Unfamiliar face matching:
Decision-making and improvement

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

Natalie Wendy Gentry

School of Psychology

University of Kent

April 2019

Abstract

Research has consistently found unfamiliar face matching to be a highly error prone task. Yet, little is known about the decision-making process that underlies this task. Furthermore, methods of training observers to improve accuracy have demonstrated mixed success. Therefore, the experiments reported in this thesis investigated how matching decisions to pairs of unfamiliar faces are made (Chapter 2), and evaluated a novel method for improving face matching accuracy (Chapters 3 and 4). Chapter 2 examined whether identifications are based on a series of smaller assessments for individual facial features and if so, how these evaluations are combined to reach an overall decision, by comparing decisions to whole faces with those to isolated feature regions. Individual facial features were found to influence the classification of the whole face disproportionately, but performance was best when all features were presented as an integrated whole. This thesis also explored whether matching performance could be improved by providing observers with clearly-labelled examples (Chapters 3 and 4). The benefit of examples was explored at an individual level and revealed that observers who were low-performing at baseline improved with the help of examples (Chapter 3). This examples advantage was maintained after the examples were removed, generalised to previously unseen stimuli taken from the same set as the target pairs, and also demonstrated some generalisation to stimuli from a new set with different characteristics. Chapter 4 then used eye-tracking to evaluate how examples were utilised during matching tasks, but did not reveal a clear improvement with the provision of examples. The different pattern of results may have been due to fundamental task differences introduced by the eye-tracking methodology. Thus, further research is required to fully explore the feasibility of the examples manipulation as a method for improving unfamiliar face matching.

Acknowledgements

I would like to give thanks to Dr Markus Bindemann for his excellent and dedicated supervision. Also, to Professor Robert Johnston for starting me on my PhD journey. I would like to express my gratitude to my friends Jaimee, Matt and Will for keeping me sane and especially to Hannah for her unswerving patience and endurance of every rant, tirade and meltdown. I also want to convey special thanks to Gina for her constant reassurance, advice and for always being a sympathetic ear.

Finally, thank you most of all to my parents and my brother for their unwavering and unconditional support and encouragement.

Declaration

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

……

Natalie W. Gentry

Publications

Within this thesis Chapters 2 and 3 have been presented at conferences and Chapter 3 is currently under review for publication.

Chapter 2

Gentry, N. W., Bindemann, M., & Johnston, R. A. (2016). *The eyes have it! The role of specific facial features for identity matching*. Poster presented at the September meeting of BPS Cognitive Section, Barcelona, Spain.

Chapter 3

Gentry, N. W. & Bindemann, M. (2017). *Example face pairs improve identity-matching in low-performing individuals*. Poster presented at the July meeting of the Experimental Psychology Society, Reading, UK.

Gentry, N. W. & Bindemann, M. (under review). Examples improve face-matching accuracy. *Journal of Applied Research in Memory and Cognition*.

Table of contents

Chapter 1

General introduction

1.1 Introduction

Security personnel at international borders are routinely required to conduct identity comparisons between photographs on travel documents and their bearers, to determine whether they represent an identity-match (the same person) or an identity-mismatch (two different individuals). In psychology, this task of contrasting simultaneously-presented faces unknown to the viewer is known as unfamiliar face matching. This task is performed on a large-scale in applied settings. For example, over 75 million passengers travel through Heathrow airport alone each year (Heathrow Airport Limited, 2018). Furthermore, unfamiliar face matching is not limited to border control, but is also required in numerous other scenarios including age verification for products, such as alcohol and tobacco, entry into restricted areas and the identification of criminals from video footage in courtrooms. Although face matching in secure settings is becoming increasingly automated, these systems still require a human operator to supervise and override decisions where necessary (see FRONTEX, 2015). However, despite the wide application of this task, a growing body of research has demonstrated that unfamiliar face matching can be highly error prone (for reviews see, e.g., Fysh & Bindemann, 2017a; Robertson, Middleton, & Burton, 2015).

The difficulty of unfamiliar face matching was initially highlighted by one-in-ten tasks, where observers are required to match a target face to the corresponding member of a ten-photograph line-up, or state they are not there in target-absent line-ups (see Figure 1.1). In an early study, observers were required to match a high-quality video still of a target to the correct member of a photo line-up (Bruce et al., 1999). When the target was present in the line-up, observers identified the wrong face on around 10% of trials and incorrectly stated that the target was not in the line-up on approximately 20% of trials. On target-absent trials, participants erroneously selected a member of the line-up on 30% of trials. A number of studies have since reported similar accuracy levels for this task of around 70% (see, e.g.,

Bruce, Henderson, Newman, & Burton, 2001; Burton, Miller, Bruce, Hancock, & Henderson, 2001; Megreya & Burton, 2006a, 2008). Accuracy levels drop further if the pose or expression of the target is changed (Bruce et al., 1999). Even if observers are informed that the target will be present in every line-up, accuracy only increases to approximately 80% (Burton et al., 2001). These high error rates are surprising, as the majority of these studies utilised high-quality images for comparison, which were taken on the same day and did not present substantial differences between the photographs. One possible explanation is that the cognitive load of processing eleven faces may be too great, which may account for the high error rates found for one-in-ten matching tasks (Megreya & Burton, 2008). However, accuracy levels are similarly low when the number of faces in the line-up is decreased to eight (Henderson, Bruce, & Burton, 2001) or five (Megreya, Bindemann, Havard, & Burton, 2012).

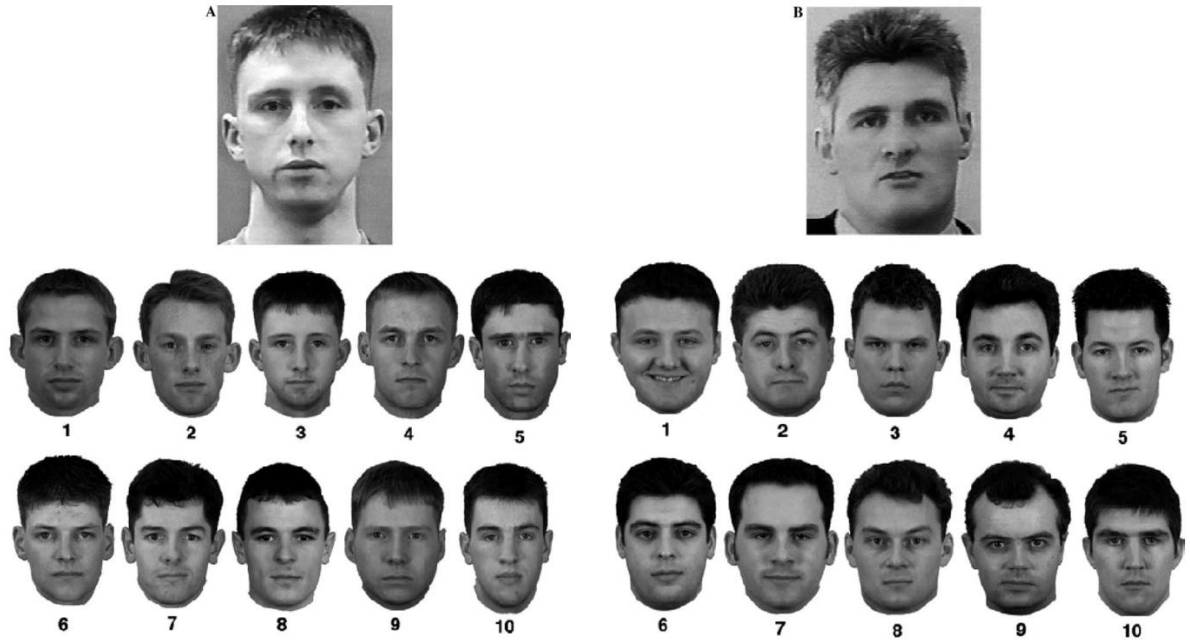

FIGURE 1.1. *Example of a target-present (A) and a target-absent (B) line-up (reproduced from Bruce et al., 1999). Observers are required to indicate whether the target is present in each line-up and if so, which number photograph depicts the target.*

