaffery & Burton, 2016). The stimuli of the KFMT are based on such findings, to characterise this dimension of face matching in operational settings, and thus provide a more ecologically valid measure of performance in this task. The KFMT is intended as a complementary resource to be used alongside more optimised measures such as the GFMT, which comprise same-day photographs of identity match pairs, and therefore estimate accuracy as a best-case scenario.

The construction of the KFMT is first described, after which data are provided to compare performance with established tests of face processing. In Experiment 1,

performance between the short versions of the KFMT and GFMT was compared, to demonstrate the greater difficulty of our test. This is followed by a second experiment, in which observers completed a longer version of the KFMT, along with two different established tests of face processing; the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) and the Cambridge Face Perception Test (CFPT; Duchaine et al., 2007).

Test Construction

To assemble the Kent University Face Database (KUFD), 252 volunteer participants (182 females, 70 males) were recruited to have their photograph taken in exchange for a small fee. Each session took place in an evenly lit laboratory, where participants were photographed across various poses and whilst bearing a neutral expression. In the same session, participants were also recorded with a camcorder rotating their heads to look in different directions. Additionally, each participant consented to the use of their Student ID photograph, which was retrieved from the University's online Student Data System. These ID photographs are not constrained by expression, pose, or image-capture device, and therefore represent an important source of variability. The ID photographs were acquired a minimum of three months prior to the laboratory photograph. The mean time interval between acquisition of the laboratory photograph and the ID photograph, across all participants, was approximately 8.8 months (SD = 10.5).

With these stimuli, two versions of the KFMT were created. The short version consists of 40 Caucasian identity pairs (20 males, 20 females) from the KUFD. Each pair comprises a high-resolution portrait (Fujifilm FinePix S2980, 14-megapixel) and a student ID photograph. The portrait images were cropped to depict only the target's

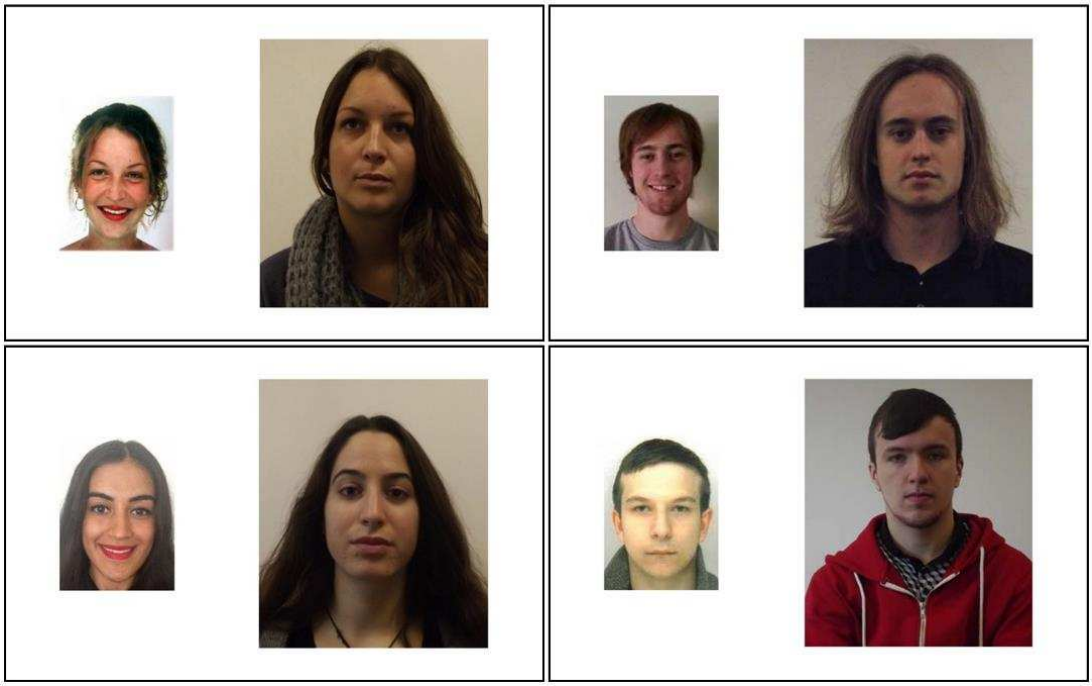

Figure 2.1. Example match (top row) and mismatch (bottom row) pairs from the KFMT.

head and shoulders, rescaled to a size of 283x332 pixels at a resolution of 72-ppi, and were placed on the right hand side of a blank white canvas. The student ID photographs measured 142x192 pixels with a resolution of 72-ppi and were positioned to the left of the digital photographs. Thus, each image pair in the KFMT comprises an optimised target photograph taken under controlled conditions, analogous to a passport photograph, but also an ambient photograph in which targets are depicted across a variety of poses, and with different facial expressions. Of the 40 image pairs that feature in the short version of the KFMT, 20 depict the same identity, whilst the remainder depict different individuals. To create these mismatch trials, target images were paired by the experimenters based on their visual similarity with regard to hair

colour, face and eyebrow shape. Example match and mismatch pairs are displayed in Figure 2.1.

The long version of the KFMT comprises 220 face pairs (166 females, 54 males) from the KUFD. Analogous to the short version, identity pairs in this test also comprise a digital portrait, which was cropped to depict only a target's head and shoulders, alongside a student ID photograph. Of these 220 image pairs, 200 depict the same identity, whilst the remainder comprise the 20 identity mismatch pairs that also feature in the short version. In contrast to the short version of the KFMT, which featured only Caucasian faces, some identity pairs in the longer version were also of Asian, Afro-Caribbean, and Middle-Eastern descent. The purpose of this longer test is to further encapsulate the difficulty of face-matching conditions in operational contexts such as passport control, by featuring a greater number of trials, and infrequent mismatches.

Experiment 1

This experiment compared performance on the short versions of the KFMT and the GFMT. Normative data show that average performance for the GFMT is around 80-90%, with individual accuracy ranging from near-chance to perfect (see, e.g., Burton et al., 2010). For the KFMT to be a useful resource in face-matching research, by providing a more challenging identification test than the GFMT, it is important to establish such a difference in performance.

Method

Participants

Sixty students (40 females, 20 males) from the University of Kent, with a mean age of 20.3 years (SD = 3.6), participated in this study in exchange for course credit or a small fee. All participants were British residents and reported normal or corrected-to-normal vision. This study was conducted in accordance with the ethical guidelines of the British Psychological Association.

Stimuli and procedure

The short version of the KFMT was employed for this comparison, as this comprises the same number of identity match trials (20) and mismatch trials (20) as the short version of the GFMT (see Burton et al., 2010). In contrast to the KFMT, one face in each pair of the GFMT consists of a digital photograph image, whilst the other comprises a still-image extracted from high-quality video footage. In each pair, targets are depicted from the front, whilst bearing a neutral expression and under even lighting. All GFMT faces are shown in greyscale, and are presented side-by-side at a width of 350 pixels, with a resolution of 72-ppi.

This experiment was run using PsychoPy software (Peirce, 2007). All participants completed the short versions of the KFMT and the GFMT, the order of which was counterbalanced across observers. Each trial was preceded by a 1-second fixation cross, which was then replaced by a stimulus pair. Observers responded using one of two keys on a standard computer keyboard. To measure best-possible accuracy, performance was self-paced in both tasks. No feedback on accuracy was provided during the experiment. As an additional measure to establish the test-retest reliability of the KFMT, 30 participants (16 females, 14 males) from the total sample completed this task twice, with a mean interval of 7.2 days (SD = 0.9) between test sessions.

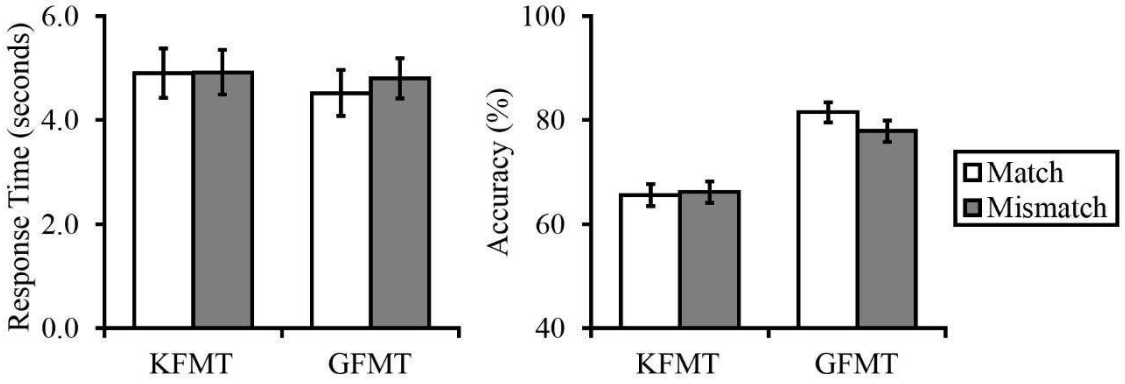

Figure 2.2. Mean correct response times and percentage accuracy scores for match and mismatch trials in the short versions of the KFMT and GFMT. Error bars represent the standard error of the mean.

Results

Response times

For each observer, mean correct response times were calculated for both tests. These are displayed in Figure 2.2 and suggest that response latencies were comparable between both tasks, with observers on average taking 5.1 and 5.6 seconds to respond on the KFMT and GFMT, respectively. To analyse these data more formally, a 2 (test: KFMT vs. GFMT) x 2 (trial type: match vs. mismatch) within-subjects analysis of variance (ANOVA) was conducted, which did not find an effect of test, $F(1,59) = 0.58$, $p = 0.45$, $\eta_p^2 = 0.01$, or of trial, $F(1,59) = 0.48$, $p = 0.49$, $\eta_p^2 = 0.01$, and these factors did not interact, $F(1,59) = 0.45$, $p = 0.51$, $\eta_p^2 = 0.01$.

Accuracy

Mean percentage accuracy for both tasks was analysed next. This is also displayed in Figure 2.2 and shows that accuracy in the KFMT was 66% for both match and mismatch trials. By comparison, overall performance in the GFMT was 80%, with

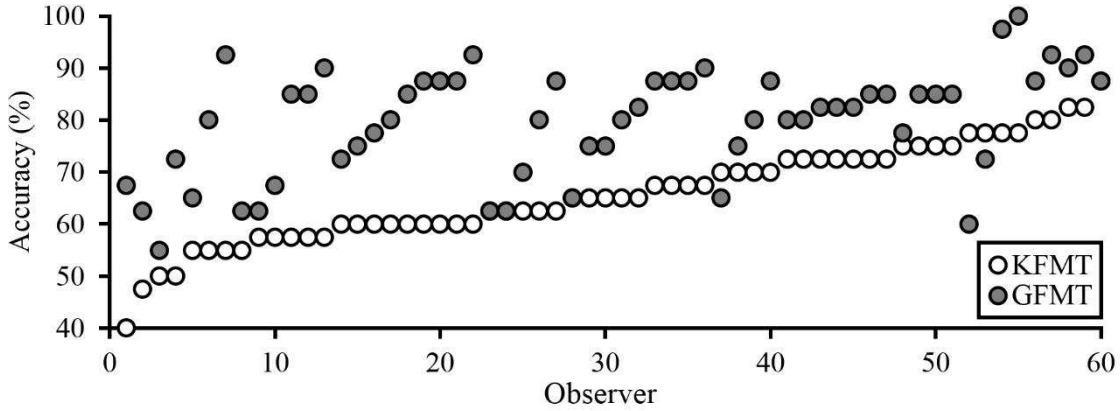

Figure 2.3. Individual data, based on overall accuracy for the KFMT and GFMT, ordered from least to most accurate observer.

slightly higher accuracy on match (82%), compared to mismatch trials (78%). This converges with the baseline level of accuracy for the GFMT in normative studies (e.g., Burton et al., 2010).

To compare performance in these tasks, a 2 (test) x 2 (trial type) within-subjects ANOVA was conducted, which did not reveal an effect of trial type, $F(1,59) = 0.32$, $p = 0.58$, $\eta_p^2 = 0.01$, or an interaction, $F(1,59) = 1.62$, $p = 0.21$, $\eta_p^2 = 0.03$, but showed that accuracy was considerably higher on the GFMT than on the KFMT, $F(1,59) = 104.73$, $p < 0.001$, $\eta_p^2 = 0.64$. This difference is confirmed by an inspection of the individual data in Figure 2.3, which shows that only three of 60 observers performed worse in the GFMT than the KFMT. Despite these differences, overall accuracy correlated for the KFMT and GFMT, $r(58) = 0.45$, $p < 0.001$, which indicates that these tasks measure similar underlying face processes.

Next, the test-retest reliability of the KFMT was analysed. Across sessions 1 and 2, overall accuracy was 66% and 67%, respectively. Overall performance across both test sessions was positively correlated, $r(28) = 0.67$, $p < 0.001$. In addition, a positive relationship was found between sessions for accuracy on match trials, $r(28) =$

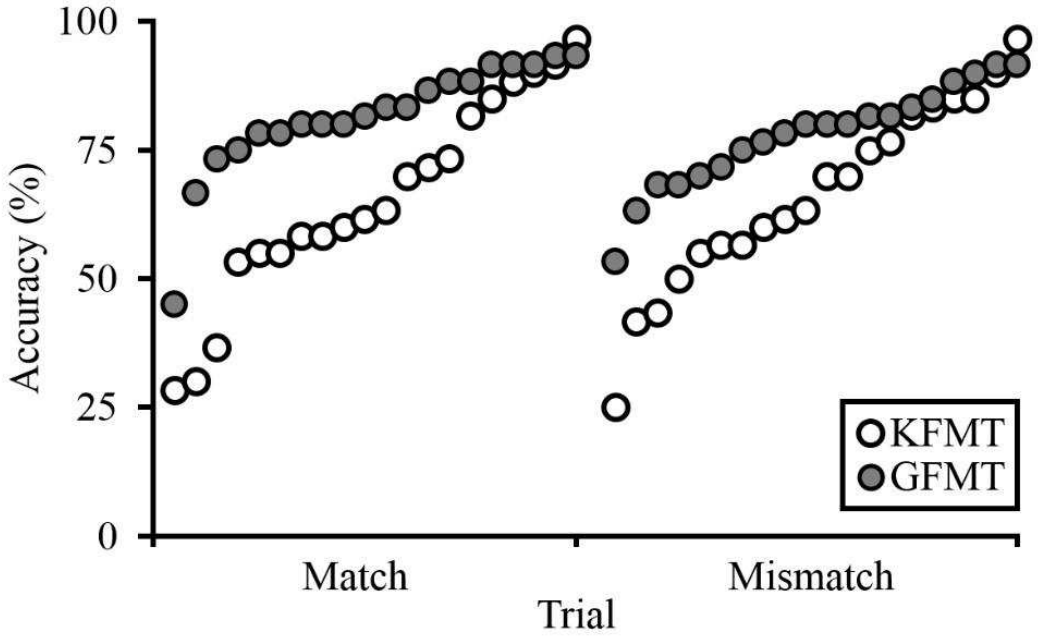

Figure 2.4. Percentage accuracy for individual items on the KFMT and the GFMT, ordered from least to most accurate.

0.68, $p < 0.001$, and on mismatch trials, $r(28) = 0.79$, $p < 0.001$. Together, these analyses indicate that the KFMT exhibits high test-retest reliability.

Finally, accuracy was analysed by item to illustrate the range of performance across different face pairs. This is illustrated for match and mismatch stimuli in Figure 2.4, ordered by item accuracy. These data reiterate that the KFMT is consistently more difficult than the GFMT, and produces a greater range in accuracy across items. In contrast to the GFMT, this range is such that some match pairs are more likely to be classified as identity mismatches, and vice versa.

d-prime and criterion

For completeness, the percentage accuracy data were also converted into signal detection measures of sensitivity (d') and response bias (criterion). Paired-sample t-tests revealed that *d'* was higher for the GFMT than the KFMT, $t(59) = 10.97$, $p <$

0.001. Criterion was comparable for these tests, $t(59) = 1.22$, $p = 0.23$, and was close to zero, both $ts \leq 1.18$, $ps \geq 0.24$.

Discussion

This experiment compared performance in a novel test of face matching – the KFMT – with the established GFMT. Accuracy on the KFMT was comparable between match and mismatch trials, and was at 66%. By comparison, overall performance on the GFMT was 80%, with slightly higher accuracy on match (82%) than mismatch (78%) trials. Converging with existing research (e.g., Bindemann, Avetisyan, et al., 2012; Burton et al., 2010; Estudillo & Bindemann, 2014), considerable individual differences emerged in both tests, with accuracy ranging from 40-88% in the KFMT, and 55-100% in the GFMT. However, the difference between tests was robust at an individual level, with only three of 60 participants recording lower accuracy on the GFMT than the KFMT. Performance between tests was positively correlated, which indicates that, despite the increased difficulty of the KFMT, both tests measure similar face processes. In addition, performance on the KFMT also correlated for observers who completed the task twice, with a week's interval between sessions. This suggests that the KFMT reliably measures similar face-matching processes over time.

These data suggest that the KFMT provides a complementary test for the GFMT that could be employed when face-matching accuracy needs to be assessed under more challenging conditions, for example, to mimic more closely applied settings such as passport control. However, in such settings, the number of to-be-matched faces typically exceeds 40 trials, and mismatches occur infrequently. To further understand performance under such conditions, a second experiment was

conducted to establish accuracy for the long version of the KFMT. To provide a comparison, this was followed by two other established tests of face processing, the CFMT (Duchaine & Nakayama, 2006) and the CFPT (Duchaine et al., 2007), which measure unfamiliar face recognition and unfamiliar face processing ability, respectively. If the KFMT provides a robust construct, then it should also correlate with these tests.

Experiment 2

In this experiment, observers completed the longer version of the KFMT, comprising 200 match trials and 20 mismatch trials. Current research shows that when matching optimised GFMT faces for a prolonged period, observers develop a response bias to erroneously classify pairs as identity matches (Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann, Fysh, et al., 2016). If the KFMT produces behavioural effects comparable to the GFMT, then such a response bias should also be found here, strengthening the results of Experiment 1.

In addition, observers also completed the CFMT (Duchaine & Nakayama, 2006) and the CFPT (Duchaine et al., 2007) upon completion of the KFMT. In contrast to the face matching task of the KFMT, the CFMT measures recognition memory for newly learned faces, whereas the CFPT requires the ordering of sequences of highly-similar face morphs. However, these three tests are unified on the basis that all focus on the identity processing of unfamiliar faces. The CFMT and CFPT have been used widely and are typically employed to assess impairments in face processing (see, e.g., Bobak et al., 2017; Ulrich et al., 2017; White et al., 2017), as well as superior recognition ability (Bobak, Hancock, et al., 2016; Bobak et al., 2017; Russell et al.,

2009). Thus, the CFMT and CFPT provide suitable tests against which performance in the KFMT can be compared.

Method

Participants

Fifty students (10 males, 40 females) from the University of Kent, with a mean age of 19.5 years (SD = 3.0), participated in this study in exchange for course credit. None of these had participated in Experiment 1. All were British residents and reported normal or corrected-to-normal vision.

Stimuli and procedure

KFMT. The long version of the KFMT comprises 220 trials, of which 200 depict the same identity, whilst the remainder are the same 20 mismatch pairs that feature in the short version of this test. These stimuli were evenly divided into four blocks of 55 trials (50 match trials, 5 mismatch trials), which were counterbalanced across observers. Administering the task in this way ensures that mismatch trials were distributed evenly throughout the task, but also allows for the opportunity to observe changes in performance over time. However, to create the impression of one continuous task, no breaks were administered between blocks. As before, this task was run on PsychoPy software (Peirce, 2007), with observers responding using one of two keys on a standard computer keyboard. At the beginning of the task, observers were informed that there would be fewer mismatch than match trials. No time pressure or feedback was administered throughout this task, and observers were encouraged to be as accurate as possible.

CFMT. Following the KFMT, observers completed the CFMT (Duchaine & Nakayama, 2006). Stimuli in this task comprise images of six male targets, along with 46 foil identities. All faces are cropped so that features such as hair and facial outline are removed, and depict evenly-lit targets bearing a neutral expression. In the first block of this task, participants study three different orientations of a single target face for three seconds, and are then required to identify the target from a three-face array containing one of the study images and two distractor faces. This is repeated for each target. In the second block, observers study six different but concurrent target faces for 20 seconds, and are then required to identify a given target from a three-face array containing two distractors and a previously-unseen view of a target face. The final block of this task is conceptually similar to Block 2, but with the addition of Gaussian noise over stimuli to further increase the difficulty of this task.

CFPT. Finally, observers completed the CFPT (Duchaine et al., 2007). On each trial, a mid-profile view of a target face is presented, along with six further faces which were created by morphing the target with six individuals by varying amounts. Observers are required to arrange these faces in order of similarity to the target face, with accuracy reflecting the number of deviations from the correct order. This task consisted of 16 trials in total, each of which lasted for a maximum duration of 60 seconds, after which a trial was terminated and the next was initiated. Additionally, half of these trials depicted upright faces, which were randomly intermixed with the remaining eight trials, in which faces were presented upside-down.