The one-in-ten task requires observers to process a large number of faces at once, which may perhaps explain the difficulty of this task. In order to reduce this cognitive load, researchers have also examined unfamiliar face matching using one-to-one matching. This task is more representative of how unfamiliar face matching typically takes place in specific real-life settings such as passport control. However, similarly to the one-in-ten task, one-to-one face matching is also error-prone (e.g., Bindemann, Avetisyan, & Rakow, 2012; Burton, White, & McNeill, 2010; Fysh & Bindemann, 2018a). In an early demonstration of this, cashiers were required to verify the identity of confederates using photo credit cards (Kemp, Towell, & Pike, 1997). The cashiers incorrectly rejected 10% of genuine card holders and accepted over half of the fraudulent cards which depicted a different individual to the holder. This suggests that in real-life settings, accuracy is likely to be poor.

In laboratory-controlled tasks, higher levels of accuracy have been found. For example, requiring observers to match high-quality photographs to live actors produces accuracy rates of around 85% (Megreya & Burton, 2008), which is comparable to accuracy on one-in-ten tasks (e.g., Bruce et al., 2001; Burton et al., 2001). Matching two still images of unfamiliar faces is similarly difficult (see, e.g., Bindemann, Avetisyan, & Blackwell, 2010; Özbek & Bindemann, 2011). In the Glasgow Face Matching Test (GFMT; Burton et al., 2010), observers are required to match two high-quality same-day (for match trials) greyscale images, with a neutral expression and frontal pose taken on two different cameras (see Figure 1.2). Even under these optimised conditions, accuracy is still only around 80%. Thus, given the important security applications of unfamiliar face matching, there is a need to better understand the decision process behind unfamiliar face matching and ultimately improve accuracy for this task.

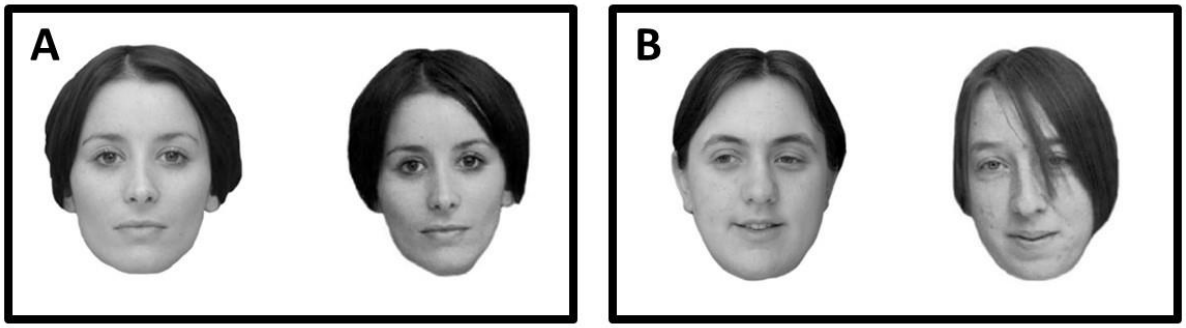

FIGURE 1.2. *Example identity-match (A) and identity-mismatch (B) pairs taken from the GFMT (Burton et al., 2010). The photographs were taken under optimised conditions using two different cameras (on the same day for match trials).*

In this review, I will explore why unfamiliar face matching is a difficult task and how matching decisions to unfamiliar faces might be made. I will then consider individual differences in face matching ability and how these suggest that improving performance on this task is theoretically possible. Finally, I will discuss existing attempts to increase matching accuracy and train observers to improve their unfamiliar face matching ability.

1.2 Why is unfamiliar face matching difficult?

Face matching may initially seem to be an easy task, as it is sometimes confused with *image* matching (Jenkins & Burton, 2011). Identity verification of an individual is trivially easy when given two identical images to compare (see Figure 1.3). However, the task becomes considerably more challenging when given two different images, as these can display substantial variation in the appearance of a person (see, e.g., Jenkins, White, Van Montfort, & Burton, 2011). An illustration of the importance of this variation comes from Bindemann and Sandford (2011), who presented observers with three different ID cards of the same individual and found that observers generally correctly matched the IDs with the target on 57% of trials. However, less than 40% of participants were able to correctly match all *three* ID cards to the target, indicating that the majority of observers were unaware that all three IDs belonged to the same person.

FIGURE 1.3. *Two images of the same individual can display substantial variation, even if these images are taken on the same day. Thus, matching images of the same person is considerably easier when provided with two identical photos of the target compared to two different images of them.*

Another reason the difficulty of face matching may be underestimated is that people are generally very good at matching *familiar* faces, such as those of friends, colleagues, or famous people. However, face matching is much more difficult for unfamiliar faces. Matching is faster and more accurate for familiar compared to unfamiliar faces (see, e.g., Bruce et al., 2001; Clutterbuck & Johnston, 2002, 2005; Young, McWeeny, Hay, & Ellis, 1986). Even when to-be-matched images are of poor quality (e.g., stills taken from low-resolution CCTV footage), matching accuracy for familiar faces remains high, whereas accuracy for unfamiliar faces deteriorates significantly (Bruce et al., 2001).

Matching unfamiliar faces may be more difficult than matching familiar faces because observers might be overly reliant on external features such as hairstyle, hairline and face shape to make identifications for unfamiliar faces (see, e.g., Bruce et al., 1999; Kemp, Caon, Howard, & Brooks, 2016; Megreya & Bindemann, 2009). By contrast, for familiar faces, internal features such as the eyes, nose and mouth appear to be more important for identification (Bonner, Burton, & Bruce, 2003; Campbell, 1999; Clutterbuck & Johnston, 2002). For example, accuracy for unfamiliar faces deteriorates more when external features are obscured compared to when internal features are hidden (Bruce et al., 1999). However, when attempting to match difficult face pairs, accuracy can increase by 5% when external

features are removed, as observers are forced to utilise information from the internal features to make a decision (Kemp et al., 2016). Thus, over-reliance on external features, which can easily be altered or changed, can be detrimental to matching accuracy (see, e.g., Kemp et al., 1997).

However, this external feature dependency is not present for all cultures. Egyptian observers are more accurate at matching unfamiliar faces with only the internal features displayed, whereas British observers are more accurate at matching unfamiliar faces with only the external features displayed (Megreya & Bindemann, 2009). This 'headscarf effect' has been attributed to observer experience with face identification based on internal features (see, e.g., Megreya, Memon, & Havard, 2012; Wang et al., 2015). In the middle-east, headscarves are typically worn, which can obscure external features such as face shape and hair. Thus, the Egyptian observers are likely to have had more experience using internal features to make identifications. These studies indicate that matching is likely to be more difficult for unfamiliar than familiar faces, as observers display an over-reliance on external features when greater examination of internal features may be required to make identifications.

1.3 Deterioration of unfamiliar matching accuracy

Under optimised conditions (i.e., with high-quality, same-day photographs to be matched), errors are made on 10-20% of trials on average (see, e.g., Bindemann, Avetisyan et al., 2010; Burton et al., 2010). These error rates are already problematic for large-scale security operations such as passport control (Dhir, Singh, Kumar, & Singh, 2010; Jenkins & Burton, 2008a). For example, over 200,000 passengers travel through Heathrow airport each day and if 10-20% errors were made, it could lead to a substantial number of cases being classified incorrectly. However, accuracy declines even further under less favourable

conditions. For example, face matching is more difficult with poor quality images, such as stills or video clips taken from CCTV footage (e.g., Bruce et al., 2001; Henderson et al., 2001). For low-resolution images, matching performance is close to chance whereas errors are substantially reduced for high-resolution images (Bindemann, Attard, Leach, & Johnston, 2013).

Matching errors also increase when to-be-matched faces are displayed from different viewpoints (e.g., Bruce et al., 1999; Favelle, Hill, & Claes, 2017). For example, pairs depicting one face in frontal view and the other in profile view can reduce mismatch accuracy by 10% (Estudillo & Bindemann, 2014). This may be due to different identity-related information being available for each view (Diamond & Carey, 1986). For example, a frontal view provides more information about the configuration of features whereas a profile view gives a greater indication of the depth of features. Uneven or unusual lighting also negatively impacts face matching performance (e.g., Favelle et al., 2017; Longmore, Liu, & Young, 2008). For instance, matching accuracy is reduced for bottom-lit faces compared to top-lit faces (Hill & Bruce, 1996).

While these factors are less likely to be a problem for strictly controlled passport photos, other conditions which are more likely to occur in applied settings can also be detrimental to unfamiliar face matching performance, such as when images have different facial expressions (see, e.g., Bruce, 1982; Bruce et al., 1999). In fact, simply embedding photographs within a passport frame containing biographical information appears to be sufficient to reduce accuracy (McCaffery & Burton, 2016), as does the usage of images taken months or years apart (see Figure 1.4), as is usually the case with a passport photograph and its bearer (see, e.g., Fysh & Bindemann, 2018a; Megreya, Sandford, & Burton, 2013). To illustrate the potential impact of this problem, matching performance can deteriorate by 20% when observers are required to match two images that are taken several months apart, rather

than images taken on the same day (Megreya et al., 2013). Similarly, matching accuracy is reduced by around 10% when using video footage that is a week old compared to footage taken on the same day for target present line-ups (Davis & Valentine, 2009).

FIGURE 1.4. *When applying for a driving licence in the UK, it is possible to take the photo from your current passport. The photograph on the driving licence on the left was taken from a passport when the individual depicted was 13 years old. The photograph on the identity card on the right is from a current student card. There is an approximate six-year time difference between the two photos and substantial variation between them. Yet, both are acceptable forms of ID currently used by the individual depicted.*

Observers' performance also deteriorates under conditions typically experienced by security personnel, such as when faces need to be matched over extensive time periods (e.g., Alenezi & Bindemann, 2013; Alenezi, Bindemann, Fysh, & Johnston, 2015). While accuracy for images of the same person can be maintained over time, mismatch accuracy declines substantially, such that after 1000 trials, this decreases to only 51% (Alenezi et al., 2015). These mismatch trials are representative of real-life imposters (i.e., individuals who are using falsely obtained ID), who may be attempting to bypass security for criminal reasons. Thus, this finding is especially problematic for real-life security scenarios where imposter detection is of vital importance. A further factor that is important for such real-life settings and which can impact face matching accuracy is time-pressure (Bindemann, Fysh, Cross, & Watts, 2016; Fysh & Bindemann, 2017b; Özbek & Bindemann, 2011; Wirth & Carbon, 2017), given that security personnel often operate under processing time targets (Border Force, 2018). For

example, mismatch accuracy was found to decline by 10% when observers were given two seconds to make a matching decision, compared to when they had ten seconds (Bindemann et al., 2016). Thus, there is a need to develop means of improving unfamiliar face matching performance in order to reduce the negative impact of these detrimental conditions that are typical in applied settings such as passport control.

1.4 Decision making in unfamiliar face matching

Whilst numerous studies have demonstrated that unfamiliar face matching is a difficult task (for reviews, see, e.g., Fysh & Bindemann, 2017a; Robertson, Middleton et al., 2015), far less is known about the decision process that underlies the identity verification of unfamiliar faces. In the related field of person *recognition*, there is a general consensus that faces are processed in a holistic manner, and thus, an overall decision is based on the entire stimulus with all parts integrated. Evidence for this comes from the composite effect and the part-whole effect (see Figure 1.5). In the composite effect, observers struggle to match part of a face which comprises of two different identities for the top and bottom halves when these are aligned, compared to when these are misaligned (see, e.g., Goffaux & Rossion, 2006; Le Grand, Mondloch, Maurer, & Brent, 2004; McKone, 2004). The part-whole effect results from observers recognising features more easily in the context of a face than in isolation (see, e.g., Donnelly & Davidoff, 1999; Tanaka & Farah, 1993; Tanaka & Sengco, 1997). Nevertheless, while these studies may give insight into how faces are processed when face recognition is required, it is not clear how such findings apply to face matching.

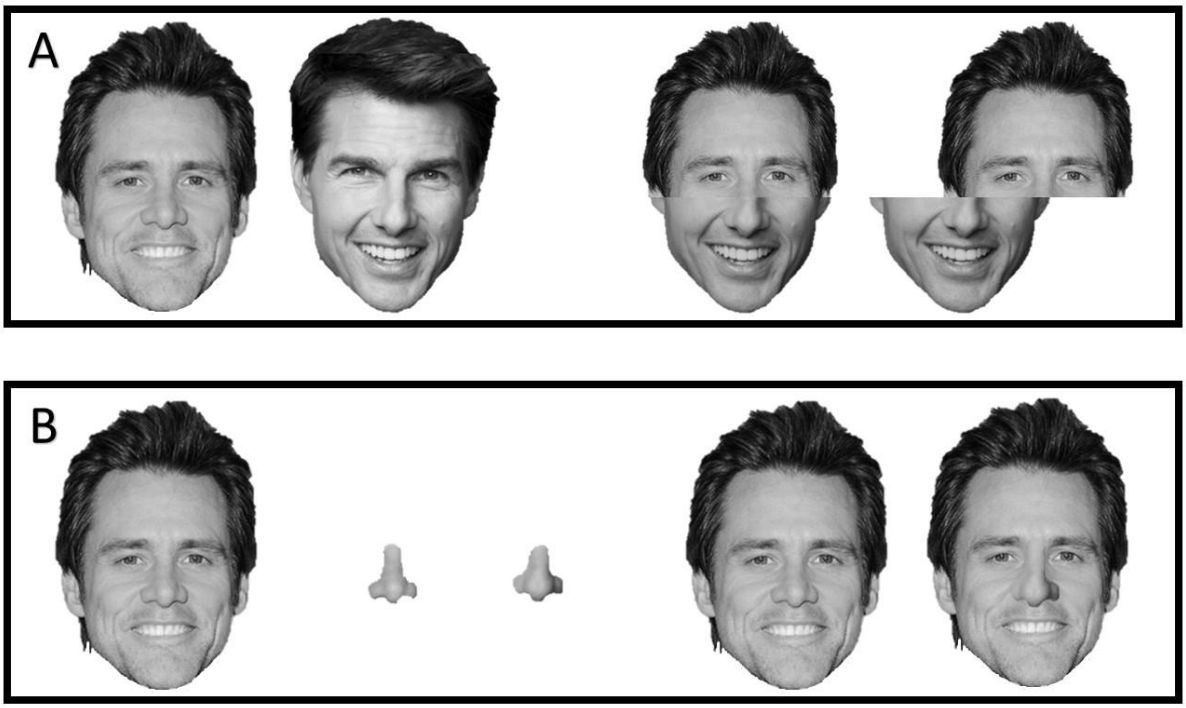

FIGURE 1.5. *Examples of the composite effect (A) and the part-whole effect (B). In the composite effect, it is more difficult to recognise the top half of the face (Jim Carrey) when aligned with the bottom face (Tom Cruise) than when the two are misaligned. In the part-whole effect, it is easier to select which is the correct nose for Jim Carrey when seen in the context of a face compared to when the noses are viewed in isolation.*