Results

KFMT

Response times

Mean correct response times were analysed first and are displayed in Figure 2.5. These reflect that match response times were faster than mismatch response times throughout the task. To analyse these data formally, a 2 (trial type: match vs. mismatch) x 4 (block: 1, 2, 3, 4) within-subjects ANOVA was conducted, which revealed an effect of trial type, $F(1,49) = 6.76$, $p < 0.05$, $\eta_p^2 = 0.12$, due to faster responses on match trials than on mismatch trials. In addition, an effect of block was found, $F(3,147) = 4.03$, $p < 0.01$, $\eta_p^2 = 0.08$. However, none of the pairwise comparisons between blocks were significant following the Bonferroni adjustment, all $ps \geq 0.08$, and these factors did not interact, $F(3,147) = 0.45$, $p = 0.71$, $\eta_p^2 = 0.01$.

Accuracy

Average accuracy for match and mismatch trials across blocks was 78% and 64%, respectively. However, the data depicted in Figure 2.5 reflect that over Blocks 1 through 4, performance on mismatch trials deteriorated from 74% to 57%, whereas accuracy on match trials increased from 71% to 82%. A 2 (trial type) x 4 (block) within-subjects ANOVA was conducted to investigate this variation in performance across blocks. This did not reveal a main effect of block, $F(3,147) = 1.19$, $p = 0.32$, $\eta_p^2 = 0.02$, but of trial type, $F(1,49) = 7.90$, $p < 0.01$, $\eta_p^2 = 0.14$, and a significant interaction, $F(3,147) = 10.64$, $p < 0.001$, $\eta_p^2 = 0.18$.

Analysis of simple main effects revealed that this was due to a deterioration in accuracy on mismatch trials, $F(3,47) = 4.00$, $p < 0.05$, $\eta_p^2 = 0.20$, which was lower in Blocks 3 and 4 compared to Block 1, both $ps < 0.05$, but was comparable between all other blocks, all $ps \geq 0.08$. The improvement in performance on match trials across blocks was also significant, $F(3,47) = 11.19$, $p < 0.001$, $\eta_p^2 = 0.42$, with higher accuracy in the final block compared to all other blocks, all $ps < 0.05$, as well as in

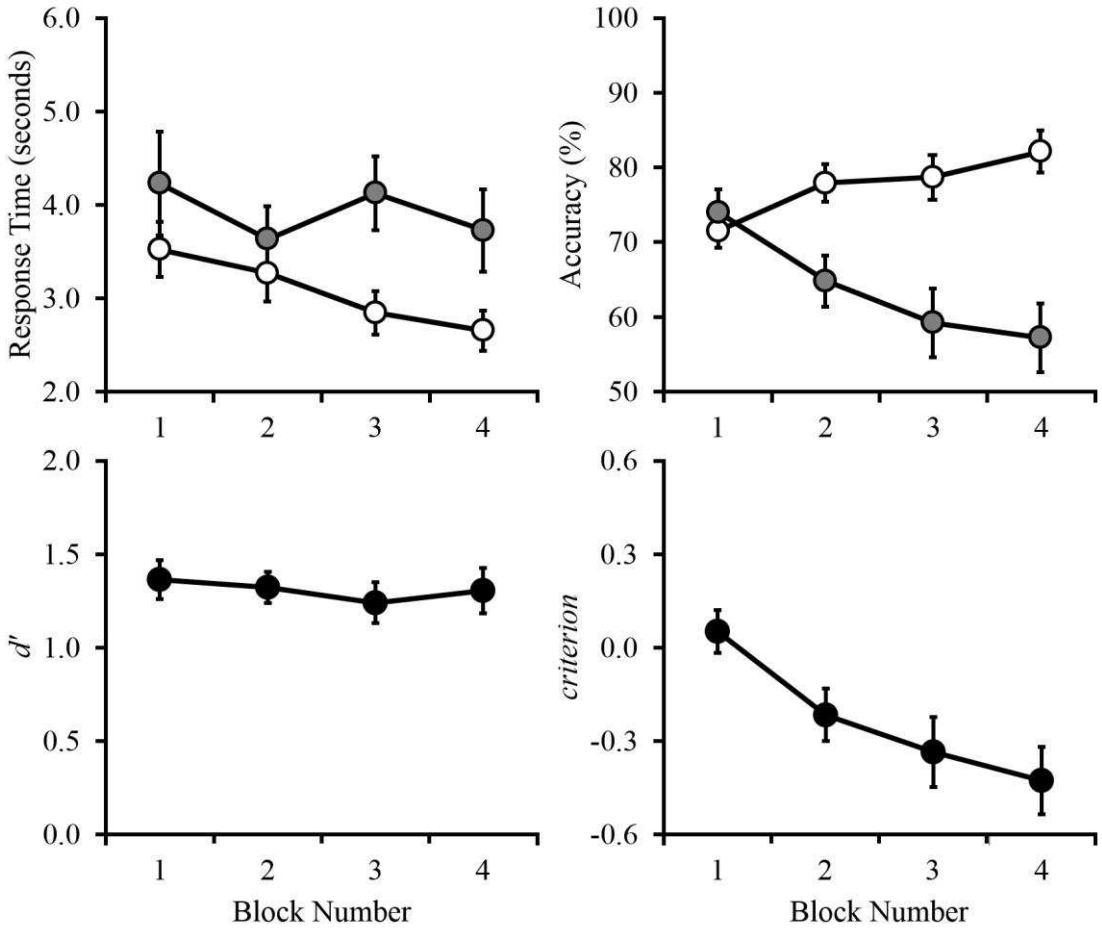

Figure 2.5. Mean correct response times, percentage accuracy, d', and criterion on the long version of the KFMT. Open markers denote match trials, and grey markers denote mismatch trials. Error bars represent the standard error of the mean.

Blocks 2 and 3 compared to Block 1, both ps < 0.001. However, performance was similar between Blocks 2 and 3, p = 1.00. In addition, performance on match and mismatch trials was comparable in the first block, $F(1,49) = 0.34$, $p = 0.56$, $\eta_p^2 = 0.01$, but accuracy on match trials was superior in the second, $F(1,49) = 6.15$, $p < 0.05$, $\eta_p^2 = 0.11$, third, $F(1,49) = 8.13$, $p < 0.01$, $\eta_p^2 = 0.14$, and final block, $F(1,49) = 14.72$, $p < 0.001$, $\eta_p^2 = 0.23$.

d-prime and criterion

Percentage accuracy scores were also converted into *d'* and criterion. A one-factor within-subjects ANOVA did not reveal an effect of block for *d'*, $F(3,147) = 0.28$, $p = 0.84$, $\eta_p^2 = 0.01$, but for criterion, $F(3,147) = 12.45$, $p < 0.001$, $\eta_p^2 = 0.20$. Pairwise comparisons revealed that this reflects a greater tendency to classify face pairs as identity matches in Blocks 2, 3, and 4, compared to Block 1, all ps < 0.01. To confirm this bias, one-sample t-tests were conducted to compare criterion to zero in each block. This revealed that criterion was comparable to zero in the first block, $t(49) = 0.76$, $p = 0.45$, but was reliably below zero in the second, $t(49) = 2.57$, $p < 0.05$, third, $t(49) = 2.98$, $p < 0.01$, and final block, $t(49) = 3.91$, $p < 0.001$.

CFMT and CFPT

Accuracy

Overall accuracy on the CFMT was at 76%, which is comparable to the average score of 80% in its normative tests (see, Duchaine & Nakayama, 2006). Average performance was at ceiling in the first block of this test, with 98% correct identifications, but deteriorated to 73% and 62% in Blocks 2 and 3, respectively.

On the CFPT, the average number of deviations (errors) from the correct order across all trials was 51.4. Performance was considerably better on upright than on inverted trials, with 35.5 versus 67.2 deviations, $t(49) = 13.75$, $p < 0.001$. The number of errors in the CFPT correlated negatively with accuracy on the CFMT, $r(48) = -0.41$, $p < 0.01$, reflecting that face memory is positively associated with face perception ability.

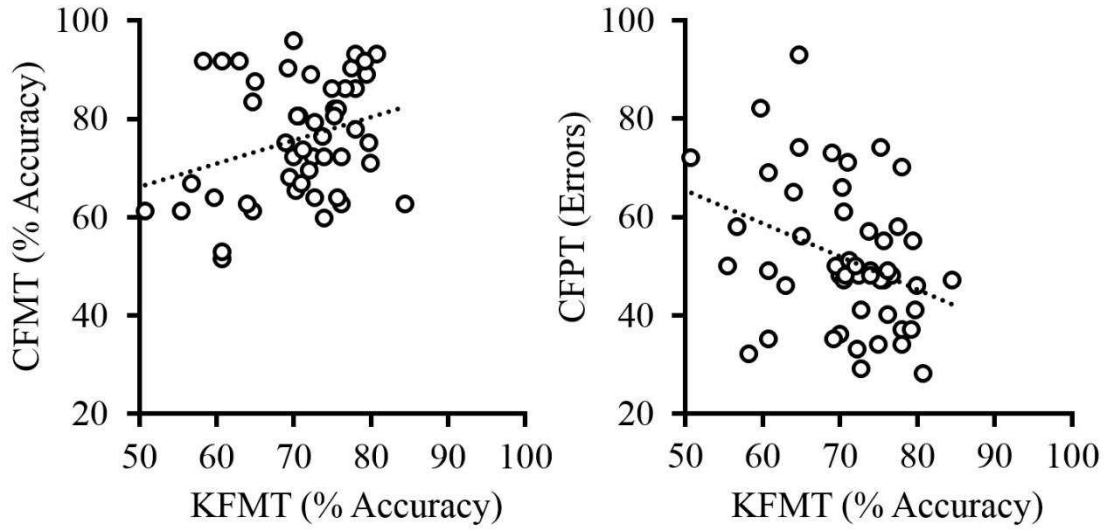

Figure 2.6. Scatter plots for overall performance on the KFMT versus the CFMT and the CFPT.

Correlations with the KFMT

To explore whether variation in performance on the KFMT was reflective of general ability in face memory and face processing, a correlation analysis was performed with the CFMT and CFPT (see Figure 2.6). This revealed a positive relationship between the KFMT and CFMT, $r(48) = 0.29$, $p < 0.05$, and a negative relationship between accuracy on the KFMT and the number of errors in the CFPT, $r(48) = -0.34$, $p < 0.05$.

Discussion

In this experiment, observers completed a longer version of the KFMT, as well as the CFMT and the CFPT. Overall performance in the KFMT was 70%, with 78% and 64% accuracy for match and mismatch trials, respectively. This is slightly higher than on the short version of the KFMT, in which overall accuracy was 66%. However, in the long version of this task, accuracy on mismatch trials deteriorated substantially

across blocks, from 74% to 57%. Conversely, performance on match trials improved, from 71% to 82%. This pattern is reflected by a shift in criterion, which indicates that a response bias to classify an increasing number of faces as identity matches emerged over time. Such a bias has also been found in recent work using the optimised stimuli of the GFMT, but with initial accuracy levels that exceed 80% (Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann, Fysh, et al., 2016). The data from Experiment 2 therefore converge with Experiment 1 to indicate that the KFMT provides a more challenging test for face matching than the GFMT, but preserves the behavioural characteristics of this test. In addition, accuracy in the KFMT also correlated with the CFMT and CFPT (see Duchaine & Nakayama, 2006; Duchaine et al., 2007). This demonstrates further that unfamiliar face matching performance on the KFMT utilises mechanisms similar to those employed for unfamiliar face memory in the CFMT and unfamiliar face perception in the CFPT.

General Discussion

This chapter presents the KFMT as a new test of face matching and examined its characteristics across two experiments. Performance on the KFMT correlated with the GFMT in Experiment 1, and also followed the accuracy profile that is found over longer experiments with this test in Experiment 2 (see, Alenezi & Bindemann, 2013; Alenezi et al., 2015). This indicates that both tests measure similar underlying processes. However, face-matching accuracy was substantially lower on the KFMT than on the GFMT, by 14%, and this effect was robust on both an individual level and by item. In addition, performance on the short version of the KFMT correlated for observers who completed this test one week apart. This demonstrates that this task measures the same processes between separate testing sessions with high reliability.

Finally, Experiment 2 showed that performance on the KFMT was associated with the CFMT and CFPT, which also measure aspects of unfamiliar face-identity processing.

Taken together, these results indicate that the KFMT is a psychometrically-stable test of unfamiliar-face matching, but the variability in the face photographs of its stimulus pairs provides a more challenging identification test than the established GFMT, which is based on optimised stimuli for person identification. It should be noted that these conclusions are based on samples here that feature an unequal sex ratio. However, sex differences exert only a numerical effect of around 5% on face-matching performance (see Megreya, Bindemann, et al., 2011), which is small compared to the very broad individual differences in face-matching accuracy between observers of the same sex (see, e.g., Burton et al., 2010; Megreya & Bindemann, 2013).

The aim of the KFMT is to facilitate further research to understand face-matching performance in the context of passport control. It is suggested that this makes the KFMT a valuable research resource to investigate factors that cannot be explored fully with the optimised identification conditions that are provided by the GFMT (see, e.g., Bindemann, Fysh, et al., 2016; Dowsett & Burton, 2015). In the next chapter, the more challenging conditions provided by the KFMT are utilised to investigate the effect of time pressure on face-matching accuracy.

Recent work (e.g., Bindemann, Fysh, et al., 2016) exploring this factor has found that when observers match faces from the GFMT, time pressure exerts only a small numerical effect on performance, of less than 11%. However, it is possible that this modest deterioration in accuracy was due to the relatively high performance that is generally observed on the GFMT. Consequently, the stimuli employed by Bindemann, Fysh, et al. (2016) may have lacked the sensitivity necessary to fully

exhibit the effects of time pressure on face-matching accuracy. Consequently, in the next chapter the more challenging conditions provided by the KFMT are utilised to investigate whether time pressure exerts a greater effect on face-matching performance when this task is completed under more the challenging conditions that are encountered in applied settings.

Chapter 3

Effects of Time Pressure and Time Passage on Face-Matching Accuracy

Introduction

The previous chapter presented the KFMT, which is intended to provide a more challenging set of conditions under which face-matching accuracy can be measured. The potential utility of this test is reflected in research that has found that under optimised conditions, such as those provided by the GFMT, some effects might be difficult to investigate due to ceiling performance (see, e.g., Bobak, Dowsett, et al., 2016; Dowsett & Burton, 2015; see also, Burton, 2013). An important purpose of the KFMT is to therefore investigate factors that might exert only a small effect on performance under optimised conditions, but might compound accuracy to a much greater extent when this task is facilitated under more demanding conditions. This aspect of the KFMT should make it possible to estimate more closely the impact that certain factors might have on performance at passport control.

One such factor is time pressure. This factor is of practical importance to applied settings, but so far, has only received limited attention in face-matching research. The aim of the current chapter, therefore, is to utilise the more challenging conditions provided by the KFMT, to explore whether the detrimental effects of time pressure are exaggerated under conditions that more closely approximate those at border control, such as when considerable within-person variability is present between representations of the same person.

Passport officers must often process high volumes of passengers within short timeframes. In the UK, for example, a key performance target for passport officers is to process 95% of passengers from the European Union (EU) and European Economic Area (EEA) within 25 minutes of joining a passport-control queue on arrival. Similarly, Australian passport officers aim to process 92% of passengers within 30 minutes. Available information suggests that these passenger processing time targets are frequently missed (see, e.g., Australia Customs and Border Protection Service, 2015; Home Affairs Committee, 2012; ICI, 2014, 2015; Toynbee, 2016). This indicates that passport officers regularly experience high levels of time pressure when processing travellers.

So far, only a few studies have investigated the effect of time pressure on face-matching accuracy. Research currently suggests that under optimised conditions, faces should be viewed for at least two seconds (O'Toole, Phillips, et al., 2007; Özbek & Bindemann, 2011), but that accuracy can benefit from longer viewing durations under more taxing conditions (O'Toole et al., 2012; White, Phillips, et al., 2015). Taken together, these findings suggest that allowing observers flexibility in the amount of time allocated to each trial, depending on the difficulty of a face-pair stimulus, could reduce errors in this task.

This is an important consideration within the context of passport control, where passport officers can devote more time to processing difficult pairs of faces, provided that this lost time can be either recouped on subsequent trials, or additional time has been accumulated through speeded decisions earlier on. This was recently investigated in one study, where time pressure was administered flexibly using a novel paradigm (Bindemann, Fysh, et al., 2016). In this paradigm, observers used two onscreen displays – a speed gauge and a progress bar – to adjust their response speed to complete

each block within a given time target. One important feature of this paradigm is that observers could use these displays to allocate more or less time to a given pair of faces, depending on how far through the block they were, and whether they were on course to meet a time target. The researchers found that under increasing time pressure, face-matching accuracy deteriorated, but improved when time pressure receded, indicating that high time pressure reduces face-matching performance. However, a separate effect of time passage was also observed, whereby observers became more likely to erroneously classify face pairs as identity matches as they progressed throughout the task. This match response bias converges with two other studies, where stimuli were also optimised but responses were self-paced (Alenezi & Bindemann, 2013; Alenezi et al., 2015), and suggests that a key factor in face matching is also the passage of time.

These findings raise some important concerns surrounding face matching at passport control, where large numbers of travellers are matched under time pressure that is administered over a sustained duration. However, the effect of time pressure observed by Bindemann, Fysh, et al. (2016) was numerically small (less than 11%) and it was difficult to specify a consistent time pressure cut-off at which performance deteriorated. Moreover, response times were consistently below 2.5 seconds, even when up to ten seconds were available per trial. These findings might arise as Bindemann, Fysh, et al. (2016) employed the highly optimised stimuli from the GFMT to measure best-possible accuracy under time pressure. Person identification in relevant applied settings necessitates, for example, the detection of infrequent identity mismatches (Bindemann et al., 2010; Papesh & Goldinger, 2014), and the matching of a passport bearer with a face photograph that was taken many months or years earlier (see Megreya et al., 2013). As a consequence, the extent to which time pressure

impacts face-matching accuracy under conditions such as these remains unclear. In this chapter, therefore, the effect of time pressure on face-matching accuracy is explored under the more challenging conditions provided by the KFMT.

Experiment 3

In this experiment, observers matched pairs of faces under time pressure. This was administered via two onscreen displays, which were constantly updated to reflect a person's average response time and the number of trials remaining. These displays were devised as an analogy to passport control at airports, where passport officers are subject to strict passenger processing time targets and can see the number of passengers in a queue that remain to be processed. In the current paradigm, the combined information provided by these displays indicated whether observers were on track to complete a block within a required timeframe. Across five blocks, time pressure systematically increased from ten to two seconds, or decreased in the reverse order. In addition, this chapter employed stimuli drawn from the KUFD as described in Chapter 2, with each pair comprising one high-quality face photograph taken under controlled conditions, and a non-controlled student ID photograph that was taken a minimum of three months earlier. These stimuli were used to more closely explore the impact of time pressure on face-matching accuracy, given that when observers match optimised faces, time pressure exerts only a small numerical effect on performance (see Bindemann, Fysh, et al., 2016). To further encapsulate face-matching conditions in practical settings, mismatches occurred infrequently in this task (see Bindemann et al., 2010; Papesh & Goldinger, 2014). The aim of this design is therefore to indicate how much time observers require to match a challenging set of stimuli, by revealing a cut-off between time pressure and accuracy. Considering that face-matching

performance also varies over the duration of the task (Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann, Fysh, et al., 2016), the data were also analysed as a function of time passage, by investigating how performance varies over the course of the experiment independently of time pressure.

Method

Participants

Eighty undergraduates from the University of Kent (17 males, 63 females), with a mean age of 20.5 years (SD = 4.4) participated in this study in exchange for course credit or a small fee. Sample size was based on previous studies in this field (e.g., Bindemann, Fysh, et al., 2016); a post-hoc analysis also confirmed that this sample size was sufficient to obtain power that satisfies the recommended level of 0.80 (Cohen, 1988). All participants reported normal (or corrected-to-normal) vision. The experiments conducted in this chapter were approved by the Ethics Committee of the School of Psychology at the University of Kent, and was conducted in accordance with the ethical guidelines of the British Psychological Association.

Stimuli

The stimuli in this study consisted of 200 face pairs from the KUFD, comprising 185 identity matches and 15 mismatches. Each pair comprised a controlled image of a target facing forwards with a neutral expression, which was taken using a 14-megapixel digital camera, against a plain white background under even lighting. These photographs were cropped to depict a target's head and shoulders, and were scaled to a size of 283x332 pixels at a resolution of 72-ppi, before being placed on the

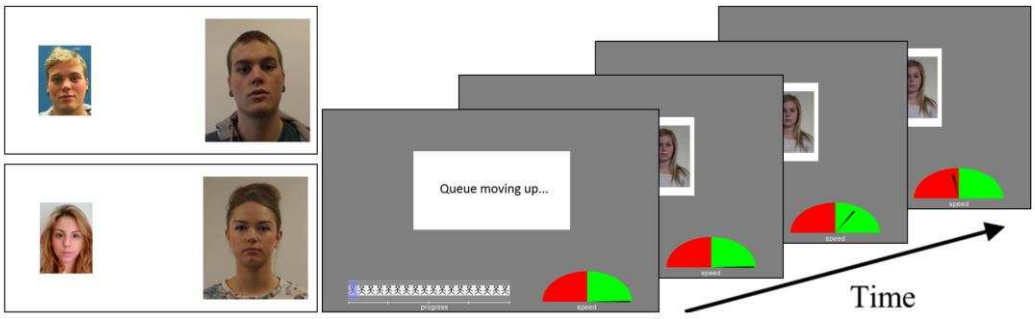

Figure 3.1. Example identity match (top) and mismatch (bottom) pairs used in the study (left), and an illustration of the stimulus screen and speed displays (right).

right-hand side of a plain white canvas. The second image in each pair consisted of a student ID photograph which was retrieved from the University of Kent's online Student Data System, and was taken a minimum of three months before the controlled image. These images were rescaled to a size of 142x192 pixels, and were also presented at an image resolution of 72-ppi, before being placed on the left-hand side of the controlled photographs. Mismatching pairs were created by selecting faces that were visually similar regarding hair colour, face and eyebrow shape. These stimuli were divided across five blocks of 40 trials (37 identity matches, and three mismatches), with no face appearing more than once.