So how might matching decisions for unfamiliar faces be made? It is possible that decisions to pairs of unfamiliar faces are made in a similar holistic way to face recognition decisions (see, e.g., Donnelly & Davidoff, 1999; Goffaux & Rossion, 2006; McKone, 2004; Tanaka & Sengco, 1997). Observers may therefore base the overall decision on integrated information from the whole face. Alternatively, it is possible that an overall decision is based on a series of 'smaller' judgements to individual features, which can then be combined to reach a matching decision for the whole face. Unfamiliar face matching accuracy correlates moderately with the Matching Familiar Figures Test (MFFT, see Figure 1.6), which assesses object processing (Megreya & Burton, 2006b). The MFFT requires individuals to select the identical line-drawing from an array of similar images to the target. Typically, only one feature varies on each drawing from the target (e.g., the chimney stack on a boat). Therefore, in order to be successful on the MFFT, piecemeal, section-by-section processing is required.

This suggests that unfamiliar face matching may entail similar processing, where observers have to make judgements to individual features in order to make an overall face matching decision. If unfamiliar face matching decisions are made in this way, decisions to individual features may predict accuracy for the whole face.

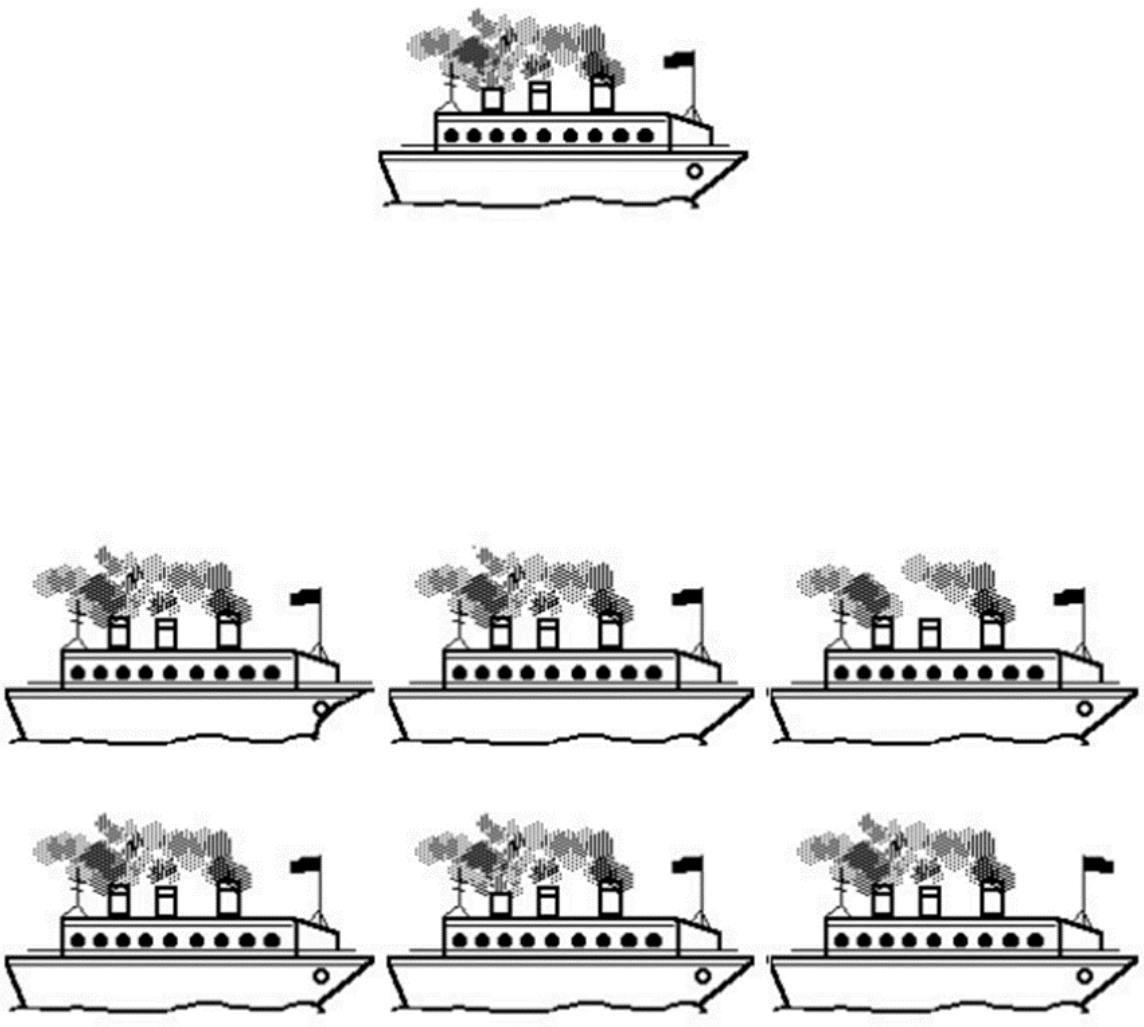

FIGURE 1.6. *Illustration of a target image and object array taken from the MFFT (Megreya & Burton, 2006b). The line-drawings in the array typically only vary from the target by one feature (e.g., the direction of the flag), except for the matching image (centre drawing, bottom row).*

In studies of person recognition, which assess memory for previously seen faces, researchers have attempted to pinpoint specific features that are diagnostic for person

identification. One feature that seems especially important for recognition accuracy is the eyes (see, e.g., Gilad, Meng, & Sinha, 2009; Tanaka & Farah, 1993; Vinette, Gosselin, & Schyns, 2004). Recognition accuracy is higher for the eyes (76%) than for the nose (64%) or the mouth (63%), when observers are required to recognise these features in isolation after learning name associations for the whole faces (Tanaka & Farah, 1993). The bubbles technique (see, e.g., Gosselin & Schyns, 2001), which limits the face regions available for processing, has also indicated the importance of the eyes for face recognition. When participants view faces partially obscured by masks containing small holes or "bubbles", accuracy is better for masks which allow the eyes to be utilised (Vinette et al., 2004). In addition, faces in photographic negative are also easier to recognise if the eyes are not negated (Gilad et al., 2009). However, this benefit is not found for other features such as the mouth. While these studies highlight the importance of the eyes for face recognition, alternative features have also been suggested as markers for accurate face recognition. These include, but are not limited to, the hair (Ellis, 1986; O'Donnell & Bruce, 2001), the eyebrows (Peissig, Goode, & Smith, 2009; Sadr, Jarudi, & Sinha, 2003), and the nose (Hills, Cooper, & Pake, 2013; Hsiao & Cottrell, 2008). Although these studies can discern features which are associated with successful face recognition, it is unclear how these might inform unfamiliar face matching.

At present, limited evidence exists in terms of whether there are any 'key' facial features that are more important for successful face matching and whether these can determine performance. Nevertheless, in a recent study, observers were required to determine perceptual differences in facial features to determine which of these were most 'critical' for matching (Abudarham & Yovel, 2016). Observers were required to judge features on a predefined scale in a feature tagging task (e.g., rate how large the mouth was for a given face) and also compare the features across two different faces (e.g., establish which face in a pair

had the widest jaw). These two measures were then combined to determine perceptual sensitivity for different features. Observers displayed high perceptual sensitivity for lip thickness, hair colour and eye colour, and thus these features were deemed to be the most important for face matching.