Time pressure displays

Time pressure was implemented via two additional onscreen displays, which were presented below the stimuli (for an illustration, see Figure 3.1). One of these displays comprised a queue index indicating the number of trials remaining in the current block. This depicted a row of person icons, to represent a queue of people, and a superimposed progress bar, which advanced on each completed trial. The second display was a semi-circular speed gauge which informed participants as to whether

they were on track to meet a time target for completing the block. This was evenly divided into a green and a red zone. A dynamic needle was also presented in this display, and reflected whether participants were responding within a given time target (green zone) or were failing to meet this target (red zone). The location of the needle was updated every 100 milliseconds, so that observers could monitor the depletion and accruement of available time in real-time. The position of the needle within the speed gauge was based on a person's average response speed, calculated across the number of completed trials in a block, in comparison to the same number of trials multiplied by the set mean time target (i.e., ten, eight, six, four, or two seconds), and was proportional to how far participants were behind or ahead of the target time. These displays were reset at the beginning of each block.

Procedure

This experiment was run using PsychoPy software (Peirce, 2007). Each trial was preceded by a 1-second interval screen displaying the message "Queue moving up…", signalling the onset of the next trial. During this interval, the speed gauge and progress bars remained onscreen, so that observers could monitor their progress and adjust their speed accordingly. This interval screen was replaced with a stimulus display, which remained onscreen until a response was submitted. Participants responded by using one of two keys on a standard computer keyboard, and were instructed to be as accurate as possible at the beginning of the task, as well as between each block.

Participants completed 200 trials, which were counterbalanced across five blocks of 40 face pairs (37 identity matches and three mismatches). At the beginning of the task, participants were instructed that there would be fewer mismatching than

matching pairs, but were not informed of the exact ratio. Time pressure was implemented by adjusting the average amount of time that had to be spent on each trial to complete a block within a time target. The order of time pressure was counterbalanced across blocks, such that the available time per trial varied systematically from ten, eight, six, four, and two seconds, or vice versa.

These time targets were reflected by the needle within the speed display, which resided in the green zone if an observer was on track to complete a block within the time target, but entered the red zone if a time target was breached. The queue display was updated upon completion of each trial, reflecting how many trials remained in the block. Participants were briefed about these displays at the beginning of the experiment, and were instructed to use these to adjust their response speed accordingly. Specifically, participants were informed that it was acceptable for the needle to enter the red zone if they took more time on some of the trials, provided that lost time could be recouped on later trials. This could be achieved by responding faster on subsequent trials, so that the needle was (back) in the green zone by the end of each block.

Results

Time Pressure

Response times

The response time data were first broken down according to the level of time pressure that was imposed in each block. This showed that all observers complied with the time pressure demands of the task, with the slowest participant taking on average 9.0, 7.3, 5.6, 3.5, and 2.0 seconds in Blocks 1-5, respectively. Next, the data were broken down further into mean correct response times on match and mismatch trials,

which are depicted in Figure 3.2, and were analysed using a 2 (trial: match vs. mismatch) x 5 (time pressure: 10, 8, 6, 4, 2 seconds) within-subjects analysis of variance (ANOVA). This revealed an effect of trial, $F(1,48) = 12.62$, $p < 0.01$, $\eta_p^2 = 0.21$, which was due to faster responses on match trials. In addition, there was an effect of time pressure, $F(4,192) = 20.50$, $p < 0.001$, $\eta_p^2 = 0.30$. Bonferroni-adjusted pairwise comparisons showed that responses were fastest in the 2-second block, all ps < 0.001, followed by the 4-second block, all ps < 0.05, but were comparable between the 6-, 8-, and 10-second blocks, all ps ≥ 0.27. The interaction between time pressure and trial was not significant, $F(4,192) = 1.33$, $p = 0.26$, $\eta_p^2 = 0.03$.

Accuracy

Next, the percentage accuracy data for each time pressure condition were calculated. These scores are also depicted in Figure 3.2, and reflect that under two seconds of time pressure, accuracy on mismatch trials deteriorated to 58%, whilst performance on match trials appeared comparable across all time pressure conditions. A 2 (trial) x 5 (time pressure) within-subjects ANOVA found an effect of trial, due to higher accuracy on match trials, $F(1,79) = 20.12$, $p < 0.001$, $\eta_p^2 = 0.20$, as well as an effect of time pressure, $F(4,316) = 3.91$, $p < 0.01$, $\eta_p^2 = 0.05$. Bonferroni-adjusted comparisons showed that this was due to higher accuracy when time pressure was ten seconds, compared to four and two seconds, both ps < 0.05. The difference in accuracy between ten and eight seconds was also approaching significance, $p = 0.05$. However, no other comparisons were significant, all ps ≥ 0.60, and these factors did not interact, $F(4,316) = 2.26$, $p = 0.06$, $\eta_p^2 = 0.03$.

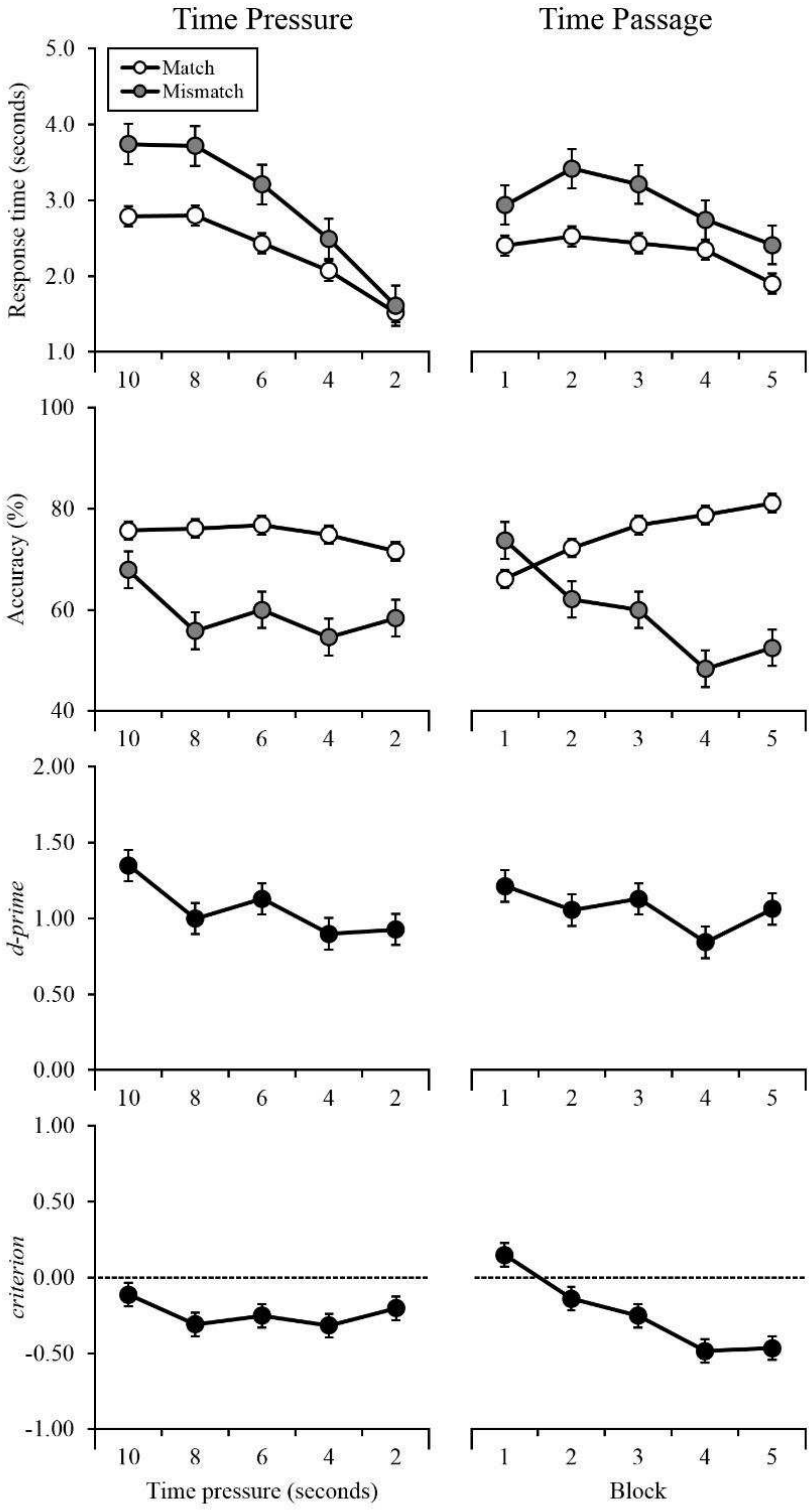

Figure 3.2. Mean correct response times, percentage accuracy, d', and criterion across time pressure conditions, as well as over the passage of time, for Experiment 3. Open markers denote match trials, and grey markers denote mismatch trials. Error bars represent the standard error of the mean.

d-prime and criterion

For completeness, the percentage accuracy data were also converted to signal detection measures *d'* and criterion, to measure overall sensitivity (accuracy) and response bias, respectively. For *d'*, a one-way within-subjects ANOVA revealed a small but significant effect of time pressure, $F(4,316) = 3.66$, $p < 0.01$, $\eta_p^2 = 0.04$, due to lower sensitivity under four and two seconds of time pressure compared to the 10-second condition, both ps < 0.05. However, sensitivity was comparable between all other blocks, all ps ≥ 0.09.

The analogous analysis of criterion did not find an effect of time pressure, $F(4,316) = 2.06$, $p = 0.09$, $\eta_p^2 = 0.03$, indicating that observers' response patterns did not vary across time pressure conditions. However, this does not rule out the possibility that a response bias was present throughout the task. To explore this further, therefore, criterion in each block was compared to zero using a series of one-sample t-tests. This revealed that response criterion was close to zero under ten seconds of time pressure, $t(79) = 1.48$, $p = 0.14$, but was reliably below zero when time pressure was eight, $t(79) = 4.07$, $p < 0.001$, six, $t(79) = 3.42$, $p < 0.01$, four, $t(79) = 4.13$, $p < 0.001$, and two seconds, $t(79) = 2.46$, $p < 0.05$. These results indicate that a match response bias was present in each time pressure condition except the 10-second block.

Time Passage

Response times

Next, the data were analysed according to time passage. For this purpose, the data were collapsed across increasing and decreasing time pressure conditions and analysed by block order. As with the analysis of time pressure, mean correct response times were analysed first, and are displayed in Figure 3.2. A 2 (trial: match vs.

mismatch) x 5 (block: 1, 2, 3, 4, 5) within-subjects ANOVA revealed an effect of trial, $F(1,48) = 12.62$, $p < 0.01$, $\eta_p^2 = 0.21$, due to faster responses on match trials. In addition, an effect of block was found, $F(4,192) = 4.50$, $p < 0.01$, $\eta_p^2 = 0.09$, due to faster responses in Block 5 compared to Blocks 2, 3, and 4, all ps < 0.05. However, no further comparisons were significant, all ps ≥ 0.20, and trial type did not interact with block, $F(4,192) = 0.99$, $p = 0.42$, $\eta_p^2 = 0.02$.

Accuracy

Percentage accuracy scores were calculated for each block, collapsed across order of time pressure. These data are also depicted in Figure 3.2, and reflect that performance on mismatch trials deteriorated across blocks, from 74% in Block 1 to 53% in Block 5. A 2 (trial) x 5 (block) within-subjects ANOVA revealed an effect of trial, $F(1,79) = 20.10$, $p < 0.001$, $\eta_p^2 = 0.20$, as well as of block, $F(4,316) = 2.46$, $p < 0.05$, $\eta_p^2 = 0.03$, and a significant interaction, $F(4,316) = 24.59$, $p < 0.001$, $\eta_p^2 = 0.24$.

Simple main effects analysis for this interaction revealed that accuracy on match trials improved across blocks, $F(4,76) = 29.67$, $p < 0.001$, $\eta_p^2 = 0.61$, with Bonferroni-adjusted pairwise comparisons showing that accuracy was higher in all blocks following the first and second block, all ps < 0.01, as well as in the final block compared to Block 3, $p < 0.01$. However, performance was comparable between Blocks 4 and 5, and between Blocks 3 and 4, both ps ≥ 0.19. The deterioration on mismatch trials was also significant, $F(4,76) = 8.66$, $p < 0.001$, $\eta_p^2 = 0.31$, with worse accuracy in all blocks following Block 1, all ps < 0.05, as well as in Block 4 compared to Block 2, $p < 0.01$. Performance was comparable between the remaining blocks, all ps ≥ 0.10.

A simple main effect of trial was also found in the second, $F(1,79) = 5.73$, $p < 0.05$, $\eta_p^2 = 0.07$, third, $F(1,79) = 13.52$, $p < 0.001$, $\eta_p^2 = 0.15$, fourth, $F(1,79) = 41.57$, $p < 0.001$, $\eta_p^2 = 0.35$, and final blocks, $F(1,79) = 33.80$, $p < 0.001$, $\eta_p^2 = 0.30$, reflecting higher accuracy on match compared to mismatch trials. By contrast, mismatch accuracy was higher than match accuracy in Block 1, but this difference failed to reach significance, $F(1,79) = 3.91$, $p = 0.05$, $\eta_p^2 = 0.05$.

d-prime and criterion

Percentage accuracy scores were again transformed into d' and criterion. The analysis of *d'* revealed that overall sensitivity was comparable across blocks, $F(4,316) = 2.00$, $p = 0.10$, $\eta_p^2 = 0.03$. However, there was an effect of block on criterion scores, $F(4,316) = 26.28$, $p < 0.001$, $\eta_p^2 = 0.25$, due to a significantly lower response criterion in Blocks 4 and 5 compared to all preceding blocks, all ps < 0.05, as well as in Blocks 2 and 3 compared to the first block, both ps < 0.001. However, criterion was comparable between the second and third block, $p = 0.84$.

As before, these scores were also compared to zero. One-sample t-tests revealed that criterion was above zero in Block 1, $t(79) = 2.42$, $p < 0.05$, due to a higher number of mismatch responses at the beginning of the task. By contrast, criterion was below zero in the second, $t(79) = 2.06$, $p < 0.05$, third, $t(79) = 3.42$, $p < 0.01$, fourth, $t(79) = 6.14$, $p < 0.001$, and final block, $t(79) = 5.60$, $p < 0.001$. These results indicate that over time, observers became increasingly likely to classify faces as identity matches.

Discussion

This experiment investigated the effects of time pressure and time passage on face-matching performance. Time pressure appeared to specifically impact performance on mismatch trials, whereby accuracy deteriorated as the average time target per trial was reduced. This is evident from *d'*, which reflected that performance was worst in the 4- and 2-second condition. Response criterion did not vary across the different levels of time pressure, but was reliably below zero in all conditions following the 10-second condition, reflecting that observers were more prone to classifying stimuli as identity matches in the 8-, 6-, 4-, and 2-second conditions. This bias could be attributed to observers' knowledge that mismatches would be occurring less frequently than matches over the task. However, other research has shown that even when match and mismatch trials occur with equal frequency, a similar response bias emerges, but is exacerbated by time pressure (Bindemann, Fysh, et al., 2016).

In addition, an effect of time passage was also observed in this experiment, whereby accuracy on match trials improved from 66% to 81% between Blocks 1 and 5, but also deteriorated on mismatch trials from 74% to 53%. This pattern reflects a shift in response criterion and shows that observers adopted a bias to classify more face pairs as identity matches over time.

These findings are consistent with those of Bindemann, Fysh, et al. (2016), where face matching was most error-prone under two seconds of time pressure. Likewise, performance in the current experiment was lowest under 4- and 2-second time targets, but did not differ between these conditions. In addition, these findings converge with studies that found a response bias to emerge over time, whereby observers make an increasing number of erroneous identity match responses over the course of an experiment (Alenezi & Bindemann, 2013; Alenezi et al., 2015).

Experiment 4

Experiment 3 indicates that time pressure and time passage exert distinct effects on face-matching performance. Time pressure reduces accuracy (%) and sensitivity (*d'*) as the time available to match faces decreases, but does not affect observers' decision criterion. By contrast, sensitivity (*d'*) is not affected by time passage, but criterion decreases over the course of the experiment, reflecting a bias to make increasingly more identity-match decisions. However, not all aspects of the results were clear-cut. For example, the time pressure analysis also revealed a match response bias (criterion) in all conditions except for the 10-second condition, and a marginally non-significant interaction of time pressure and trial type.

The aim of Experiment 4 was therefore two-fold. Firstly, this experiment sought to replicate the distinct effects that time pressure and time passage appear to exert in Experiment 3. Secondly, the aim of Experiment 4 was to also clarify marginal effects, such as the non-significant interaction of time pressure and trial type. Considering that observers' mean response times were substantially below the target time of the 10-second condition in Experiment 3, this condition was excluded in Experiment 4. In turn, this exclusion enabled us to increase the number of data points for each time pressure condition, by distributing surplus trials across the remaining blocks. Thus, in Experiment 4 observers completed four blocks of face-matching trials, where time pressure varied between eight, six, four, and two seconds.

Method

Participants

Sixty undergraduates (10 males, 50 females) with a mean age of 20 years (SD = 3.3) participated in this experiment in exchange for course credit or a small fee.

None of these had participated in the previous experiment, and all reported normal, or corrected-to-normal vision.

Stimuli and procedure

As in the previous experiment, this experiment featured 200 face pairs extracted from the KUFD. One identity match from Experiment 3 was replaced with a mismatch trial, resulting in 184 match trials, and 16 mismatches. These were evenly divided over four blocks of 50 face pairs (46 match, 4 mismatch), and were counterbalanced across participants, with no pair appearing more than once for each observer.

The procedure was identical to the previous experiment, except for the difference that instead of five blocks where time pressure increased or decreased from ten to two seconds, this task comprised four blocks, with time targets varying systematically from eight to two seconds. To further encapsulate time pressure, the interval between each trial was reduced to 500ms, and observers could receive up to three verbal prompts per block. These consisted of "please speed up", "you must speed up", and "go faster!", and were only issued if the needle was in the red zone at 25%, 50%, or 75% block completion, respectively. All other aspects of the procedure, such as the speed gauge and the progress bar, were unchanged.

Results

Time Pressure

Response times

As in Experiment 3, response times were analysed first. The slowest observer took 7.3, 5.1, 3.3, and 1.4 seconds to complete the 8-, 6-, 4-, and 2-second condition,

respectively. These data were next broken down into mean correct response times on match and mismatch trials, which are depicted in Figure 3.3. A 2 (trial: match vs. mismatch) x 4 (time pressure: 8, 6, 4, 2 seconds) within-subjects ANOVA revealed that responses on match trials were faster than on mismatch trials, $F(1,52) = 17.00$, $p < 0.001$, $\eta_p^2 = 0.25$. There was also an effect of time pressure, $F(3,156) = 50.75$, $p < 0.001$, $\eta_p^2 = 0.49$, due to faster responses in the 2-second condition compared to the 4-, 6-, and 8-second conditions, all ps < 0.001. In addition, responses in the 4-second condition were faster than in the 6- and 8-second conditions, both ps < 0.001. The difference between the 6- and 8-second conditions was approaching significance, $p = 0.05$. These factors did not interact, $F(3,156) = 0.82$, $p = 0.48$, $\eta_p^2 = 0.02$.

Accuracy

Percentage accuracy scores for this experiment are also depicted in Figure 3.3, and show that under four and two seconds of time pressure, accuracy on mismatch trials deteriorated to 54%, whilst performance on match trials remained comparable across all time pressure conditions. To analyse these data, a 2 (trial) x 4 (time pressure) within-subjects ANOVA was conducted. This revealed an effect of time pressure, $F(3,177) = 3.71$, $p < 0.05$, $\eta_p^2 = 0.06$, as well as an effect of trial, $F(1,59) = 10.58$, $p < 0.01$, $\eta_p^2 = 0.15$, and an interaction, $F(3,177) = 4.72$, $p < 0.01$, $\eta_p^2 = 0.07$.