In another recent study, participants were required to successively rate the similarity of different features in pairs of faces, before making an overall matching judgement, to determine whether similarity ratings of specific features are related to whole face accuracy (Towler, White, & Kemp, 2017). Eleven different features were considered (face shape, ears, forehead, eyes, nose, cheek area, mouth, jawline, mouth area, chin and scars/blemishes). Using this approach, the ears were found to be the most diagnostic facial feature for successful matching. Furthermore, the ears were rated as the most useful feature for face matching by forensic examiners completing the task.

Instructing participants to focus on specific facial features has also informed how the contributions of different features to an overall whole face decision can differ (Megreya & Bindemann, 2018). After completing an initial accuracy assessment, observers were told to concentrate on either the eyebrows, eyes, or ears when making an overall matching decision for pairs of faces. Focusing on the eyebrows improved task performance, but attending the eyes had no impact on overall accuracy and performance declined when observers concentrated on the ears. These studies converge to suggest that individual features can strongly contribute to an overall matching decision and thus, decision making for pairs of unfamiliar faces may be based on featural processing. However, these studies also prescribe different features that are important for face matching, and hence it is possible that accuracy for a whole face is not dependent on one specific feature.

1.5 Individual differences in face matching ability

Despite error levels for unfamiliar face matching tasks typically being around 10-20% (see, e.g., Bindemann, Avetisyan et al., 2010; Burton et al., 2010; Özbek & Bindemann, 2011), a growing body of research has found that there is also between-subject variation in unfamiliar face matching ability (for a review, see Lander, Bruce, & Bindemann, 2018). These individual differences are pronounced, such that on the Glasgow Face Matching Test (GFMT, Burton et al., 2010), which has been used extensively to examine face matching ability, accuracy for individual observers varies from near-to-chance (i.e., 53%) to perfect (see Figure 1.7). Similarly, in the Kent Face Matching Test where observers are required to match images taken from student ID cards with a photograph taken at least three months later, accuracy ranges from below chance (40%) to 88% (Fysh & Bindemann, 2018a). Moreover, large variation in individual performance has been found for a number of unfamiliar face matching tasks (see, e.g., Bindemann, Brown, Koyas, & Russ, 2012; Estudillo & Bindemann, 2014; Kemp et al., 1997; Megreya & Bindemann, 2013; White, Kemp, Jenkins, Matheson, & Burton, 2014). These findings suggest that as some individuals are able to obtain perfect accuracy for face matching tasks (see, e.g., Burton et al., 2010), it should be possible to train low-performing individuals to improve their task performance.

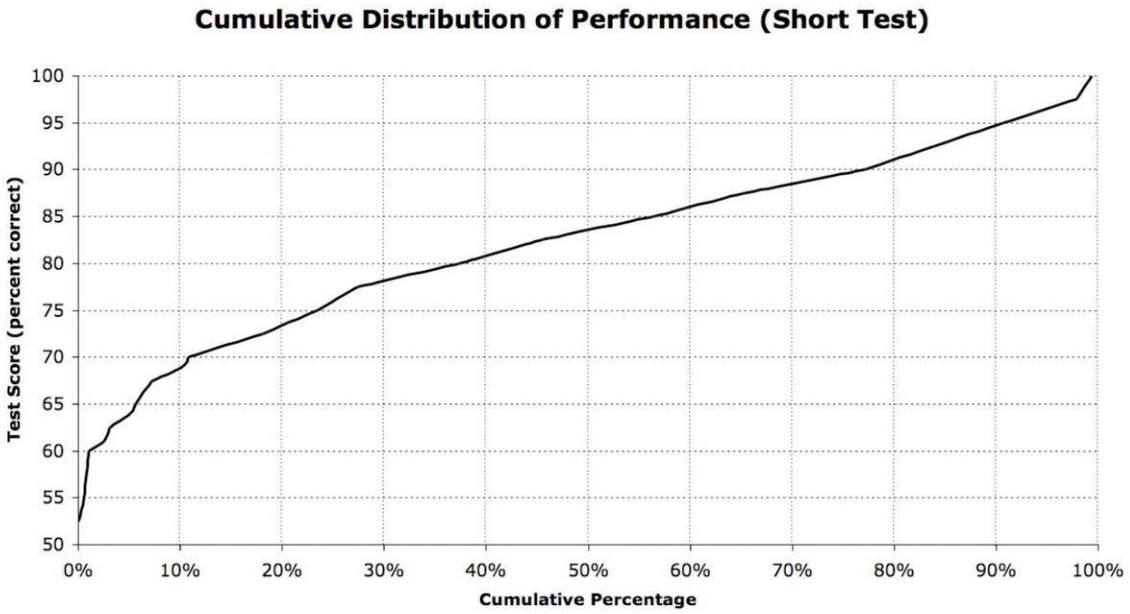

FIGURE 1.7. *An illustration of individual differences in performance on the GFMT (reproduced from Burton et al., 2010). Individuals range from near chance to perfect accuracy.*

As well as *inter*-observer variation, there is also considerable *intra*-observer variation for this task (see, e.g., Bindemann, Avetisyan et al., 2012). Observers can perform inconsistently across different days, with perfect accuracy on one day and many errors on another, and the same observer can make different matching decisions for the same face pairs on different days (Bindemann, Avetisyan et al., 2012). Individual performance for the same stimuli also declines over time (Alenezi et al., 2015). This variation within- and between-observers (see, e.g., Bindemann, Avetisyan et al., 2012; Bruce et al., 1999; Estudillo & Bindemann, 2014; Fysh & Bindemann, 2018a; Kemp et al., 1997) suggests that some professional matchers, such as passport control officers, may also be poor at matching unfamiliar faces.

1.6 Professional face matchers and individuals with exceptional face processing ability

It is especially important that individuals working in security, such as passport officers and police officers, are not only accurate but also consistent at this task. These individuals would be expected to do particularly well at this task because first, they have had substantially more practice than the general public and second, it is likely that they would have received training in order to do this task. Researchers have therefore been interested in investigating how these 'professional' matchers perform compared to other groups.

There are very few published studies that incorporate the police or passport officers due to the number of security restrictions in place. Nevertheless, one such study compared passport officers to student participants on an unfamiliar face matching task (White, Kemp, Jenkins, Matheson et al., 2014). They found that the passport officers were no more accurate than the students and in fact both groups performed poorly. However, passport officers took longer to make a decision for a face pair. In a more recent study, passport officers particularly struggled with the mismatching pairs and mistakenly accepted up to 25% of these stimuli as matches (Wirth & Carbon, 2017). This may be a reflection of the fact that in security scenarios such as passport control mismatches are rare. However, mismatch frequency has been found not to impair task performance (Bindemann, Avetisyan et al., 2010). Furthermore, experience has been found to be a poor indicator of task accuracy both for passport officers (e.g., White, Kemp, Jenkins, Matheson, et al., 2014) and other professionals who are required to routinely match unfamiliar faces (e.g., Papesh, 2018). As research with passport officers is limited, it is difficult to build a clear picture of how they may differ from the general public. However, it is clear that there is a need to find ways of improving unfamiliar face matching due to its importance in security scenarios.

While passport officers may not demonstrate superior performance to lay student participants, other professional groups have shown improved performance. In the UK, if

doubts regarding a face identity match occur in court, trained facial image analysts are sometimes used to assess the similarity of the two images (Bobak, Dowsett, & Bate, 2016). These individuals have been found to possess superior face matching skills when compared to individuals without facial image training (e.g., Norell et al., 2015; White, Phillips, Hahn, Hill, & O'Toole, 2015). 'Super-recognisers' employed by the Metropolitan Police Force have also been found to display enhanced face processing abilities (Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). However, these groups of professionals also display large individual differences in ability (see, e.g., White, Dunn, Schmid, & Kemp, 2015). In a recent study, forensic examiners (individuals with extensive training who perform rigorous face comparisons that can be used to assist expert testimony in courtrooms), facial reviewers (individuals trained to perform less thorough comparisons that can be utilised in law-enforcement) and super-recognisers outperformed fingerprint examiners and students on a challenging matching task incorporating images which were relatively unrestricted in terms of lighting, expression and appearance (Phillips et al., 2018). However, there were also substantial individual differences in performance for all groups. These differences were such that there was considerable overlap in the performance, which, for all groups, ranged from perfect (or nearly perfect for the student group) to around chance (see Figure 1.8).

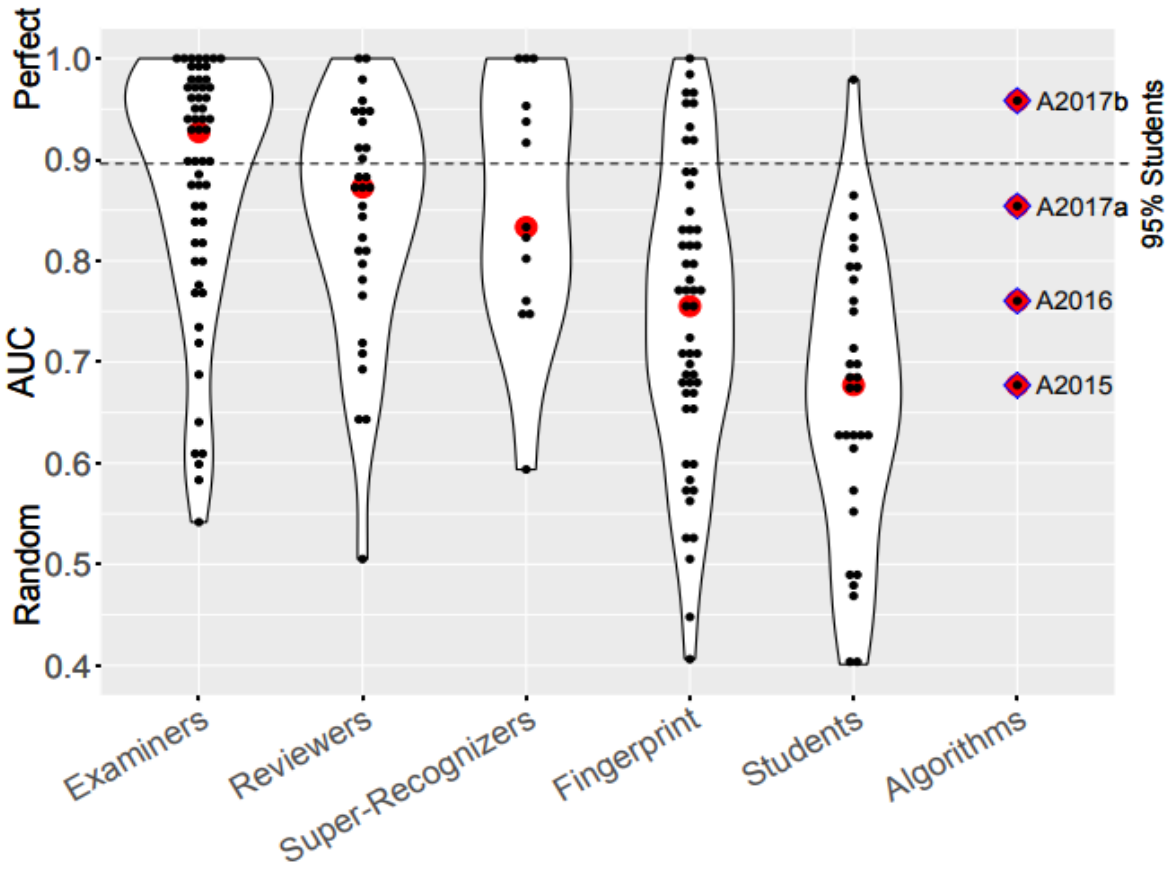

FIGURE 1.8. *Illustration of individual differences in accuracy for forensic examiners, facial reviewers, super-recognisers, fingerprint examiners, students for a difficult face matching task (reproduced from Phillips et al., 2018). Computer algorithms A2017b, A2017a, A2016, A2015 are also plotted on the right-hand side. Task performance was measured using the area under the receiver operating curve (AUC), which accounts for both hits and false positives made during the task. Although the three 'professional' face matching groups perform better on average than the fingerprint examiners and students, there is substantial overlap in task performance when individual differences are considered.*

Studies conducted on non-professionals have also identified individuals who display exceptional face identification abilities (see, e.g., Bobak, Hancock, & Bate, 2016; Russell, Duchaine, & Nakayama, 2009). Bobak, Dowsett, et al. (2016) tested the face matching ability of super recognisers using the GFMT. However, normative performance on the GFMT has been found to be reasonably high, around 80-90% (Burton et al., 2010). Therefore, the authors also used a more difficult face matching task, the Models Face Matching Test (MFMT), to assess the participants (see, e.g., Dowsett & Burton, 2015). The MFMT requires individuals to match images of models who have undergone a change in appearance from one

image to the next, such as a different hairstyle. Bobak, Dowsett, et al. (2016) found that on the GFMT, super recognisers were more accurate overall (97%) than the control group (87%). They also found that super recognisers outperformed the control participants on the more difficult MFMT, obtaining overall accuracy levels of 83% and 64% respectively. As some groups of individuals are able to obtain high levels of accuracy on unfamiliar face matching tasks (e.g., Bobak, Dowsett et al., 2016), it suggests that improvement on unfamiliar face matching tasks is possible and thus, theoretically it should be possible to train individuals to improve their unfamiliar matching accuracy.

1.7 Improving unfamiliar face matching

As unfamiliar face matching is an error prone task, a large body of research has focused on developing ways to try and improve accuracy. This research can be divided into two primary approaches; stimulus-focused and observer-focused. Stimulus-based approaches seek to increase matching accuracy by providing improved face representations such as caricatures (McIntyre, Hancock, Kittler, & Langton, 2013) and averaged faces (e.g., Burton, Jenkins, Hancock, & White, 2005) that could be used to redesign photo-ID. Observer-based approaches seek to increase accuracy on standard face matching tests (e.g., GFMT) using methods such as combining performance of multiple observers (e.g., White, Burton, Kemp, & Jenkins, 2013) and providing motivational incentives to increase performance (e.g., Moore & Johnston, 2013). Both of these approaches will be discussed in further detail in this section.

1.7.1 Stimulus-focused approaches

Matching unfamiliar faces from photographic ID may be difficult because the photos contain limited information for making an accurate matching decision (Jenkins & Burton, 2011). This could be viewed as a data-limited problem (Norman & Bobrow, 1975), where the

given stimuli do not provide sufficient information to determine the correct response. If unfamiliar face matching is a data-limited problem, it may not be possible to obtain consistently high levels of accuracy for this task when only a few face stimuli are provided (Jenkins & Burton, 2011). In turn, providing multiple images of faces to be matched has been found to improve unfamiliar matching accuracy (White, Burton, Jenkins, & Kemp, 2014). Observers were shown one, two, three, or four images of the faces to be matched. Accuracy increased in line with the number of photos displayed, though this improvement was limited to match trials only. However, when shown three different ID cards depicting the same person, observers were also more likely to identify the correct target when the faces were presented concurrently (85%) than sequentially (60%) (Bindemann & Sandford, 2011). These studies suggest that provision of a single photograph of a target may well be insufficient for maximising matching accuracy. In contrast, provision of multiple images of an identity allows observers to see how a person's face can change, thus seeing the variation in images of a single individual can make the task easier. However, the implementation of multiple images for face matching is likely to require more processing time so that the observers can make use of the additional photographs, which may in turn reduce the efficiency of high-pressed security services.