Simple main effects analysis revealed that performance on match and mismatch trials was comparable in the 6-second, $F(1,59) = 1.64$, $p = 0.21$, $\eta_p^2 = 0.03$, and 8-second conditions, $F(1,59) = 2.23$, $p = 0.14$, $\eta_p^2 = 0.04$, whereas accuracy was higher on match trials in the 4-second, $F(1,59) = 15.90$, $p < 0.001$, $\eta_p^2 = 0.21$, and 2-second conditions, $F(1,59) = 15.36$, $p < 0.001$, $\eta_p^2 = 0.21$. In addition, there was a simple main effect of time pressure on mismatch trials, $F(3,57) = 5.07$, $p < 0.01$, $\eta_p^2 =$

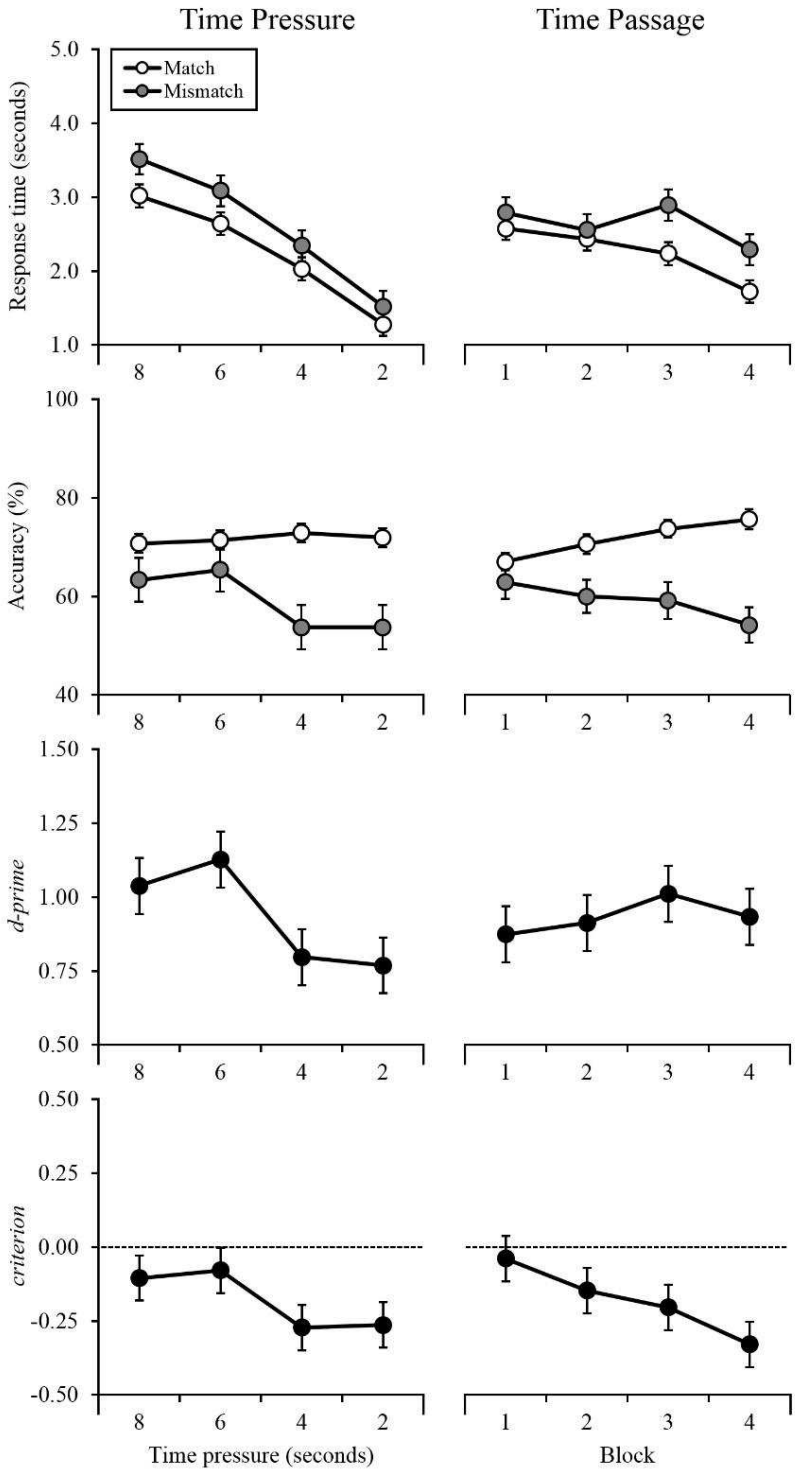

Figure 3.3. Mean correct response times, percentage accuracy, d', and criterion across time pressure conditions, as well as over the passage of time, for Experiment 4. Open markers denote match trials, and grey markers denote mismatch trials. Error bars represent the standard error of the mean.

0.21. Bonferroni-adjusted pairwise comparisons showed that this was due to higher accuracy in the 6-second condition compared to the 4-second condition, $p < 0.05$. However, performance was comparable between all other conditions, all $ps \geq 0.07$. There was no effect of time pressure on match trials, $F(3,57) = 0.68$, $p = 0.57$, $\eta_p^2 = 0.03$.

d-prime and criterion

The percentage accuracy data were converted to signal detection measures *d'* and criterion to measure overall performance and response bias. For *d'*, ANOVA found an effect of time pressure, $F(3,177) = 4.00$, $p < 0.01$, $\eta_p^2 = 0.06$, due to worse performance under four seconds of time pressure compared to six seconds, $p < 0.05$. However, sensitivity was comparable across all other blocks, all $ps \geq 0.07$. The analogous analysis of criterion also revealed an effect of time pressure, $F(3,177) = 4.04$, $p < 0.01$, $\eta_p^2 = 0.06$, which was due to a shift in response criterion between the 4- and 6-second block, $p < 0.05$. No other comparisons reached significance, all $ps \geq 0.11$.

In an additional step, the criterion scores for each time pressure condition were also compared to zero using one-sample t-tests. This revealed that criterion was comparable to zero under eight, $t(59) = 1.31$, $p = 0.20$, and six seconds of time pressure, $t(59) = 1.05$, $p = 0.30$, but was reliably below zero under four, $t(59) = 3.63$, $p < 0.01$, and two seconds of time pressure, $t(59) = 3.58$, $p < 0.01$. This shows that under strict time pressure targets of four and two seconds, observers exhibit a bias to classify more face pairs as depicting the same person.

Time Passage

Response times

Next, the data were analysed according to time passage. These data are displayed in Figure 3.3, and reflect that responses generally became faster over time. A 2 (trial: match vs. mismatch) x 4 (block: 1, 2, 3, 4) within-subjects ANOVA revealed an effect of trial, $F(1,52) = 17.00$, $p < 0.001$, $\eta_p^2 = 0.25$, due to faster responses on identity match trials. There was also an effect of block, $F(3,156) = 2.99$, $p < 0.05$, $\eta_p^2 = 0.05$, due to faster responses in the final block compared to the third, $p < 0.01$. However, no further comparisons were significant, all $ps \geq 0.28$, and these factors did not interact, $F(3,156) = 2.40$, $p = 0.07$, $\eta_p^2 = 0.04$.

Accuracy

To determine whether performance was declining over time, percentage accuracy scores were next examined for each block. Breaking down the data in this way revealed that accuracy on mismatch trials decreased from 63% in Block 1, to 54% in Block 4. Conversely, performance on identity match trials improved over time, from 67% in Block 1 to 76% in Block 4. A 2 (trial: match vs mismatch) x 4 (block: 1, 2, 3, 4) within-subjects ANOVA did not reveal an effect of block, $F(3,177) = 0.23$, $p = 0.88$, $\eta_p^2 = 0.00$, but an effect of trial, $F(1,59) = 10.58$, $p < 0.01$, $\eta_p^2 = 0.15$, and a significant interaction, $F(3,177) = 5.24$, $p < 0.01$, $\eta_p^2 = 0.08$.

Simple main effects analysis for this interaction revealed that performance on match and mismatch trials was comparable in Block 1, $F(1,59) = 0.88$, $p = 0.35$, $\eta_p^2 = 0.02$, but was significantly higher on match trials in the second, $F(1,59) = 5.28$, $p < 0.05$, $\eta_p^2 = 0.08$, third, $F(1,59) = 8.42$, $p < 0.01$, $\eta_p^2 = 0.13$, and fourth block, $F(1,59) = 18.21$, $p < 0.001$, $\eta_p^2 = 0.24$. In addition, a simple main effect of block was found on match trials, $F(3,57) = 9.75$, $p < 0.001$, $\eta_p^2 = 0.34$. Bonferroni-adjusted comparisons

showed that accuracy was higher in Block 4 compared to Blocks 1 and 2, both ps < 0.001, and in Block 3 compared to Block 1, $p < 0.001$. However, performance was comparable between the second and third, and the second and first block, both ps ≥ 0.09. The deterioration on mismatch trials was not significant, $F(3,57) = 1.70$, $p = 0.18$, $\eta_p^2 = 0.08$.

d-prime and criterion

The percentage accuracy data were again converted into *d'* and criterion. For *d'*, ANOVA did not reveal an effect of block, $F(3,177) = 0.41$, $p = 0.75$, $\eta_p^2 = 0.01$. However, this effect was present for criterion, $F(3,177) = 5.81$, $p < 0.01$, $\eta_p^2 = 0.09$, which was lower in the final block, compared to Block 1, $p < 0.01$. No other comparisons were significant, all ps ≥ 0.10. One-sample t-tests were conducted to compare the criterion scores in each block to zero. This analysis revealed that criterion was comparable to zero in Block 1, $t(59) = 0.58$, $p = 0.56$, but was reliably below zero in the second, $t(59) = 2.06$, $p < 0.05$, third, $t(59) = 2.53$, $p < 0.05$, and final block, $t(59) = 4.00$, $p < 0.001$. This indicates that a bias emerged after Block 1 to classify face pairs as identity matches.

Discussion

This experiment found that face-matching performance again deteriorated under time pressure targets of four and two seconds. Numerically, this effect accounted for 11% of errors on mismatch trials between the 8- and 4-second conditions and provides evidence that time pressure is detrimental to the detection of mismatching identities. Moreover, a reduction in *d'* was observed between the 6- and 4-second conditions, in conjunction with a match response bias. Overall, these

findings suggest that four seconds represents a possible mean cutoff time at which face matching deteriorates, due to a bias to classify face pairs as identity matches.

Converging with the previous experiment, the separate analysis of time passage also revealed this match response bias across blocks. Due to this bias, performance on match trials improved from 67% to 76% in Blocks 1-4. This provides further evidence that over the passage of time, observers become more prone to perceive two faces in a pair as the same identity.

General Discussion

This chapter investigated the effects of time pressure on face-matching accuracy. Across two experiments, time pressure was administered flexibly via two onscreen displays that allowed observers to monitor whether they were on track to meet a time target, or were required to speed up (see Bindemann, Fysh, et al., 2016, for a similar design). The effect of time pressure was clearest in response times, which decreased gradually across the six, four, and two second conditions in both experiments. Importantly, both experiments also revealed an effect of time pressure on accuracy, where *d'* deteriorated at four seconds relative to more liberal time targets, but was comparable to time targets of two seconds. However, in numerical terms, these effects were relatively small, accounting for only 7% and 5% additional errors in the 4-second compared to the 10-second and 6-second condition in Experiments 3 and 4, respectively.

These findings converge with recent work where time pressure exerted only a small effect on face-matching performance, and accounted for less than 11% of errors (Bindemann, Fysh, et al., 2016). It was reasoned a priori that this small effect might have been due to the optimised stimuli employed in this research, for which accuracy

is generally high (see, e.g., Burton et al., 2010; Estudillo & Bindemann, 2014). Contrary to this prediction, however, the current study obtained a comparable effect of time pressure on face-matching performance. This occurred in a context where general performance was considerably poorer than that observed by Bindemann et al. (2016). This poor general performance converges with additional work where to-be-compared stimuli portrayed more within-person variation (Megreya et al., 2013), and mismatches were rare (Papesh & Goldinger, 2014). Taken together, these findings suggest that time pressure only exerts a relatively moderate effect on face matching, both with optimized stimuli (Bindemann, Fysh, et al., 2016) and under the more taxing conditions of the current experiments.

It is worth noting that these results were obtained in a context where average response times were consistently below the target threshold in all time pressure conditions. In the 8-second condition, for example, average response times of 3.0 and 3.2 seconds were obtained for Experiment 3 and 4, respectively. These fast responses are surprising given that observers were instructed at the beginning to use the onscreen displays to adjust their speed accordingly. Similar response patterns were observed by Bindemann, Fysh, et al. (2016), who found that response times were consistently below 2.5 seconds even when ten seconds were available per trial. The researchers considered whether this could be due to a lack of motivation from student observers to fully utilise the available time on each trial. An alternative explanation could be that observers consistently underestimated the difficulty of matching unfamiliar faces. This makes sense when considering studies where observers generalise their ability to match and identify familiar faces, which is comparatively high, to the more difficult identification of unfamiliar faces, and so fail to anticipate errors that arise in such tasks (Bindemann, Attard, & Johnston, 2014; Ritchie et al., 2015). This is also supported by

evidence that passport officers take longer than students in face-matching tasks but are not more accurate (White, Kemp, Jenkins, Matheson, et al., 2014). At present, however, it is unclear how observers allocate their processing time on a trial-by-trial basis in face matching. Research suggests that some expert observers incur greater benefits than student controls when additional time is provided (White, Phillips, et al., 2015). This indicates that there is an effective strategy of time allocation in face matching, and should be explored in future research.

Although only a small number of errors could be attributed to time pressure in the current study, a strong effect of time passage was also consistently detected in both experiments. This was characterised by a match response bias that emerged over time, and accounted for up to 21% of errors on mismatch trials. In numerical terms, the passage of time therefore appears to exert a more detrimental effect on face-matching accuracy than time pressure, particularly on the detection of identity mismatches. This time-passage effect has also been demonstrated in three other studies, where mismatch accuracy deteriorated to below chance levels when optimised faces were matched under self-paced conditions (Alenezi & Bindemann, 2013; Alenezi et al., 2015) and under time pressure (Bindemann, Fysh, et al., 2016). This therefore appears to be a robust effect, although its cause remains unclear (see Alenezi et al., 2015).

In this chapter, it is notable that the effects of time pressure and time passage were obtained through separate analysis, for which the data were ordered either by the time pressure conditions, which were counterbalanced across observers, or by block order in the experiment. These data transformations as well as the different characteristics of time pressure and time passage demonstrate that these effects are qualitatively different, but can concurrently influence face matching. This raises concerns for applied settings that rely on face matching, such as person identification

at passport control. In those settings, personnel experience time pressure frequently (see, e.g., Home Affairs Committee, 2012; ICI, 2014, 2015; Toynbee, 2016) whilst also performing face matching over prolonged periods. Time pressure effects may be exacerbated further in applied settings by the requirement to check additional person information, such as names, nationality and travel documents (see Lee, Vast, & Butavicius, 2006; McCaffery & Burton, 2016). The experiments in this chapter, which encompassed only 200 trials per participant and required face matching only, may therefore still underestimate the impact of time pressure and time passage in applied settings (see Alenezi & Bindemann, 2013; Alenezi et al., 2015).

One solution to the concerns raised by these findings is the implementation of Automated Border Control (ABC) at passport control. These systems are becoming increasingly ubiquitous in applied settings, and use state-of-the-art face recognition algorithms to verify travellers' identities. Importantly, the operation of these algorithms should be unaffected by factors such as time pressure and time passage. However, given that the true accuracy of ABC systems remains unknown in operational contexts, a human operator is always present to ensure that the algorithm does not make an incorrect identification, such as incorrectly accepting an impostor identity as a match, or vice versa (FRONTEX, 2015a). The accuracy of this human-computer interaction is explored in Chapter 4, to investigate the extent to which human operators are biased by the identification decisions of algorithms, and whether a human can reliably override an incorrect identity judgement.

Chapter 4

Human-Computer Interaction in Face Matching

Introduction

The previous chapters demonstrate that practically-relevant factors such as within-target variation, time pressure, and time passage, compound human performance in face-matching tasks. Automated Border Control (ABC) systems present a potential solution to this problem. In the UK, for example, "Electronic Passport Gates", or "e-Gates", are now installed in most major airports. These e-Gates employ state-of-the-art facial recognition algorithms that compare live travellers to a digital photograph that is stored on their passports, and are unaffected by factors that impact human capacity for face matching, such as time pressure (Bindemann, Fysh, et al., 2016), time passage (Alenezi & Bindemann, 2013; Alenezi et al., 2015), and sleep deprivation (Beattie et al., 2016). These benefits are corroborated by studies in which face recognition algorithms have achieved perfect or near-perfect performance in benchmark tests (see, e.g., Phillips et al., 2010; see also Jenkins & Burton, 2008b).

Despite these advantages, however, it remains difficult to establish the accuracy of automatic facial recognition systems in applied contexts. For example, algorithms outperform human observers in tests that are considered to be of easy and moderate difficulty (O'Toole, Phillips, et al., 2007; O'Toole et al., 2012). However, under more challenging conditions that more closely approximate passport control, such as when to-be-compared stimuli are photographed on different days, these

algorithms perform comparably to some observers (O'Toole et al., 2012; Phillips & O'Toole, 2014), and are defeated by expert matchers (White, Phillips, et al., 2015). Some studies have also reported instances where face recognition algorithms failed to score even a single hit in matching tasks, whilst humans were well above chance (Rice, Phillips, Natu, An, & O'Toole, 2013). Together, these findings indicate that algorithms are not yet fully capable of supplanting humans at border control.

Currently, e-Gates function under the supervision of human operators, who manage exceptions such as when the system cannot fully resolve a traveller with their passport photograph. A further key responsibility of these operators is to prevent the system from incorrectly accepting a mismatching identity or incorrectly rejecting a genuine match (FRONTEX, 2015a, 2015b). Such errors are projected to occur only rarely, with the false acceptance of impostors estimated to occur on 0.1% of trials, and the false rejection of identity matches on 5-10% of trials (FRONTEX, 2015b). These error rates are not represented in applied contexts, where e-Gates have been reported to reject high volumes of identity matches (ICI, 2014; Watt, 2016), and to falsely accept some egregious mismatches, such as men as women (http://www.bbc.co.uk/news/uk-england-manchester-12482156; ICI, 2011). These surprising errors indicate that the interaction between humans and e-Gates is crucial for the accuracy of person identification at passport control.

The accuracy of this human-computer interaction is currently unknown. Research shows that human decisions in face matching can be biased by external factors, such as the concurrent presentation of biographical information (McCaffery & Burton, 2016). In addition, observers appear to possess limited insight into their identification decisions, to the extent that they will affirm ownership of decisions that in fact negate their previous responses (Sauerland et al., 2016). Together, these studies

indicate that humans might be unreliable at detecting incorrect identifications made by e-Gates in applied contexts.

So far, only limited research has explored this issue. In one study, the face-matching decisions of humans and algorithms were aggregated together, resulting in near-perfect performance (O'Toole, Abdi, Jiang, & Phillips, 2007). However, the judgements of human observers in this study were independent of those made by algorithms, which differs from applied settings, where human operators instead validate a priori judgements by e-Gates. A more recent study investigated the performance of facial review staff, who use state-of-the-art face recognition algorithms to process new passport applications (White, Dunn, et al., 2015). In this task, the algorithm compares the face of a passport applicant across a database of existing passport holders to prevent fraudulent applications from being processed. The algorithm returns eight candidates who most closely resemble the applicant, which are then studied by the human operator to ensure that the applicant's photograph does not match that of any existing passport holders. Importantly, the researchers found that the accuracy of facial review staff actually limited the success of the algorithm, which could reliably return a matching identity from a database of over a million candidates.

This research reflects that person identification accuracy might not benefit from this human-computer interaction. However, crucial differences between the role of facial review staff and human operators at passport control make it difficult to generalise White, Dunn, et al.'s (2015) findings to the latter context. For example, facial review staff are required to check a single candidate image against eight highly similar face photographs, to safeguard against fraudulent passport applications. By contrast, the operators of e-Gates at passport control perform a secondary comparison on pairs of faces, to verify that the system has made the correct decision. As a

consequence, the question of whether this interaction between humans and algorithms improves identity verification at passport control remains unresolved.

This question is explored in the current chapter. Across three experiments, observers matched pairs of faces that were labelled as depicting the "same" person, "different" individuals, or that were "unresolved". Labels that provided a same or different resolution were generally consistent with the faces shown. However, a small percentage of these also provided inconsistent information. In these cases, match trials were incorrectly labelled as different individuals, and mismatch trials were labelled as depicting the same person. Unresolved trials were chosen as an analogy to the exceptions at e-Gates when a traveller cannot be matched by the algorithm, and thus must be processed by the human operator. The aim of Experiment 5 was to determine accuracy with these trial labels. In subsequent experiments, it was investigated how performance is further affected when observers do not encounter any inconsistent labels until later in the task (Experiment 6), as well as whether feedback encourages further compliance with these labels, and reduces the detection of inconsistent trial labels (Experiment 7).

Experiment 5

In this experiment, observers matched pairs of faces that were labelled onscreen as belonging to the "same" person, "different" individuals, or as "unresolved" identity pairings. At the start of the task, observers were informed that most, but not all, of these labels provided consistent information, and so it was important that they provided the final identification decision on each trial. The aim of this first experiment was to determine whether observers' face-matching decisions are biased by external information, such as when face-pair stimuli are labelled as the same

person or different individuals. As in Chapters 2 and 3, the stimuli employed in this chapter were extracted from the KUFD, and thus portrayed considerable within-person variability (see, e.g., Jenkins et al., 2011; Megreya et al., 2013), and mismatches were infrequent (Bindemann et al., 2010; Papesh & Goldinger, 2014). This design should indicate whether human performance in face matching is reduced by inconsistent trial information, as an analogy to human-computer interaction at passport control.