Passport photographs must depict a neutral expression, even though smiling is a more typical facial expression (Jenkins et al., 2011). In a recent study, observers were required to match pairs of open-mouth smiling faces and pairs of faces with a neutral expression (Mileva & Burton, 2018). Observers were more accurate at matching the open-mouth smiling faces than the faces with a neutral expression. This improvement was found for both match and mismatch pairs. Smiling may change the face in idiosyncratic ways and provide additional information such as teeth shape, which could make matching easier. This is in line with previous research that found smiling images of celebrities were rated as more representative

of a given identity, than images with a neutral expression (Jenkins et al., 2011). Thus, presenting observers with more variation in images such as the idiosyncratic changes produced by smiling, could improve matching accuracy.

Furthermore, averaged faces, which capture more stable face characteristics, can improve matching performance (see, e.g., Burton et al., 2005; Jenkins & Burton, 2008a, 2011; White, Burton et al., 2014). A single photograph may only provide limited information for face matching. However, averaged faces take into account a number of images of a person and encompass identity-related information, while removing more irrelevant face variation (see Figure 1.9). Averaged faces are generated by creating an average texture for the face by taking the mean RGB values for each pixel of the images to be used in the average. This texture is then morphed on to an average shape for the face, produced by placing and aligning feature landmarks for each of the images (see Burton et al., 2005 for more detail on the averaging process).

The naming of celebrity faces is faster when observers are shown an averaged image of them compared to a single photograph (Jenkins & Burton, 2011). Face recognition software used to unlock smartphones is also more accurate when based on an averaged face than a single photograph (Robertson, Kramer, & Burton, 2015). Moreover, averaged images also produce a higher hit rate than single photographs. For example, the FaceVACS recognition system accurately matched 100% of the averaged celebrity faces to the corresponding identity within the database, compared to a hit rate of only 54% for single images of the same celebrities (Jenkins & Burton, 2008a). Thus, producing more stable face representations which incorporate variation from a multitude of images of an identity for use on photographic ID, is likely to improve unfamiliar face matching performance. However, the process of creating an average image requires the use of multiple images of an individual. The more images are used to create an average, the more effective averages appear to become

(Burton et al., 2005). Hence, averaged faces may be difficult to implement for photographic ID such as passports and driving licences due to the large number of images required to produce them.

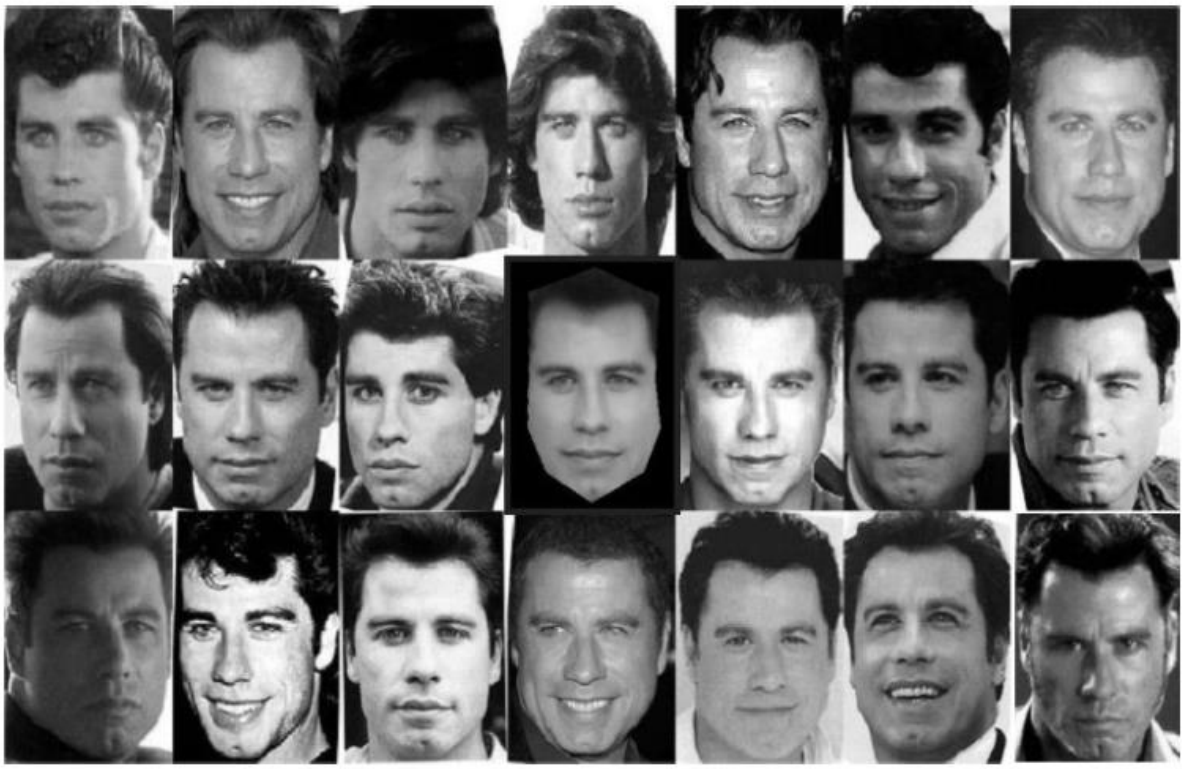

FIGURE 1.9. *Twenty images of John Travolta have been used to produce the averaged image in the centre (reproduced from Jenkins & Burton, 2008a). The individual photographs display substantial within-person variation, but the averaged image is a more stable representation of identity which can be more easily identified.*

Another type of face representation which may be used as an alternative to photographs is to use caricatures. In the related field of person recognition, caricaturing faces has been found to improve memory for and the recognition of faces (Deffenbacher, Johanson, Vetter, & O'Toole, 2000; Lee, Byatt, & Rhodes, 2000; Schulz, Kaufmann, Walther, & Schweinberger, 2012). Caricatures exaggerate the most distinctive features of faces and can make them more identifiable. Caricaturing has been applied to unfamiliar face matching to improve accuracy (McIntyre et al., 2013). Slight caricature was found to improve matching

accuracy on a one-in-ten line-up task. However, more extensive caricaturing increased the likelihood of participants responding that a target is not present in a line-up (McIntyre et al., 2013). Hence, using images with low levels of caricature may help to improve matching accuracy compared to normal photographs as they can help to increase the distinctiveness of faces and so make them easier to match. However, caricatures are very time consuming to produce, as every face has different distinctive features that would need exaggerating to produce an effective caricature. Furthermore, excessive caricaturing may harm matching accuracy.

1.7.2 Observer-focused approaches

While providing improved face representations such as averaged faces and caricatures for matching can increase accuracy, these methods may be difficult and time-consuming to implement for photo-ID. Observer-focused approaches to improving face matching seek to improve performance on tasks which replicate how face matching is typically performed in real-world scenarios. One such method is providing motivational incentives, in order to manipulate observer behaviour during matching tasks (e.g., Bobak, Dowsett et al., 2016; Moore & Johnston, 2013). When observers were offered a food incentive for above-average task performance, the incentive did not increase accuracy on match trials but improved performance for mismatch trials in the motivation condition (92%) compared to the control group (82%) (Moore & Johnston, 2013). Providing financial inducements as motivation also increases task accuracy (Bobak, Dowsett et al., 2016). However, such incentives are likely to be expensive and impractical to implement in real-world security scenarios, such as passport control due to the scale of these operations. Furthermore, passport officers, who should already be more motivated to be accurate, do not outperform student participants (White, Kemp, Jenkins, Matheson et al., 2014). Thus, while motivational incentives improve

accuracy in lab-based tasks, they are likely to have limited applied value for real-world matching scenarios.