Method

Participants

Thirty undergraduates (11 males, 19 females) with a mean age of 20 years (SD = 3.8) studying at the University of Kent participated in this research in exchange for course credit. All reported normal (or corrected-to-normal) vision. This study was approved by the Ethics Committee of the School of Psychology at the University of Kent, and was conducted in accordance with the ethical guidelines of the British Psychological Association.

Stimuli

Stimuli in this chapter comprised 210 pairs of faces that were extracted from the KUFD. Of these, 15 were mismatching identities, and the remaining 195 were identity matches. One photo in each pair consisted of a controlled image, in which targets were photographed against a plain white background under even lighting and whilst bearing a neutral expression. These photographs were cropped to depict the target's head and shoulders, and were scaled to a size of 283x332 pixels at a resolution of 72-ppi, before being placed on the right-hand side of a plain white canvas. The second image consisted of a student ID photograph that was retrieved with permission

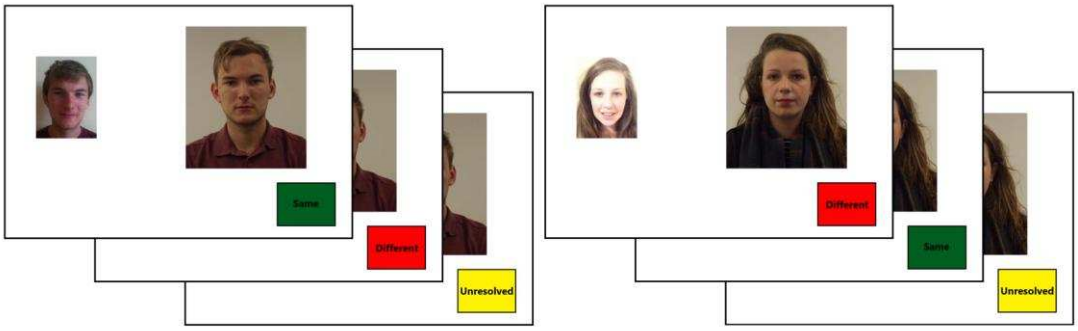

Figure 4.1. An example pair of matching (left) and mismatching (right) identities with consistent, inconsistent, and unresolved identity labels.

from the University of Kent's online Student Data System. These images were unconstrained in target expression, pose, and lighting, and therefore contribute an important source of variability to each stimulus pair. These photographs were scaled to a size of 142x192 pixels at a resolution of 72-ppi, and were presented to the left of the controlled images. Mismatching pairs were created by pairing identities that were visually similar in terms of hair colour, face shape, and eyebrow shape.

Each trial label measured 137x101 pixels and was positioned in the bottom right corner of the screen. These labels were green, red, or yellow, and displayed the message "same", "different", or "unresolved", respectively. These stimuli were counterbalanced over 15 versions of the task, to ensure that each identity was depicted with a consistent, inconsistent, and unresolved label. See Figure 4.1 for an example match and mismatch pair across each label condition.

Procedure

This experiment was run using PsychoPy software (Peirce, 2007). Trials were divided evenly over three blocks of 70 face pairs (65 matches, 5 mismatches), which proceeded without any breaks. At the beginning of the task, observers were instructed

that an identity judgement had already been supplied for each face pair, and that whilst the majority of these would be correct, some would be inaccurate. It was therefore important that observers checked each identity pair carefully before submitting the final decision.

Each trial was preceded by a 1-second fixation cross. This was then replaced with a stimulus pair that was labelled onscreen as "same", "different", or "unresolved". The majority of the trial labels (60%) provided consistent information about the face pair. However, 20% of the labels were also inconsistent, in that they displayed the incorrect solution to the onscreen faces. The remaining 20% of trial labels were unresolved, such that observers were required to independently decide whether two faces depicted the same person or two different individuals. Thus, for the 65 identity matches in a block of 70 trials, 39 were presented with a consistent identification label, 13 with an inconsistent label, and another 13 with an unresolved label. Equally, for the five identity mismatches in each block, three were presented with a consistent identification label, one with an inconsistent label, and another with an unresolved label.

Results

Accuracy

To begin with, mean percentage accuracy scores for consistent, inconsistent, and unresolved match and mismatch trials were analysed. To maximise the number of data points in this analysis, the accuracy data were collapsed across the three blocks of the experiment. Cross-subject means are depicted in Figure 4.2 and reflect that, between consistent and inconsistent labels, accuracy deteriorated by 18% and 22% on match and mismatch trials, respectively. Performance with unresolved trial labels fell

between consistent and inconsistent trials for identity matches, but was more comparable for consistent and unresolved trials for identity mismatches.

To analyse these data formally, a 2 (trial type: match vs. mismatch) x 3 (trial label: consistent, inconsistent, unresolved) within-subjects analysis of variance (ANOVA) was performed. This revealed an effect of trial type, $F(1,29) = 5.12$, $p < 0.001$, $\eta_p^2 = 0.15$, and of trial label, $F(2,58) = 12.07$, $p < 0.001$, $\eta_p^2 = 0.29$, and an interaction between these factors, $F(2,58) = 3.23$, $p < 0.05$, $\eta_p^2 = 0.10$. Bonferroni-adjusted pairwise-comparisons between match and mismatch trials revealed that performance was superior on match trials with consistent, $F(1,29) = 8.47$, $p < 0.01$, $\eta_p^2 = 0.23$, and inconsistent labels, $F(1,29) = 6.30$, $p < 0.05$, $\eta_p^2 = 0.18$. Performance between unresolved match and mismatch trials was comparable, $F(1,29) = 0.63$, $p = 0.43$, $\eta_p^2 = 0.02$. More importantly, simple main effects analysis for the interaction of trial type and trial label revealed that performance on match trials was affected by the trial labels, $F(2,28) = 6.40$, $p < 0.01$, $\eta_p^2 = 0.31$, with higher accuracy on trials with consistent labels as opposed to when the labels were inconsistent or unresolved, both $ps < 0.01$. In addition, accuracy was also worse on inconsistent compared to unresolved match trials, $p < 0.01$. A simple main effect of label type was also found for mismatch trials, $F(2,28) = 6.61$, $p < 0.01$, $\eta_p^2 = 0.32$, with reduced accuracy when labels were inconsistent compared to when these were consistent or unresolved, both $ps < 0.01$. However, performance was comparable between consistent and unresolved mismatch trials, $p = 1.00$.

As an additional step to this analysis, one-sample t-tests were conducted to compare performance on consistent match and mismatch trials to 100%, given that this could be achieved by resolutely following the trial labels. This showed that

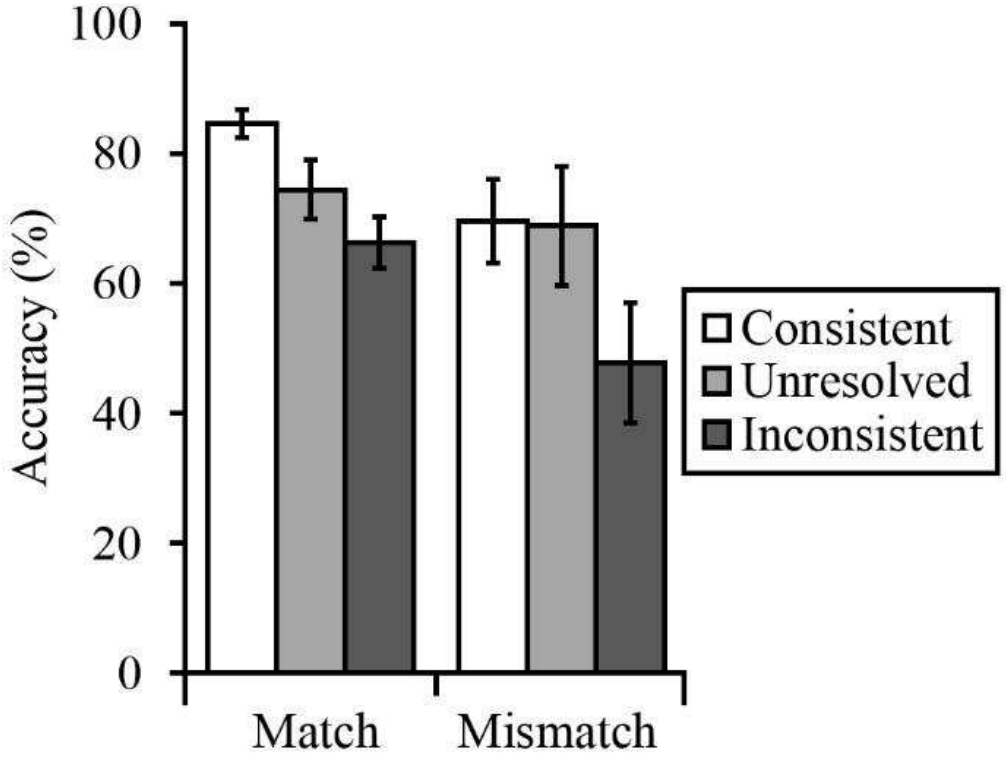

Figure 4.2. Percentage accuracy scores for Experiment 5. Error bars represent the standard error of the mean.

accuracy was significantly below ceiling for both match, $t(29) = -6.93$, $p < 0.001$, and mismatch trials, $t(29) = -8.23$, $p < 0.001$. Next, performance on inconsistent trials was compared to 50%, which represents the point at which the trial labels and the facial information within stimuli influenced observers' decisions equally for these trials. Scores above 50% would reflect that responses were more influenced facial information than by trial labels, whereas the opposite of this would be true for scores below 50%. For match trials, accuracy was significantly above 50%, $t(29) = 3.55$, $p < 0.001$, but on mismatch trials, performance was comparable to 50%, $t(29) = -0.36$, $p = 0.72$.

d-prime and criterion

The percentage accuracy data were also converted into signal detection scores for overall sensitivity (*d'*) and response bias (criterion). For *d'*, a one-way ANOVA revealed an effect of trial label, $F(2,58) = 11.37$, $p < 0.001$, $\eta_p^2 = 0.28$, due to significantly lower *d'* on inconsistent trials, of 0.41, compared to 1.73 and 1.36 on consistent and unresolved trials, respectively, both ps < 0.01. However, sensitivity was comparable between consistent and unresolved trials, $p = 0.35$. The analogous analysis of criterion also revealed an effect of trial label, $F(2,58) = 3.63$, $p < 0.05$, $\eta_p^2 = 0.11$, with criterion shifting from -0.31 on inconsistent trials, to -0.25 and -0.09 on consistent and unresolved trials, respectively. However, none of the pairwise comparisons for this effect were significant following Bonferroni adjustment, all ps ≥ 0.06.

Response times

For completeness, mean correct response times were also analysed for match and mismatch trials for each label category. On match trials, response times increased from 3.21 seconds when trial labels were consistent, to 5.08 and 4.16 seconds when trial labels were inconsistent and unresolved, respectively. Response times on mismatch trials increased from 4.48 seconds when the trial labels were consistent, to 4.93 and 5.47 seconds when the labels were inconsistent and unresolved, respectively. These data reflect that responses were quickest when trial labels were consistent, but took longer when labels were misleading or did not resolve a given trial. However, a 2 (trial type) x 3 (trial label) within-subjects ANOVA did not reveal an effect of trial type, $F(1,22) = 3.91$, $p = 0.06$, $\eta_p^2 = 0.15$, or an effect of trial label, $F(2,44) = 1.48$, $p = 0.24$, $\eta_p^2 = 0.06$, and these factors did not interact, $F(2,44) = 0.17$, $p = 0.84$, $\eta_p^2 = 0.01$.

Discussion

In this experiment, observers matched faces that were labelled onscreen as depicting same or different identities or as unresolved identifications. Performance was considerably more accurate when these labels provided information that was consistent with the identities of the depicted face pairs. For example, accuracy on mismatch trials deteriorated from 70% when these were labelled as depicting different identities to 48% when these faces were labelled as belonging to the same person. Similarly, performance on consistently-labelled match trials deteriorated from 85% to 66% when the labels indicated that the faces depicted different individuals. For trials that were labelled as unresolved, accuracy was similar between match and mismatch trials, at 74% and 69%, respectively. These effects were corroborated by the analysis of d', which showed that errors increased considerably when to-be-matched faces were inconsistently-labelled.

Together, these findings indicate that observers' face-matching decisions are biased by a priori external identity judgements, such as same- and different-identity labels. This converges with recent work showing that observers' face-matching decisions can be compromised when led to believe that two faces depict the same person (Menon, White, & Kemp, 2015b), as well as research demonstrating that the concurrent presentation of biographical information alongside face-pair stimuli biases observers towards erroneous match responses (McCaffery & Burton, 2016). In contrast to these studies, the current experiment observed such biasing effects with a paradigm designed to mimic human-computer interaction at passport control.

Although the observed interaction shows that the trial labels influenced responses in this task, comparing accuracy on inconsistently-labelled trials to 50%

reflected that observers' decisions were still largely influenced by the facial information within stimuli. Similarly, performance on trials for which the labels provided consistent information was below 100%, indicating that observers were reluctant to adhere fully to the trial labels. This makes it difficult to apply these findings to human-computer interaction at passport control, where e-Gates are expected to function with high accuracy (FRONTEX, 2015a). Consequently, human operators are likely to be more trusting of the algorithms' decisions in such settings. To encapsulate this, a second experiment was conducted, which sought to encourage compliance with the trial labels by replacing all inconsistent labels in Block 1 to provide consistent information, thereby making it possible to achieve 100% accuracy in this block by resolutely following the trial labels.

Experiment 6

The previous experiment shows that accuracy deteriorated by around 20% on match and mismatch trials for which the labels provided inconsistent identification information. However, observers also rejected nearly a quarter of labels that actually displayed consistent information. One explanation for this could be that encountering misleading labels at the beginning of the task may have discouraged observers from trusting the information that was provided by the trial labels. To investigate this possibility, all inconsistent labels in the first block of Experiment 6 were replaced to provide consistent information. The aim of this new design was to encourage observers to trust the trial labels at the beginning of the task. If this manipulation is successful, then performance on consistent labels in Block 1 should be very high, given that these labels provide the correct solution with 100% accuracy. This high rate of compliance is expected to coincide to produce high accuracy in Blocks 2 and 3, on trials for which

the trial labels provide consistent information. If so, however, then this should also coincide with considerably worse accuracy on inconsistently-labelled trials in Blocks 2 and 3. Moreover, this effect might be particularly pronounced on mismatch trials, given that these occur less frequently than inconsistent match trials.

Method

Participants, stimuli, and procedure

Thirty new participants from the University of Kent (5 males, 25 females) with a mean age of 19.2 years (SD = 1.2) participated in this experiment in exchange for course credit or a small fee. All participants reported normal (or corrected-to-normal) vision, and none had participated in Experiment 5.

The stimuli and procedure in this experiment were identical to the previous experiment, except that all labels in Block 1 that provided inconsistent information were replaced to now be consistent. The number of unresolved trials was unchanged, and Blocks 2 and 3 were identical to the second and third blocks in Experiment 5.

Results

Accuracy

Again, mean percentage accuracy scores were calculated for match and mismatch trials according to whether labels provided consistent, inconsistent, or unresolved information. Because Block 1 did not feature any inconsistent trial labels, performance in this block was analysed separately first. Cross-subject means for consistent and unresolved match and mismatch trials can be found in Figure 4.3. These data were analysed using a 2 (trial type: match vs. mismatch) x 2 (trial label: consistent vs. unresolved) within-subjects ANOVA, which did not reveal an effect of trial type,

$F(1,29) = 1.85$, $p = 0.19$, $\eta_p^2 = 0.06$, but of trial label, $F(1,29) = 6.19$, $p < 0.05$, $\eta_p^2 = 0.18$, due to higher accuracy on consistently-labelled trials. The interaction was not significant, $F(1,29) = 0.36$, $p = 0.55$, $\eta_p^2 = 0.01$. These accuracy data were also compared to 100% using one-sample t-tests. Accuracy was significantly below 100% on both match, $t(29) = -8.15$, $p < 0.001$, and mismatch trials, $t(29) = 5.21$, $p < 0.001$.

Next, a cross-experimental comparison was conducted on the percentage accuracy data for consistent trials only from Block 1 of Experiments 5 and 6. This analysis should reveal whether replacing inconsistent trial labels with consistent labels in the first block of Experiment 6 resulted in higher accuracy. However, a 2 (trial type) x 2 (experiment: 5 vs. 6) mixed-factor ANOVA did not reveal an effect of experiment, $F(1,58) = 0.51$, $p = 0.48$, $\eta_p^2 = 0.01$, or an interaction of experiment and trial type, $F(1,58) = 0.37$, $p = 0.55$, $\eta_p^2 = 0.01$.

The data of main interest were how observers performed on trials for which the labels provided consistent, inconsistent, and unresolved information. These scores were collapsed across Blocks 2 and 3, and are also depicted in Figure 4.3. These data show a decline in accuracy on consistent match trials, from 85% to 65%, when these were labelled as different identities. In addition, accuracy on mismatch trials deteriorated from 65% when the labels provided consistent information, to 40% when these were labelled inconsistently. Finally, accuracy on unresolved match and mismatch trials was 73% and 60%, respectively. A 2 (trial type) x 3 (trial label) within-subjects ANOVA revealed significantly higher accuracy on match trials compared to mismatch trials, $F(1,29) = 22.95$, $p < 0.001$, $\eta_p^2 = 0.44$. An effect of trial label was also found, $F(2,58) = 8.59$, $p < 0.01$, $\eta_p^2 = 0.23$. Bonferroni-adjusted pairwise comparisons revealed that this was due to worse accuracy on inconsistent trials compared to consistent, $p < 0.01$, and unresolved trials, $p < 0.05$. However,

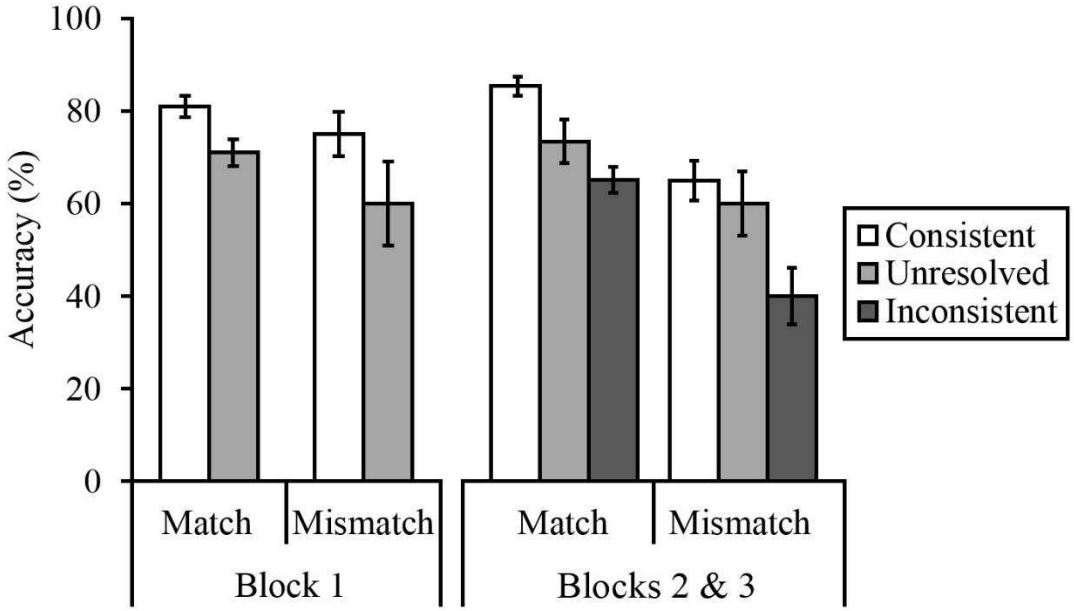

Figure 4.3. Percentage accuracy scores for Experiment 6. Error bars represent the standard error of the mean.

performance was comparable between consistent and unresolved trials, $p = 0.15$. The interaction between trial type and trial label was not significant, $F(2,58) = 1.25$, $p = 0.29$, $\eta_p^2 = 0.04$.

A cross-experimental comparison was performed on the percentage accuracy data collapsed across Blocks 2 and 3 of Experiments 5 and 6. A 2 (experiment) x 2 (trial type) x 3 (trial label) mixed-factor ANOVA did not reveal an effect of experiment, $F(1,58) = 0.80$, $p = 0.38$, $\eta_p^2 = 0.01$, which did not interact with trial type, $F(1,58) = 0.02$, $p = 0.90$, $\eta_p^2 = 0.00$, or with trial label, $F(2,116) = 0.22$, $p = 0.80$, $\eta_p^2 = 0.00$. The three-way interaction was also not significant, $F(2,116) = 0.03$, $p = 0.98$, $\eta_p^2 = 0.00$.

Finally, one-sample t-tests showed that accuracy on consistently-labelled match, $t(29) = -6.93$, $p < 0.001$, and mismatch trials, $t(29) = -8.23$, $p < 0.001$, was

significantly below 100%, reflecting that observers were not resolutely following the trial labels. For inconsistently-labelled trials, accuracy on match trials was above 50%, $t(29) = 3.23$, $p < 0.01$. However, accuracy on mismatch trials was comparable to 50%, $t(29) = -1.44$, $p = 0.16$. These findings suggest that for match trials, the facial information in stimuli exerted a greater influence on observers' decisions than the trial labels, but these factors affected identity judgements comparably on mismatch trials.

d-prime and criterion

For completeness, *d'* and criterion scores were also analysed. In Block 1, a paired-sample t-test revealed that *d'* was superior on consistently-labelled trials, at 1.76, compared to on unresolved trials, for which *d'* was 0.95, $t(29) = 2.4$, $p < 0.05$. Criterion scores were near-identical for consistent and unresolved trials, at -0.11 and -0.15, respectively, and were statistically comparable, $t(29) = 0.28$, $p = 0.78$.

Next, *d'* and criterion for Blocks 2 and 3 were analysed. Collapsed across these blocks, *d'* was worse on inconsistent trials, at 0.11, compared to 1.68 on consistent trials. For unresolved trials, *d'* was 1.03. A one-way ANOVA revealed these differences between trial labels to be significant, $F(2,58) = 8.67$, $p < 0.01$, $\eta_p^2 = 0.23$, with reduced sensitivity when trial labels were inconsistent, both $ps < 0.05$. However, *d'* was comparable between consistent and unresolved trial labels, $p = 0.12$. The analogous analysis of criterion did not reveal an effect of trial label, $F(2,58) = 1.90$, $p = 0.16$, $\eta_p^2 = 0.06$, due to similar criterion scores of -0.39 on consistent and inconsistent labels, and -0.18 when trial labels were unresolved.

Response times

Finally, mean correct response times for Experiment 6 were analysed. In Block 1, a 2 (trial type) x 2 (trial label: consistent vs. unresolved) within-subjects ANOVA did not find an effect of trial label, $F(1,16) = 1.97$, $p = 0.18$, $\eta_p^2 = 0.11$, but revealed longer response times of 5.42 seconds on mismatch trials, compared to 4.13 seconds on match trials, $F(1,16) = 5.35$, $p < 0.05$, $\eta_p^2 = 0.25$. These factors did not interact, $F(1,29) = 2.17$, $p = 0.15$, $\eta_p^2 = 0.07$.

Collapsed across Blocks 2 and 3, a 2 (trial type) x 3 (trial label) within-subjects ANOVA also did not reveal an effect of trial label, $F(2,30) = 0.70$, $p = 0.50$, $\eta_p^2 = 0.05$, but again of trial type, $F(1,15) = 6.08$, $p < 0.05$, $\eta_p^2 = 0.29$, which was also due to slower response times of 4.28 seconds on mismatch trials, compared to 3.05 seconds on match trials. Again, the interaction was not significant, $F(2,30) = 1.64$, $p = 0.21$, $\eta_p^2 = 0.10$.

Discussion

As in Experiment 5, this experiment showed that providing inconsistent information through trial labels reduced identification performance. Across Blocks 2 and 3, this effect accounted for 23% more errors on inconsistent trials compared to when the labels were consistent, and 14% more errors than on unresolved trials. Again, *d'* was reduced on trials for which the labels provided inconsistent information, but was comparable for consistent and unresolved trials. Considered together, these findings converge with those of Experiment 5, and replicate the detrimental effects of inconsistent trial labels.

On trials for which the labels provided inconsistent identity information, performance exceeded 50% on match trials, but was comparable to this level on mismatch trials. In addition, accuracy was below ceiling on consistently-labelled

match and mismatch trials in Block 1, and across Blocks 2 and 3. These findings indicate that again, observers were generally reluctant to completely trust the onscreen labels on a high number of trials, but reached an identification decision based on the facial information of the stimuli. This is further supported by the cross-experiment comparison, which indicated that observers were not more trusting of the trial labels in Block 1, despite the absence of inconsistent trial labels in this block of Experiment 6, and made a similar number of errors on inconsistently-labelled trials in Blocks 2 and 3. In a further attempt to persuade observers to trust the trial labels, a third experiment was conducted, in which all inconsistent labels in Blocks 1 and 2 were replaced to provide consistent information, and feedback on accuracy was administered in the first block of the task. The aim of this manipulation was to provide observers with a compelling reason to trust the trial labels. It was then assessed in Block 3 whether this further exacerbated accuracy when the trial labels were inconsistent.

Experiment 7

Experiments 5 and 6 show that face matching is more difficult if an incorrect solution is presented onscreen. However, both experiments also featured a surprisingly high error rate on trials that could be accurately resolved by following the outcome supplied by the labels. A cross-experiment comparison suggests that performance on inconsistent trials was not reduced to a greater extent when the labels in Block 1 predicted the correct answer with 100% accuracy in Experiment 6. In addition, performance was significantly below 100% for match and mismatch trials for which the trial labels provided consistent information. Together, these findings indicate that observers were still reluctant to trust the information provided by these labels. It is

possible, however, that a single block of trials provided insufficient time for observers to learn to resolutely rely on these labels.

To explore this further, all inconsistent labels in Blocks 1 and 2 were replaced to provide consistent information in Experiment 7. In addition, trial-by-trial feedback was administered in the first block whilst stimuli were still onscreen, to encourage compliance with the labels. Other research has shown that face-matching performance benefits reliably from feedback (Alenezi & Bindemann, 2013; White, Burton, et al., 2014), indicating that observers refine their strategy for comparing faces when able to monitor their performance within a session. Here, it is expected that observers become more compliant with trial labels over the course of Block 1, as they receive feedback that aligns with the trial labels. If this feedback manipulation is successful in encouraging observers to follow the trial labels, then accuracy should also be high in Block 2, given that this block also did not feature any inconsistent trial labels. In the final block, this should coincide with high accuracy on trials for which the trial labels are consistent, but result in an even greater number of errors on inconsistent trials. Moreover, it is also expected that these errors will be exaggerated on inconsistent mismatch trials, given that these occurred less frequently than inconsistent match trials.

Method

Participants, stimuli, and procedure

Thirty undergraduates studying at the University of Kent (8 males, 22 females) with a mean age of 19.6 years (SD = 1.8) participated in this study in exchange for course credit or a small fee. None of these had participated in the previous experiments, and all reported normal (or corrected-to-normal) vision.

The stimuli and procedure used in this experiment were identical to that of the previous experiment, except for the following changes. All inconsistent labels in Blocks 1 and 2 were replaced to provide consistent information, whilst the frequency of unresolved match and mismatch trials remained unchanged. In addition, onscreen feedback was provided following each response in Block 1, whilst the label and stimuli were still onscreen, and consisted of "Correct/Incorrect! These faces show the SAME person/two DIFFERENT individuals!". This feedback was withdrawn in Block 2, and Block 3 was identical to the third block in Experiments 5 and 6.

Results

Accuracy

The percentage accuracy scores for Blocks 1, 2 and 3 are shown in Figure 4.4. First, performance between Block 1 of Experiment 6 and Block 1 of Experiment 7 was compared on consistently-labelled trials only, to investigate whether feedback facilitated greater trust in the trial labels. A 2 (trial type) x 2 (experiment) mixed-factor ANOVA did not find an effect of experiment, $F(1,58) = 0.03$, $p = 0.87$, $\eta_p^2 = 0.00$, but did reveal an interaction, $F(1,58) = 8.16$, $p < 0.01$, $\eta_p^2 = 0.12$. Simple main effects analysis revealed that this was due to higher accuracy on match trials in Experiment 7, compared to Experiment 6, $F(1,58) = 18.94$, $p < 0.001$, $\eta_p^2 = 0.25$. By contrast, mismatch accuracy was comparable between experiments, $F(1,58) = 1.90$, $p = 0.17$, $\eta_p^2 = 0.03$. A simple main effect of trial type was also found in Experiment 7, $F(1,58) = 26.75$, $p < 0.001$, $\eta_p^2 = 0.32$, due to higher accuracy on match than mismatch trials, but not in Experiment 6, $F(1,58) = 1.28$, $p = 0.26$, $\eta_p^2 = 0.02$. This indicates that the feedback increased accuracy on the most frequent consistent trial – the identity matches.

Next, it was assessed whether these cross-experiment gains were maintained in Experiment 7, from Block 1 to Block 2. A 2 (trial type: match vs. mismatch) x 2 (trial label: consistent vs. unresolved) x 2 (block: 1 vs. 2) within-subjects ANOVA did not reveal an effect of block, $F(1,29) = 2.17$, $p = 0.15$, $\eta_p^2 = 0.07$, or an interaction of block with trial label, $F(1,29) = 0.08$, $p = 0.77$, $\eta_p^2 = 0.00$, or with trial type, $F(1,29) = 0.67$, $p = 0.42$, $\eta_p^2 = 0.02$. The three-way interaction was also not significant, $F(1,29) = 0.04$, $p = 0.84$, $\eta_p^2 = 0.00$. This analysis indicates that the feedback gains for consistent match trials were maintained in Block 2. Despite these feedback gains, accuracy on consistent match trials was below 100% in Block 1, $t(29) = -6.93$, $p < 0.001$, and in Block 2, $t(29) = -4.93$, $p < 0.001$. Similarly, mismatch accuracy was also below ceiling in the first, $t(29) = -6.43$, $p < 0.001$, and second block, $t(29) = -5.96$, $p < 0.001$.

The data of main interest concerned the extent to which observers were able to detect misleading trial labels in Block 3. The data depicted in Figure 4.4 reflect that on match trials, accuracy deteriorated from 94% to 62% between consistent and inconsistent labels, respectively, and from 70% to 23% on mismatch trials. Accuracy on unresolved match and mismatch trials was 85% and 37%, respectively. A 2 (trial type) x 3 (trial label) within-subjects ANOVA revealed that accuracy on match trials was significantly greater than on mismatch trials, $F(1,29) = 47.00$, $p < 0.001$, $\eta_p^2 = 0.62$. In addition, an effect of trial label was found, $F(2,58) = 13.15$, $p < 0.001$, $\eta_p^2 = 0.31$. Bonferroni-adjusted pairwise comparisons revealed that this was due to worse accuracy on trials that were labelled inconsistently, versus trials for which the labels provided consistent, $p < 0.001$, and unresolved information, $p < 0.01$. These factors did not interact, $F(2,58) = 1.63$, $p = 0.21$, $\eta_p^2 = 0.05$.

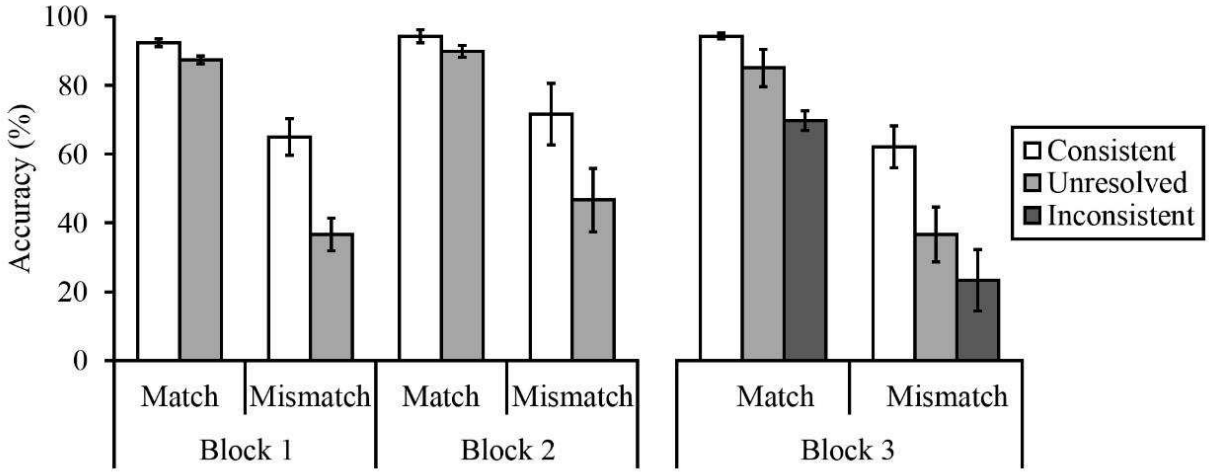

Figure 4.4. Percentage accuracy scores for Experiment 7. Error bars represent the standard error of the mean.

As in the previous experiment, a cross-experimental comparison was performed between Block 3 of this experiment and the final block of Experiment 6, to assess whether the increased compliance with trial labels exacerbated accuracy on inconsistent trials. The 2 (experiment) x 2 (trial type) x 3 (trial label) mixed-factor ANOVA did not reveal an effect of experiment, $F(1,58) = 0.35$, $p = 0.57$, $\eta_p^2 = 0.01$, which did not interact with trial label, $F(2,116) = 0.59$, $p = 0.56$, $\eta_p^2 = 0.01$. The interaction between experiment and trial type was approaching significance, $F(1,58) = 4.03$, $p = 0.05$, $\eta_p^2 = 0.07$, with higher accuracy on match compared to mismatch trials, in Experiment 6, $F(1,58) = 17.16$, $p < 0.001$, $\eta_p^2 = 0.23$, and in Experiment 7, $F(1,58) = 48.76$, $p < 0.001$, $\eta_p^2 = 0.46$. However, accuracy was comparable between experiments for match, $F(1,58) = 3.59$, $p = 0.06$, $\eta_p^2 = 0.06$, and mismatch trials, $F(1,58) = 2.28$, $p = 0.14$, $\eta_p^2 = 0.04$. The three-way interaction was not significant, $F(2,116) = 1.20$, $p = 0.34$, $\eta_p^2 = 0.02$.

Finally, comparing accuracy on trials for which labels provided consistent information revealed that accuracy was significantly below 100% on match, $t(29) = -$

6.36, $p < 0.001$, and mismatch trials, $t(29) = -6.13$, $p < 0.001$. Similarly, accuracy on match trials exceeded 50% for inconsistently-labelled stimuli, $t(29) = 3.66$, $p < 0.001$. This indicates, once again, that observers did not adhere to the information provided by the trial labels completely. However, performance was significantly below this threshold for inconsistently-labelled mismatch trials, $t(29) = -3.40$, $p < 0.01$. This suggests that observers were more likely to base the identification decisions on the trial labels than the facial information in this condition.

d-prime and criterion

Next, *d'* and criterion scores were analysed. Sensitivity increased slightly on consistent labels between Blocks 1 and 2, from 1.94 to 2.24, respectively, as well as on unresolved trials, from 0.75 to 1.16. A 2 (trial label) x 2 (block) within-subjects ANOVA did not reveal an effect of block, however, $F(1,29) = 2.20$, $p = 0.15$, $\eta_p^2 = 0.07$, but of trial label, $F(1,29) = 19.39$, $p < 0.001$, $\eta_p^2 = 0.40$, due to superior performance on consistent trials. The interaction was not significant, $F(1,29) = 0.07$, $p = 0.80$, $\eta_p^2 = 0.00$. Criterion scores appeared comparable between Blocks 1 and 2 on consistent trial labels, with -0.50 and -0.45, respectively, as well as on unresolved trials, with -0.81 and -0.69, respectively. The 2 (trial label) x 2 (block) within-subjects ANOVA did not find an effect of trial label, $F(1,29) = 0.52$, $p = 0.48$, $\eta_p^2 = 0.02$, but of block, $F(1,29) = 5.21$, $p < 0.05$, $\eta_p^2 = 0.15$, due to a greater number of match responses in Block 2 compared to Block 1. These factors did not interact, $F(1,29) = 0.09$, $p = 0.77$, $\eta_p^2 = 0.00$.

In Block 3, *d'* was at 1.95 when trial labels were consistent, but deteriorated to 0.68 and -0.26 on unresolved and inconsistent trial labels. A one-way ANOVA revealed that this trial label effect was reliable, $F(2,58) = 15.23$, $p < 0.001$, $\eta_p^2 = 0.29$,

with Bonferroni-adjusted pairwise comparisons showing that sensitivity was significantly higher on consistent trial labels compared to when these were inconsistent, $p < 0.001$, and unresolved, $p < 0.01$. However, *d'* was comparable between inconsistent and unresolved trials, $p = 0.11$. The analogous analysis of criterion for Block 3 did not reveal an effect of trial label, $F(2,58) = 0.68$, $p = 0.51$, $\eta_p^2 = 0.02$, with similar criterion scores of -0.75 and -0.78 on inconsistent and unresolved trials, respectively, and -0.60 on consistent trial labels.

Response times

Finally, the mean correct response time data were explored. Due to an insufficient number of data points on unresolved trials between Blocks 1 and 2, only performance on consistent match and mismatch trials was analysed. For the first two blocks, a 2 (trial type) x 2 (block) within-subjects ANOVA did not reveal an effect of block, $F(1,26) = 0.10$, $p = 0.75$, $\eta_p^2 = 0.00$, but of trial type, $F(1,26) = 29.48$, $p < 0.001$, $\eta_p^2 = 0.53$, due to slower response times of 4.23 seconds on mismatch trials, compared to 2.59 seconds on match trials. The interaction was not significant, $F(1,26) = 0.18$, $p = 0.67$, $\eta_p^2 = 0.01$. The final block also yielded insufficient data points for analysis, due to the low accuracy on inconsistent mismatch trials.

Discussion

This experiment provides further evidence that face matching is biased by trial labels. The comparison between the first block of Experiments 6 and 7 showed that accuracy on consistent match trials was enhanced in Experiment 7, indicating that the feedback encouraged observers to follow the trial labels. Accuracy was also similar between Blocks 1 and 2, suggesting that observers remained compliant with the labels

after the feedback was withdrawn. Although this did not result in a greater number of errors in the final block when compared with Experiment 6, a comparison with the level of accuracy that one might expect if trial labels and facial information exert equal influence on decision-making (i.e., 50%), indicates that responses on inconsistently-labelled mismatch trials were influenced to a greater extent by the trial labels. This suggests that the administration of feedback increased observers' reliance on the trial labels. This is an important finding considering this effect was observed with inconsistent mismatch trials, which were misleadingly labelled as identity matches. If a similar effect exists with e-Gates at border control, then human-computer interactions would lead to increased failure to detect the persons of most interest – the criminal identity impostors.

General Discussion

This study investigated face-matching accuracy whilst onscreen trial labels provided consistent, inconsistent, or unresolved information about to-be-matched faces. Observers were informed that most of these labels supplied the correct response, but that some would also be inaccurate as well as unresolved, and so they were required to provide the final decision on each pair. In each experiment, the trial labels impacted performance, with accuracy deteriorating considerably between consistent and inconsistent trial labels. In Experiment 5, this effect accounted for 18% and 22% more errors on match and mismatch trials, respectively. However, even when the trial labels were consistent with the trial type, accuracy was at 85% on match trials, and 70% on mismatch trials, indicating that observers were reluctant to trust the trial labels even though these provided the correct solution.

We attempted to encourage reliance on these labels in Experiment 6 by replacing all inconsistent trial labels in Block 1 to provide consistent information. However, observers remained reluctant to follow the labels in this block, rejecting 19% of consistent match trials, and 25% of mismatch trials. In addition, a similar numerical effect of trial labels was observed in Experiment 6, with accuracy depleting by 20% between consistent and inconsistent match trials, and by 25% between consistent and inconsistent mismatch trials.

In the final experiment, observers were provided with trial-by-trial feedback in Block 1, and did not encounter any inconsistent trial labels until Block 3. Compared to Experiment 6, this manipulation improved performance on consistent trials in Block 1. However, this did not result in a significantly greater number of errors in Block 3, in which accuracy deteriorated between consistent and inconsistent trial labels by 25% and 39% on match and mismatch trials, respectively.

It is perhaps surprising that the provision of trial-by-trial feedback at the beginning of Experiment 7 improved performance on consistent trials without concurrently inducing a greater number of errors on inconsistent trials, compared to Experiment 6. However, in Experiment 7 alone, accuracy on inconsistent mismatch trials was also significantly lower than 50%, suggesting that the trial labels exerted a stronger influence on observers' decisions than the facial information in stimuli. Together, these results suggest that observers' face-matching decisions are influenced by trial labels to a greater extent than by facial information when provided with compelling reasons to trust these judgements.

This trial label effect was also consistently reflected by *d'*, which provided further evidence that performance was reduced on inconsistent trials. However, such an effect was not observed for response times, which remained comparable between

consistent and inconsistent trial labels in Experiments 5 and 6. This implies that face stimuli did not undergo additional processing when these were labelled inconsistently, which could be interpreted as further evidence for the difficulty of detecting inconsistent trial labels.

Similar to White, Dunn, et al. (2015), who found that human performance in face matching curtailed the accuracy of algorithms when processing passport applications, the current study suggests that human-computer interaction at passport control, where human operators supervise e-Gates, is also error-prone. The reported experiments indicate that the commission of errors by algorithms facilitate errors in humans, given that observers were more likely to accept a mismatch, and reject an identity match, if these were labelled as the depicting same person or different individuals, respectively. This finding converges with evidence that facial identification processes are guided by information from trustworthy sources, such as experimenters, even when inaccurate (see, e.g., Johansson, Hall, Sikström, & Olsson, 2005; Menon et al., 2015b; Sagana, Sauerland, & Merckelbach, 2016; Sauerland et al., 2016). In addition, human operators are typically expected to monitor up to seven e-Gates concurrently (FRONTEX, 2015a). This raises further concerns when considering that in laboratory settings, face matching suffers considerably when observers are expected to process more than one concurrent identity (see, Megreya & Burton, 2006b; Bindemann, Sandford, Gillatt, & Avetisyan, 2012). As a consequence, it is possible that the task of human operators is substantially more challenging still than the current results suggest.

In addition to the high error rate on inconsistent trials, wherein observers incorrectly followed the trial labels, many errors also emerged on consistent trials, whereby observers incorrectly overruled the labels. This reluctance to trust the labels

is perhaps surprising, given that these generally provided consistent information. However, it is possible that observers were more reliant on the labels when the outcome of a trial was uncertain. This explanation fits with studies in which observers display an attentional bias for faces (Bindemann, Jenkins, & Burton, 2005; Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005), except when it is advantageous to attend to task-relevant non-face stimuli (Bindemann, Burton, Langton, Schweinberger, & Doherty, 2007). Given that the relevance of the trial labels in this study was dependent on whether these provided consistent or inconsistent information, it is possible that to-be-matched faces were the primary focus of observers' attention, except when the correct resolution was unclear. Ideally, this represents how the task should be performed in operational settings, with human operators reaching an independent identification decision that typically converges with the e-Gate verdict if correct, but otherwise overrules the system's resolution. The current results indicate that it is particularly difficult to avoid being influenced by trial labels.

Across the experiments reported in this chapter, performance on unresolved trials ranged from 73-88% on match trials, and 40-69% on mismatch trials. This resonates with the consistent finding that face matching is challenging when an a priori judgement is not provided. This raises additional concerns surrounding the identification accuracy of human operators of e-Gates when the system cannot adequately resolve a person with their passport photograph (FRONTEX, 2015a). However, accuracy on these unresolved trials was generally superior to when the trial labels provided inconsistent information, reflecting that it is more challenging to overrule an incorrect identity judgement than to make a correct identification independently.

In sum, the experiments presented in this chapter show that it is particularly difficult to accurately match faces when confronted with misleading identity information. Specifically, the reported experiments suggest that the commission of errors by automated systems are likely to undermine the performance of human observers, such as when an impostor is incorrectly labelled as an identity match. This has implications for human-computer interaction at passport control, where human operators verify the decisions of e-Gates. The present results indicate that humans are unreliable at safeguarding against the errors of such systems.

Chapter 5

Summary, Discussion, and Future Directions

This thesis investigated performance in forensic face matching under conditions that aimed to encapsulate some of the challenges associated with comparing travellers to passport photographs in applied settings. The first chapter provided a systematic review of face-matching research to date. This research has consistently shown that face matching is remarkably error-prone, and suffers under a number of data-limiting conditions, such as when to-be-matched faces differ in terms of lighting (Hill & Bruce, 1996; Jenkins et al., 2011; Liu et al., 2013) and pose (Estudillo & Bindemann, 2014; Hill & Bruce, 1996). In addition, accuracy deteriorates when the time interval between two face photographs increases (Jenkins et al., 2011; Megreya et al., 2013), as well as when external facial features are occluded (Bruce et al., 1999; Estudillo & Bindemann, 2014; Kemp et al., 2016; Megreya & Bindemann, 2009).

However, there is also evidence that face matching is largely dependent on the resource capacity of observers who complete this task, given that accuracy within groups of observers also deteriorates within a single prolonged session (Alenezi & Bindemann, 2013; Alenezi et al., 2015) and under time pressure (Bindemann et al., 2016). In addition, considerable differences in performance arise between observers even when data limitations are minimised (see, e.g., Bindemann, Avetisyan, et al., 2012; Burton et al., 2010; Estudillo & Bindemann, 2014; Megreya & Bindemann, 2013; White, Kemp, Jenkins, Matheson, et al., 2014; White et al., 2017). Some

observers also consistently exhibit a higher capacity for matching faces than others (Bobak, Dowsett, et al., 2016; Davis et al., 2016; White, Phillips, et al., 2015), even under high data limitations (Robertson et al., 2016). By contrast, some individuals exhibit sub-average performance even under optimised conditions (e.g., Burton et al., 2010; Estudillo & Bindemann, 2014). Together, this research suggests that data limitations imposed by stimuli in face matching are moderated by observers' resource capacity for performing this task, given that some individuals can accurately match faces even under impoverished viewing conditions (see Robertson et al., 2016).

Much of this research has been conducted under the ideal conditions that are provided by the GFMT (see Burton et al., 2010). This test comprises high-quality face images that were taken under even lighting, whilst bearing a similar expression and pose. Crucially, identity matches in the GFMT were photographed only minutes apart, but with different image capture devices, which contribute the primary source of variation between these face images (see, e.g., Burton, 2013; Noyes & Jenkins, 2017). The optimised nature of these stimuli makes it possible to isolate the effects of factors such as low mismatch frequency (Bindemann et al., 2010), time pressure (Bindemann et al., 2016), and time passage (Alenezi & Bindemann, 2013; Alenezi et al., 2015) on face matching, without being compounded by additional factors, such as within-person variability in the appearance of the depicted targets (Jenkins et al., 2011; Megreya et al., 2013). However, such variability is of practical importance when considering that passports typically remain valid through a 10-year period. In addition, some recent findings have suggested that due to typically high accuracy rates of 80%, the GFMT might lack the sensitivity to fully assess the impact of some factors on task performance (see, e.g., Bobak, Dowsett, et al., 2016; Dowsett & Burton, 2015). The optimised conditions provided by the GFMT may therefore be limited in their capacity

to estimate how some factors might impact face-matching performance at passport control, given that these stimuli do not fully reflect the difficulty of this task in applied settings (see Burton, 2013; see also Young & Burton, 2017).

The purpose of Chapter 2 was to address this limitation via the KFMT. This test aims to provide a more realistically challenging stimulus database for face-matching research, by creating identity pairs that were photographed many months apart, and which exhibit natural variation in expression, pose, and lighting. Experiment 1 measured performance in a short version of the KFMT comprising 40 items, and compared accuracy in this test with the analogous short version of the GFMT. Overall accuracy in the KFMT was 66%, which was reflected for both match and mismatch trials. By contrast, overall accuracy in the GFMT was 80%, with observers scoring 82% on match trials, and 78% on mismatch trials. These scores converge with normative test data on the GFMT (see Burton et al., 2010), but demonstrate the greater difficulty of the KFMT. Importantly, performance in these tests was strongly correlated, reflecting that the KFMT and GFMT measure similar processes, but differ in terms of difficulty. The greater difficulty of the KFMT was further reflected at the level of individual test items, as well as for the majority of observers. In addition, the KFMT exhibited high test-retest reliability, with a strong positive correlation within observers who completed this test twice, following an interval of one week. Together, these findings reflect that the short version of the KFMT is a reliable test of face matching that provides a more difficult test than the GFMT, whilst measuring similar processes.

These findings were corroborated in Experiment 2, which presented a longer version of the KFMT comprising 200 match trials and 20 mismatch trials. Overall accuracy in this task was 78% and 64% on match and mismatch trials, respectively.

However, mismatch accuracy deteriorated from 74% to 57% from the first to the final block of trials. This was driven by a match response bias that emerged over time, whereby observers erroneously classified more faces as identity matches as they progressed through the task. This bias has also been observed in research using stimuli from the GFMT (see Alenezi & Bindemann, 2013; Alenezi et al., 2015), and strongly suggests that the KFMT and GFMT share similar behavioural characteristics. In addition, performance correlated with the CFMT and CFPT, which provide established measures of face memory and face perception, respectively (see Duchaine & Nakayama, 2006; Duchaine et al., 2007). These correlations reflect that the KFMT taps into similar processes to those employed for identifying and processing faces. Moreover, these results converge with other work in which face-matching performance correlates with other measures of face memory (Burton et al., 2010; Megreya & Burton, 2006a). In addition, performance in the CFMT and CFPT appear to predict performance in the GFMT (Bobak, Hancock, et al., 2016; Robertson et al., 2016; White et al., 2017). As a consequence, these relationships provide further evidence that the KFMT is a reliable measure of face matching.

Considered together, the findings from Chapter 2 reflect that the KFMT comprises a psychometrically-stable measure of face matching, by providing a more challenging test whilst measuring similar processes to the GFMT. This is consistent with research demonstrating that face matching is more error-prone when to-be-matched stimuli are photographed months apart (Megreya et al., 2013), or are depicted in ambient settings (Jenkins et al., 2011). The KFMT is not intended as a replacement for any existing measures of face matching (for an overview of current measures, see Noyes & O'Toole, 2017). Rather, the purpose of this test is to facilitate further research with the aim of understanding how some factors might impact accuracy in applied

settings, where performance is already compounded by within-target variability and high trial numbers. The potential utility of such a resource is reflected in research showing that, due to ceiling-level performance, optimised measures of face matching may lack the sensitivity to fully explore some effects that emerge under more challenging conditions (Bobak, Dowsett, et al., 2016; Dowsett & Burton, 2015; Kemp et al., 2016). For example, recent work has shown that whilst super-recognisers outperform student observers on the GFMT, these individuals perform comparably to police identifiers who are not super-recognisers, suggesting that the GFMT lacks the sensitivity to detect these important individual differences in ability (Davis et al., 2016). Additional research has shown also that within-target variability interacts with other factors such as own-race biases to exacerbate performance further (Meissner et al., 2013). This suggests that the detrimental effects of some factors are exaggerated under more challenging conditions.

Utilising the more challenging conditions provided by the KFMT, Chapter 3 then investigated the concurrent effects of time pressure and time passage on face-matching performance. Time pressure is of practical relevance to applied settings, given that passport officers must process high passenger numbers within set time targets that are frequently missed (Home Affairs Committee, 2012; ICI, 2014, 2015; Toynbee, 2016). However, time targets in these settings apply over a large number of trials, rather than on a trial-by-trial basis. As a consequence, passport officers may flexibly allocate their response time within a queue of travellers, provided that the whole queue is processed within the required timeframe.

To operationalise time pressure in this way, we developed a novel paradigm that administered time pressure via an onscreen speed gauge and a progress bar (Bindemann, Fysh, et al., 2016). Together, these onscreen displays relayed to

observers whether they were on track to complete a given block within the required timeframe, as well as the number of trials remaining. Observers could use this information to take more time on a difficult pair of faces, provided that surplus time was available, or speed up if they were progressing too slowly. The researchers found, however, that time pressure targets of 10-2 seconds exerted only a small numerical effect on performance, accounting for fewer than 11% errors. However, these performance data were collected under the optimised conditions provided by the GFMT. As a consequence, Bindemann, Fysh, et al.'s (2016) results may underestimate the extent to which time pressure impacts face-matching performance under the more difficult conditions at passport control.

Chapter 3 sought to address this by using the same paradigm to administer time pressure, but under the more challenging conditions provided by the KFMT. In Experiment 3, observers matched faces across five blocks, under time pressure that varied systematically from ten, eight, six, four, and two seconds. Converging with Bindemann, Fysh, et al. (2016), accuracy in this task deteriorated as time pressure increased, with the most errors arising under four and two seconds of time pressure. In addition, an effect of time passage was also observed, whereby accuracy deteriorated over the duration of the task, irrespective of whether time pressure was increasing or decreasing. Further evidence for these effects was provided in Experiment 4, which replicated the main effects of time pressure and time passage, but also clarified the marginal interaction between time pressure and trial type which was approaching significance in Experiment 3.

Together, Experiments 3 and 4 provide further evidence that the KFMT is more challenging than the GFMT. However, whilst these studies found large effects of time passage, the effect of time pressure was numerically similar to that observed by

Bindemann, Fysh, et al. (2016). This indicates that time pressure exerts a unitary effect on face matching that is not exacerbated by the difficulty of to-be-matched faces, but impacts observers' resource capacity to perform this task. The reasons for this time pressure effect are not immediately obvious. However, research limiting the amount of time for which faces are viewed in matching tasks suggests that observers employ different viewing strategies for comparing face stimuli under time constraints, as opposed to when the task is self-paced (see Özbek & Bindemann, 2011). For instance, this research suggests that at least two fixations per face are necessary to best facilitate face matching. However, these fixations tend to be directed at the eye regions when faces are displayed for two seconds, but focus on a greater portion of the face, encompassing the nose and mouth when faces are matched under self-paced conditions (see, e.g., Bobak et al., 2017; Özbek & Bindemann, 2011). Considered together, these findings suggest that time constraints of two seconds or less might prompt observers to adopt a feature-based processing strategy, but that under self-paced conditions, observers appear to process faces in a more holistic manner. Taking into consideration the findings of Experiments 3 and 4, as well as recent research investigating time pressure (e.g., Bindemann, Fysh, et al., 2016), it is possible that time pressure exerts a similar effect on observers' viewing strategies in face matching. In future research, this effect could be observed directly through the use of eye-tracking methodologies, and thus presents a logical avenue for studies seeking to deconstruct observers' viewing strategies and resource allocation in this task under time pressure.

It is notable that observers in Experiments 3 and 4 consistently seemed reluctant to use the full range of time that was available, but instead completed each block well within the required timeframe. For example, mean correct response times

did not exceed four seconds for any of the time pressure conditions, even when up to ten seconds were available. By contrast, response times on mismatch trials in Experiment 2, in which observers processed a similar number of identity pairs under self-paced conditions, were above four seconds in the first and third block. This is a curious finding, and raises the possibility that the perception of time pressure, as opposed to the actual time pressure that was being administered throughout Experiments 3 and 4, drove observers to respond more quickly than was required to complete each block on time. This is consistent with studies in which context effects appear to facilitate fast average response times of less than two seconds even under self-paced conditions, provided that time constraints were first imposed (see, e.g., Özbek & Bindemann, 2011). If passport officers respond to time pressure similarly in applied settings by compromising the allocation of perceptual resources to process travellers more rapidly, then additional identification errors may be facilitated.

A further aspect of these data that remains unclear is whether specific trials required more time to process than others. Consistent with the work of Bindemann, Fysh, et al. (2016), Experiments 3 and 4 show that observers make more errors when matching faces under time pressure of four and two seconds. However, because the order of trials was counterbalanced across, but not within blocks, it cannot be determined from the current data as to whether some face pairs consistently required more time to be processed than others. Research currently shows that faces must be viewed for a minimum duration of around two seconds to facilitate comparable matching accuracy as to when the task is self-paced (see Özbek & Bindemann, 2011; White, Phillips, et al., 2015). However, this threshold also reflects an average, rather than a precise cut-off. As a consequence, some faces can be matched accurately after a viewing duration of only a single second, even though general accuracy suffers under

such constraints (e.g., Bindemann et al., 2010; O'Toole, Phillips, et al., 2007; Özbek & Bindemann, 2011). Conversely, on trials for which the outcome is less obvious, observers might require up to three seconds. Given that some expert face matchers are more accurate with viewing durations of up to 30 seconds versus two seconds (see White, Phillips, et al., 2015), it would be useful to understand how processing time in this task can be optimised to reduce identification errors. This research could reveal further information regarding the strategies used by such high-performing experts (e.g., Towler, White, & Kemp, 2017; White, Dunn, et al., 2015; White, Phillips, et al., 2015), which could contribute to the reduction of errors in passport officers, who currently take considerably longer than student observers to process faces, without incurring performance benefits (White, Kemp, Jenkins, Matheson, et al., 2014).

The experiments presented in Chapters 2 and 3 indicate that in operational settings, considerable errors in face matching arise due to within-target variation, time pressure, and time passage. A potential solution to this problem is the development of Automated Border Control (ABC) systems at passport control, such as the e-Gates that are installed in most UK and European airports. These systems use state-of-the-art face recognition algorithms to compare a traveller's face to the digital face photograph that is stored on an electronic passport. However, although these e-Gates exhibit perfect, or near-perfect performance in some benchmark tests (see O'Toole, Phillips, et al., 2007; O'Toole et al., 2012; but see also Rice et al., 2013), these systems have committed egregious errors in applied settings, by mistaking men for women (http://www.bbc.co.uk/news/uk-england-manchester-12482156; ICI, 2011), and have created unmanageable delays following high false-rejection rates (ICI, 2014; Watt, 2016). Therefore, to maximise the performance of these systems, e-Gates are supervised by a human operator whose responsibility is to ensure that the system does

not incorrectly accept an impostor identity. Some research already suggests that human-computer interaction might reduce face-verification accuracy (White, Dunn, et al., 2015), but in a task which requires passport officers to compare a face to eight highly-similar targets. However, the extent to which a human operator's ability to verify the identification made by an algorithm in a pairwise matching task remains unclear.

To address this question, Chapter 4 explored this issue across three experiments. In Experiment 5, trial labels were presented alongside each stimulus pair that provided "same" or "different" identity judgements, or "unresolved" information. Observers were instructed that the majority of these trial labels were correct, but that a small number were also inconsistent with the trial type, and thus they were required to verify consistent labels and overrule those that were inconsistent. Performance deteriorated considerably on inconsistent, compared to consistent trial labels, suggesting that observers were influenced by the identity information that these provided. Despite this effect, performance was significantly below 100% on consistently-labelled trials, and was above 50% on inconsistently-labelled trials. This indicates that although observers' identifications were guided by the trial labels, their decision processes were also largely influenced by the facial information in stimuli. Consequently, observers seemed reluctant to fully comply with the onscreen labels.

Given that in operational passport settings, human operators may be likely to trust the e-Gate decisions on the majority of trials given that these are assumed to be generally correct, the aim of Experiment 6 was to encourage compliance with these labels. This was attempted by replacing all inconsistent trial labels in the first block with labels that provided consistent information. The findings provided further evidence that observers struggle to identify whether two faces show the same person

or different individuals, by showing that accuracy was compromised even when observers were consistently provided with the correct trial solutions. In turn, additional analyses suggested that observers were still reluctant to fully adhere to the information that these provided.

In a final attempt to encourage compliance with the trial labels, Experiment 7 administered trial-by-trial feedback to observers in the first block of trials. In addition, all inconsistent trial labels in the first and second blocks were replaced to provide consistent trial information. It was expected that by receiving feedback that aligned with the trial labels, observers would become more trusting of the information that these provided. As a consequence, it was expected that this would lead to an exaggerated number of errors on inconsistently-labelled trials in the final block. Performance in Experiment 7 was enhanced on match trials when compared to Experiment 6, suggesting that observers were more compliant with the trial labels. This compliance effect was consistent between Blocks 1 and 2, reflecting that observers continued to trust the labels after the feedback was withdrawn. An additional cross-experiment comparison did not reveal a significantly greater number of errors on inconsistently-labelled trials in the final block of Experiment 7 compared to Experiment 6. However, accuracy on inconsistent mismatch trials was significantly below 50%, suggesting that observers' responses on these trials were influenced to a greater extent by the trial labels, rather than the facial information within stimuli.

Considered together, Experiments 5-7 reflect that human decisions in face matching are influenced by external information (i.e. the trial labels). This converges with early studies in which the evaluation of incorrect semantic information interfered with the identification of familiar faces (Young, Ellis, Flude, McWeeny, & Hay, 1986), as well as unfamiliar objects (Lupker, 1985). In line with these studies, it would

seem that in Experiments 5-7, the information provided by the trial labels interfered with observers' identification processes.

Overriding inconsistent trial labels appeared to be particularly difficult in Experiment 7, when observers were given a reason to follow the trial labels. This increase in compliance benefitted performance on consistently-labelled trials, but reduced accuracy considerably on inconsistently-labelled mismatch trials, whereby responses were guided to a greater extent by the information provided by the trial labels than the facial information within stimuli. These findings converge with research in which the decisions of passport issuance officers actually curtail the identification accuracy of face recognition algorithms (White, Dunn, et al., 2015). In addition, these results are consistent with research in which observers' identification decisions are biased by misleading information from experimenters, such as manipulations of observers' previous identification decisions (see Johansson et al., 2005; Sagana et al., 2016; Sauerland et al., 2016), as well as the provision of false information that two different identities depict the same person (Menon et al., 2015b). In addition, observers' decisions in face matching are also guided by responses that are made by peers (Dowsett & Burton, 2015), and are biased by the concurrent presentation of biographical information (McCaffery & Burton, 2016).

Experiments 5-7 also raise some interesting questions that are of practical importance for passport control, such as the extent to which observers can resist being biased by these judgements. For instance, observers were instructed in these experiments that the majority of labels were accurate. An interesting further experiment to run would be to instruct observers specifically to ignore the information provided by the trial labels. This should, in theory, encapsulate the manner in which human operators verify e-Gates in practical settings, whereby identifications should

be made independently of e-Gates, and align with the algorithm only if correct. If a trial label effect emerges even under these conditions, then this would suggest that it is particularly difficult to avoid incurring a bias from prior information, and thus such labels may be particularly detrimental to face-matching accuracy in applied settings. This would require a fundamental rethink of how algorithm and human identification decisions should be combined best to maximise the security of ABC systems at passport control.

In addition, the trial label paradigm employed in Experiments 5-7 could further encapsulate applied settings by featuring multiple concurrent identity pairs that require verification, with each assigned their own trial label. This would provide a closer analogy to the task of human operators of e-Gates, who are frequently expected to monitor up to seven e-Gate booths simultaneously (FRONTEX, 2015a). Considering that research already shows matching performance to deteriorate considerably when observers are required to match two concurrent identities (see Bindemann, Sandford, et al., 2012; Megreya & Burton, 2006b), it is likely that Experiments 5-7 still underestimate the difficulty of overruling incorrect identity judgements by recognition algorithms. It is additionally likely that the difficulty of this task is further compounded by the verification of biographical information (McCaffery & Burton, 2016), as well as influenced by sleep deprivation. This latter factor is experienced by passport officers who frequently work irregular night-and-day shift patterns, and has been shown to reduce insight into one's own ability to match faces (Beattie et al., 2016), and may therefore further reduce observers' capacity to override incorrect identifications.

It would also be worthwhile to combine these trial label experiments with the time pressure paradigm used in Chapter 3 (see also Bindemann, Fysh, et al., 2016).

This is due to the fact that e-Gates are frequently endorsed for their time efficiency when processing travellers, and are expected to be of a comparable speed to regular passport officers, or faster (FRONTEX, 2015a). However, to maximise this efficiency, it is important that the human operator supervising the system responds to each identity judgement rapidly. Taking into consideration the finding that time pressure reduces face-matching performance (Experiments 3 & 4), it is important to understand whether observers become more compliant with the trial labels used in Chapter 4 under such pressure. If so, then this might reduce false rejection errors on consistent trials, whilst simultaneously increasing false acceptance errors on inconsistently-labelled trials.

Considered together, the findings in this thesis reflect that face matching is particularly challenging in passport control settings. This raises some concerns surrounding the detection of impostors in such contexts, given that these individuals are a documented security concern (NCA, 2015; Stevens, 2011). A potential caveat to these findings is that the reported experiments were run with student observers, and thus may be of only limited generalisability to passport control settings. However, current research reflects that experienced passport officers perform comparably to student observers under both taxing and optimised conditions (White, Kemp, Jenkins, Matheson, et al., 2014), as well as when working with algorithms to process passport applications (White, Dunn, et al., 2015). Despite this research, it would be useful for subsequent research to explore how such professionals perform in the KFMT experiments presented in Chapter 2, as well as under the additional demands of time pressure (Chapter 3), and when required to verify concurrently-presented trial labels (Chapter 4). Such research would increase the extent to which accuracy in operational settings can be estimated from the current results.

One consistent finding throughout the reported experiments was that performance on mismatch trials was particularly low. In Experiments 5-7, observers appeared to be particularly poor at detecting these identities when trial labels suggested that the onscreen faces depicted the same person. Although these findings reflect the fragility of observers' face-matching processes (see also Sauerland et al., 2016), perhaps more troubling was the finding in Experiments 2-4, whereby accuracy on these trials deteriorated due to a match response bias that emerged over time.

Despite the consistent emergence of this match response bias across multiple studies (see Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann, Fysh, et al., 2016; see also Chapters 2 & 3), its origins are unclear. Given that the identity pairs employed in Experiments 2-4 were photographed many months apart, this bias cannot be attributed to the similarity of the same-day identity matches employed in recent studies (e.g., Alenezi & Bindemann, 2013; Alenezi et al., 2015). In addition, it is difficult to reconcile the onset of this bias over time as a product of boredom or fatigue, given that rest-breaks do not replenish accuracy (Alenezi et al., 2015). However, this bias might arise due to the depletion of observers' capacity to distinguish one identity from another. In other words, observers' perceptual tolerance for variance between identities, which underscores the processes required to distinguish one person from another, becomes eclipsed by the perceptual tolerance for variance within identities, and thus leads to a greater number of perceived identity matches.

This explanation makes some sense when considering that on identity match trials, observers are always comparing different images of the same identity. As a consequence, some tolerance for variation between faces is necessary to facilitate this task, given that no face casts the same image twice (see Jenkins & Burton, 2011). Following this logic, it is possible that trial-by-trial feedback preserves accuracy in

this task by maintaining observers' tolerance for the variation within, versus the variation between identities (see Alenezi & Bindemann, 2013). However, it remains unclear as why observers' capacity for detecting mismatching identities declines over time, and not vice versa. Research indicates that the processes required to classify match and mismatch pairs are dissociable. For example, Megreya and Burton (2007) found that accuracy on match and mismatch trials did not correlate when observers matched pairs of unfamiliar faces. Moreover, although observers become more likely to classify faces as the same person over a prolonged duration (Alenezi et al., 2015), observers with face-specific deficits show an opposite impairment, and become increasingly likely to make identity-mismatch decisions (White et al., 2017). Together, these findings reflect that the classification of matching and mismatching identities rely on different cognitive processes. When considered alongside Experiments 2-4, it is possible that for observers who are not prosopagnosic, the processes utilised for detecting mismatches are exhausted early, resulting in a greater number of match responses.

This explanation makes some sense when considering that faces form a relatively homogeneous subset, in that they all share basic properties such as shape, texture, and featural configuration. As a consequence, all faces are inherently similar, but differ crucially in the internal face regions, which are typically fixated in face-matching tasks (see Bobak et al., 2017; Bindemann et al., 2009; Özbek & Bindemann, 2011). It is possible that the perceptual analysis of such features becomes more challenging over time, resulting in a greater reliance on the external features, which already appear to dominate unfamiliar face matching (Bruce et al., 1999; Kemp et al., 2016; Megreya & Bindemann, 2009).

In line with this theory, it is possible that the shifting of attention between two face stimuli might exhaust cognitive resources for detecting differences in these features, resulting in more similar-looking target faces. This is supported by studies that show face-processing capacity to be limited to only a single face. For example, Bindemann, Burton, and Jenkins (2005) found that in a sex-classification task, the presentation of a peripheral distractor face diverted observers' attention from a task-relevant name, but not from a task-relevant face. However, observers can deliberately redirect their attention to other faces, when it is advantageous to do so (Bindemann et al., 2007). As faces also appear to be special in retaining attention (see Bindemann, Burton, Hooge, et al., 2005), it seems plausible that shifting attention between two task-relevant face stimuli places non-trivial demands on processing resources, particularly given that face matching also requires the retention of internal minutiae in working memory to facilitate identity comparison between images.

Manipulating the differences between these internal features through changes in camera distance appears to reduce face-matching performance, even on same-identity trials (Noyes & Jenkins, 2017). Additional evidence suggests that exaggerating the distinctiveness of unfamiliar faces through image caricature improves accuracy on mismatching trials, but can concurrently increase errors on same-identity trials (McIntyre, Hancock, Kittler, & Langton, 2013). This research implies that making the differences between two faces more obvious, thereby manipulating data limitations in faces, might offset the onset of this match response bias. On the other hand, modulating observers' resource capacity for this task through trial-by-trial feedback also appears to prevent this bias from emerging (see Alenezi & Bindemann, 2013). Future research should explore further strategies for diminishing this bias.

In sum, this thesis presented a more challenging test of face matching; the KFMT (Experiments 1-2), which reliably measures similar processes to the GFMT but under more realistic viewing conditions. Following this, Experiments 3-4 found within such conditions, identification accuracy on mismatch trials deteriorated under time pressure, as well as over the passage of time. Finally, Experiments 5-7 reflect that human-computer interaction in face matching might not reduce errors, but rather, the commission of errors by algorithms might in fact promote error rates on mismatch trials in human observers. Together, these findings raise concerns surrounding person identification accuracy in operational settings, reflecting that this task might be more challenging still than is already estimated (e.g., Bindemann, Fysh, et al., 2016; Burton et al., 2010). The paradigms and stimuli employed in this thesis provide scope for further research, to provide increasingly realistic analogies to operational settings. The facilitation of such research will continue to advance understanding of identification performance in such contexts.

References

Alenezi, H. M. & Bindemann, M. (2013). The effect of feedback on face-matching accuracy. Applied Cognitive Psychology, 27, 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

Attwood, A. S., Penton-Voak, I. S., Burton, A. M. & Munafò, M. R. (2013). Acute anxiety impairs accuracy in identifying photographed faces. Psychological Science, 24, 1591-1594. doi:10.1177/0956797612474021

Australian Customs and Border Protection Service, (2015). Annual Report 2014 – 2015. Retrieved February 26, 2016, from https:// www.border.gov.au/ Reportsand Publications/Documents/annual-reports/ACBPS-Annual-report-2014-15-optimised.pdf.

Bate, S., Bennetts, R., Parris, B. A., Bindemann, M., Udale, R. & Bussunt, A. (2015). Oxytocin increases bias, but not accuracy, in face recognition line-ups. Social Cognitive and Affective Neuroscience, 10, 1010-1014. doi:10.1093/scan/nsu150

Bate, S., Parris, B., Haslam, C. & Kay, J. (2010). Socio-emotional functioning and face recognition ability in the normal population. Personality and Individual Differences, 48, 239-242. doi:10.1016/j.paid.2009.10.005

Beattie, L., Walsh, D., McLaren, J., Biello, S. M., & White, D. (2016). Perceptual impairment in face identification with poor sleep. Royal Society Open Science, 3, 160321. doi:10.1098/rsos.160321

Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. Cogent Psychology, 1, 986903. doi:10.1080/23311908.2014.986903

Bindemann, M., Attard, J., Leach, A. & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar face matching. Applied Cognitive Psychology, 27, 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, 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, 277-291. doi:10.1037/a0029635

Bindemann, M., Brown, C., Koyas, T. & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. Journal of Applied Research in Memory and Cognition, 1, 96-103. doi: 10.1016/j.jarmac.2012.02.001

Bindemann, M., Burton, A. M., Hooge, I. T., Jenkins, R., & de Haan, E. H. (2005). Faces retain attention. Psychonomic Bulletin & Review, 12, 1048-1053. doi:10.3758/BF03206442

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

Bindemann, M., Burton, A. M., Langton, S. R., Schweinberger, S. R., & Doherty, M. J. (2007). The control of attention to faces. Journal of Vision, 7, 1-8. doi:10.1167/7.10.15

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

Bindemann, M. & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. Perception, 40, 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, 415-435. doi:10.1068/p6922

Bindemann, M., Scheepers, C. & Burton, A. M. (2009). Viewpoint and center of gravity affect eye movements to human faces. Journal of Vision, 9, 1-16. doi:10.1167/9.2.7

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, e0148148. doi:10.1371/journal.pone.0148148

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

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. Quarterly Journal of Experimental Psychology, 70, 201-217. doi:10.1080/17470218.2016.1161059

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M. & Miller, P. (1999). Verification of face identities from images captured on video. Journal of Experimental Psychology: Applied, 5, 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, 207-218. doi:10.1037/1076-898X.7.3.207

Bundesdruckerei, (2013). ePassport pocket guide 2013. Retrieved May 26, 2014, from https://www.bundesdruckerei.de/sites/default/files/documents/2013/08/pocketguide_epass_en.pdf.

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

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

Burton, A. M., White, D. & McNeill, A. (2010). The Glasgow Face Matching Test. Behavior Research Methods, 42, 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, 243-248. doi:10.1111/1467-9280.00144

Chiroro, P. M., Tredoux, C. G., Radaelli, S. & Meissner, C. A. (2008). Recognizing faces across continents: The effect of within-race variations on the own-race bias in face recognition. Psychonomic Bulletin & Review, 15, 1089-1092. doi:10.3758/PBR.15.6.1089

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

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillside, NJ: Lawrence Earlbaum Associates.

Dalrymple, K. A. & Palermo, R. (2016). Guidelines for studying developmental prosopagnosia in adults and children. Wiley Interdisciplinary Reviews: Cognitive Science, 7, 73-87. doi:10.1002/wcs.1374

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

Davis, J. P. & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. Applied Cognitive Psychology, 23, 482-505. doi:10.1002/acp.1490

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D. & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. Law and Human Behavior, 28, 687-706. doi:10.1007/s10979-004-0565-x

Dolzycka, D., Herzmann, G., Sommer, W. & Wilhelm, O. (2014). Can training enhance face cognition abilities in middle-aged adults?. PloS one, 9, e90249. doi:10.1371/journal.pone.0090249

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, 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. Quarterly Journal of Experimental Psychology, 69, 1-10. doi:10.1080/17470218.2015.1017513

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, 576-585. doi:10.1016/j.neuropsychologia.2005.07.001

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

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

FRONTEX (2012). Best practice operational guidelines for Automated Border Control (ABC) systems. Retrieved from http://frontex.europa.eu/assets/Publications/Research/Best_Practice_Operational_Guidelines_for_Automated_Border_Control.pdf

FRONTEX (2015a). Best practice operational guidelines for Automated Border Control (ABC) systems. Retrieved from http://frontex.europa.eu/assets/Publications/Research/Best_Practice_Operational_Guidelines_ABC.pdf

FRONTEX (2015b). Best practice technical guidelines for Automated Border Control (ABC) systems. Retrieved from http://frontex.europa.eu/assets/Publications/Research/Best_Practice_Technical_Guidelines_ABC.pdf

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

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

Home Affairs Committee, (2012). The work of the UK Border Agency (December 2011 – March 2012), HC 71. Retrieved February 29, 2016, from http://www. publications. parliament.uk/pa/cm201213/cmselect/cmhaff/71/71.pdf.

HM Passport Office, (2016). Hansard fraud rates. Retrieved May 17, 2016, from https://www.whatdotheyknow.com/request/85742/response/225707/attach/html/4/Hansard%20fraud%20rates.pdf.html.

Independent Chief Inspector of Borders and Immigration. (2011). Inspection of Gatwick Airport North Terminal: April-September 2011. Retrieved from http://icinspector.independent.gov.uk/wp-content/uploads/2012/02/Inspection-of-Gatwick-Airport-North-Terminal.pdf

Independent Chief Inspector of Borders and Immigration (ICI). (2014). An inspection of border force at Stansted Airport: May-August 2013. Retrieved December 1, 2015, from http://icinspector.independent.gov.uk/wp-content/uploads/2014/01/An-Inspection-of-Border-Force-Operations-at-Stansted-Airport.pdf

Independent Chief Inspector of Borders and Immigration (ICI). (2015). Inspection of border force operations at Heathrow Airport: June-October 2014. Retrieved February 26, 2016, from http://icinspector.independent.gov.uk/wp-content/uploads/2015/07/Inspection-of-Border-Force-Heathrow-15.07.2015.pdf

Jenkins, R. & Burton, A. M. (2008a). Limitations in facial identification: The evidence. Justice of the Peace, 172, 4-6.

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

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

Jenkins, R. & Kerr, C. (2013). Identifiable images of bystanders extracted from corneal reflections. PloS one, 8, e83325. doi:10.1371/journal.pone.0083325

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

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. Science, 310, 116-119. doi:10.1126/science.1111709

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

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, 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, 211-222. doi:10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-O

Lander, K., Humphreys, G. & Bruce, V. (2004). Exploring the role of motion in prosopagnosia: Recognizing, learning and matching faces. Neurocase, 10, 462-470. doi:10.1080/13554790490900761

Lander, K. & Poyarekar, S. (2015). Famous face recognition, face matching, and extraversion. Quarterly Journal of Experimental Psychology, 68, 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).

Liu, C. H., Chen, W., Han, H. & Shan, S. (2013). Effects of image preprocessing on face matching and recognition in human observers. Applied Cognitive Psychology, 27, 718-724. doi:10.1002/acp.2967

Longmore, C. A., Liu, C. H. & Young, A. W. (2008). Learning faces from photographs. Journal of Experimental Psychology: Human Perception and Performance, 34, 77-100. doi:10.1037/0096-1523.34.1.77

Lupker, S. J. (1985). Context effects in word and picture recognition: A reevaluation of structural models. Progress in the Psychology of Language, 1, 109-142.

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

McIntyre, A. H., Hancock, P. J., Kittler, J., & Langton, S. R. (2013). Improving discrimination and face matching with caricature. Applied Cognitive Psychology, 27, 725-734. doi:10.1002/acp.2966

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

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

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

Megreya, A. M. & Bindemann, M. (2015). Developmental improvement and age-related decline in unfamiliar face matching. Perception, 44, 5-22. doi:10.1068/p7825

Megreya, A. M. & Burton, A. M. (2006a). Unfamiliar faces are not faces: Evidence from a matching task. Memory & Cognition, 34, 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, 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, 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, 364-372. doi:10.1037/a0013464

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, 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. Quarterly Journal of Experimental Psychology, 64, 1473-1483. doi:10.1080/17470218.2011.575228

Meissner, C. A. & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. Psychology, Public Policy, and Law, 7, 3-35. doi:10.1037/1076-8971.7.1.3

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, 1287-1305. doi:10.1080/13506285.2013.832451

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

Menon, N., White, D., & Kemp, R. I. (2015b). Identity-level representations affect unfamiliar face matching performance in sequential but not simultaneous tasks. The Quarterly Journal of Experimental Psychology, 68, 1777-1793. doi:10.1080/17470218.2014.990468

Moore, R. M. & Johnston, R. A. (2013). Motivational incentives improve unfamiliar face matching accuracy. Applied Cognitive Psychology, 27, 754-760. doi:10.1002/acp.2964

National Audit Office (NAO), (2007, February 7). Identity and passport service: Introduction of ePassports. Retrieved May 26, 2014, from http:// www.nao. org. uk/wp-content/uploads/2007/02/0607152.pdf.

National Crime Agency (NCA), (2015). National strategic assessment of serious and organised crime 2015. Retrieved March 2, 2016, from http://www. nationalcrimeagency.gov.uk/publications/560-national-strategic-assessment-of-serious-and-organised-crime-2015/file.

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., & O'Toole, A. J. (2017). Face recognition assessments used in the study of super-recognisers. arXiv preprint arXiv:1705.04739.

O'Toole, A. J., Abdi, H., Jiang, F. & Phillips, P. J. (2007). Fusing face recognition algorithms and humans. IEEE: Transactions on Systems, Man & Cybernetics, 37, 1149-1155. doi:10.1109/TSMCB.2007.907034

O'Toole, A. J., An, X., Dunlop, J. & Natu, V. (2012). Comparing face recognition algorithms to humans on challenging tasks. ACM Transactions on Applied Perception, 9, 1-13. doi:10.1145/2355598.2355599

O'Toole, A. J., Phillips, P. J., Jiang, F., Ayyad, J., Pénard, N. & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces over changes in illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, 1642-1646. doi:10.1109/TPAMI.2007.1107

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, 2145-2155. doi:10.1016/j.visres.2011.08.009

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

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

Phillips, P. J., Scruggs, W. T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L., & Sharpe, M. (2010). FRVT 2006 and ICE 2006 large-scale experimental results. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 831-846. doi:10.1109/TPAMI.2009.59

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. Image and Vision Computing, 32, 74-85. doi:10.1016/j.imavis.2013.12.002

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, 2235-2243. doi:10.1177/0956797613492986

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. The Quarterly Journal of Experimental Psychology, 70, 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

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

Robertson, D., Middleton, R. & Burton, A. M. (2015). From policing to passport control. Keesing Journal of Documents & Identity, February, 3-8.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R. & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. PloS one, 11, e0150036. doi:10.1371/journal.pone.0150036

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

Sagana, A., Sauerland, M., & Merckelbach, H. (2016). The effect of choice reversals on blindness for identification decisions. Psychology, Crime & Law, 22, 303-314. doi:10.1080/1068316X.2015.1085984

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

Stevens, C. (2011). Facing up to impostor fraud. Retrieved May 26, 2014, from http:// www.icao.int/publications/journalsreports/2011/ICAO%20MRTD%20Report%20Vol.6%20No.3,%202011.pdf.

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, 504-513. doi:10.1002/acp.3218

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

Toynbee, P. (2016, October 11). Take back control? Our border force is in no fit state to do its job. The Guardian. Retrieved from https:// www.theguardian.com/ commentisfree/2016/oct/11/take-back-control-border-force-lax-security.

UK Parliament, (2016). Passports: Fraud: Written question - 36668. Retrieved July 8, 2016, from http://www.parliament.uk/business/publications/written-questions-answers-statements/written-question/Commons/2016-05-05/36668/.

Ulrich, P. I., Wilkinson, D. T., Ferguson, H. J., Smith, L. J., Bindemann, M., Johnston, R. A., & Schmalzl, L. (2017). Perceptual and memorial contributions to

developmental prosopagnosia. The Quarterly Journal of Experimental Psychology, 70, 298-315. doi:10.1080/17470218.2016.1177101

Watt, H. (2016, July 11). Setting sun blinds hi-tech cameras at Stansted border control. The Guardian. Retrieved from https://www.theguardian.com/uk-news/2016/jul/11/stansted-setting-sun-blinds-border-control-hi-tech-cameras

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, 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, 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, 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, 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, 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, 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. Quarterly Journal of Experimental Psychology, 70, 287-297. doi:10.1080/17470218.2016.1173076

Young, A. W., & Burton, A. M. (2017). Recognizing faces. Current Directions in Psychological Science, 26, 212-217. doi: 10.1177/0963721416688114

Young, A. W., Ellis, A. W., Flude, B. M., McWeeny, K. H., & Hay, D. C. (1986). Face–name interference. Journal of Experimental Psychology: Human Perception and Performance, 12, 466-475. doi:10.1037/0096-1523.12.4.466