An alternative observer-based approach to increasing face-matching accuracy is by using the 'wisdom of crowds'. Aggregating decisions of different individuals to form a group decision can lead to improved accuracy rates on unfamiliar face matching tasks (see, e.g., Balsdon, Summersby, Kemp, & White, 2018; Dowsett & Burton, 2015; Phillips et al., 2018; White et al., 2013). Previous research has found a great deal of within-participant and between-participant inconsistency on face matching tasks (e.g., Bindemann, Avetisyan et al., 2012; Bruce et al., 1999; Burton et al., 2010; Kemp et al., 1997; Megreya & Burton, 2008). Therefore, group estimates may be used to average out poor face-matching performance and lead to higher accuracy overall. Using a majority rule to reach a group decision was found to produce higher accuracy than that of any individual within the group (Balsdon et al., 2018; Phillips et al., 2018; White et al., 2013). Furthermore, allowing individuals to complete an unfamiliar face-matching task in pairs also resulted in improved accuracy compared to observers who completed the task alone (Dowsett & Burton, 2015). Thus, having more than one operator verify matching decisions in secure settings is likely to improve overall matching accuracy. However, needing multiple officers to verify decisions is also likely to reduce the efficiency of these services and so may not be possible to easily implement in real-life settings. Therefore, a more individual-based approach, such as training, may be more useful for improving accuracy in applied settings.

1.8 Training to develop better matching criteria

Another method of improving face matching performance is to provide individuals with training. Professional face matchers, such as forensic face examiners, receive feature comparison training for matching faces more effectively (see, e.g., White, Kemp, Jenkins,

Matheson et al., 2014; White, Phillips et al., 2015). These individuals have been found to outperform student groups at face matching tasks (see, e.g., Phillips et al., 2018; White, Dunn et al., 2015; White, Phillips, et al., 2015). Therefore, some researchers have utilised feature training in an attempt to increase task accuracy (see, e.g., Megreya & Bindemann, 2018; Towler, White & Kemp, 2014; Towler et al., 2017). An alternative method of training individuals to improve accuracy on unfamiliar face matching tasks is to provide them with feedback on how they have performed during the task. In real-life settings, observers are rarely given the opportunity to learn from their matching errors. Thus, providing them with feedback may also be used to improve task performance (see, e.g., Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014). Both of these training approaches will be discussed further in the following sections.

1.8.1 Training observers with features

Unfamiliar face matching correlates moderately with the Matching Familiar Figures Test (MFFT, Megreya & Burton, 2006b), that requires observers to determine which line drawing from an array matches with a target image. Images typically only differ by one feature, so in order to complete the task successfully, observers must process the images section by section. As unfamiliar face-matching performance is associated with MFFT accuracy, it is possible that unfamiliar faces are processed in a similar way whereby decisions made to individual features may be utilised to reach an overall matching decision. Moreover, facial examiners who perform face matching routinely are typically trained to use a feature comparison strategy (see, e.g., White, Kemp, Jenkins, Matheson et al., 2014; White, Phillips et al., 2015). Forensic examiners have been found to have enhanced processing skills (see, e.g., Phillips et al., 2018; White, Dunn et al., 2015; White, Phillips, et al., 2015), which may

be due to their ability to effectively use information from individual features to make an overall matching decision.

Consequently, some training schemes have attempted to improve face matching by instructing individuals how to use features to inform an overall matching decision. Forensic examiners typically conduct comparisons for facial features such as face shape to reach an overall matching decision (see, e.g., White, Phillips et al., 2015). Towler et al. (2014) investigated whether training participants to classify face shape improves unfamiliar face-matching performance. Participants were required to categorise one hundred photos of unfamiliar people using their face shape (see Figure 1.10). Observers were unaware that these photos consisted of 20 different identities with five photographs of each, two of which were identical photographs. The short version of the GFMT (Burton et al., 2010) was used to assess the participants' face-matching ability before and after commencing the face shape training. They found that there was low within-participant consistency, as on average, participants classified the same identity as having three different face shapes. Additionally, the five photos of the same identity were only classified as having the same face shape in 7% of cases. Moreover, the pairs of identical photographs were only judged as having the same face shape on roughly 50% of these occasions. These findings suggest that face shape is not critical for verifying identity and that face shape training is not effective for improving accuracy. Classification of face shape may be subjective and may be subject to variation when faces are seen from different angles. Therefore, focusing on specific *internal* features may be more likely to produce improvement in face-matching accuracy with training.

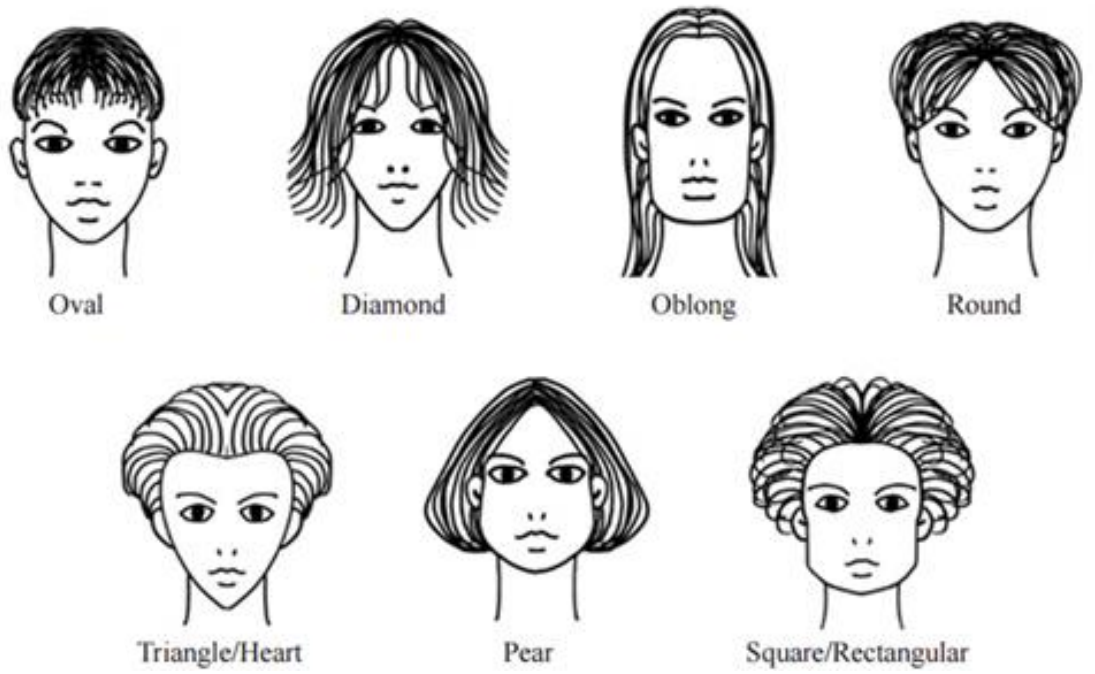

FIGURE 1.10. *Illustration of the seven face shape categories that participants were trained to recognise (reproduced from Towler et al., 2014). Observers were required to sort 100 photographs comprising five images of 20 different identities using these face shapes.*

Other training schemes focusing on facial features have been more successful (e.g., Towler et al., 2017; Megreya & Bindemann, 2018). Asking observers to sequentially rate the similarity of different face regions, such as forehead, eyes, nose and mouth, can improve accuracy for pairs of images depicting the same person (Towler et al., 2017). This training requires observers to rate 11 different facial features and so observers have to process faces in more detail than they might otherwise. Thus, encouraging observers to process faces in a way that gives greater consideration to individual features can improve task accuracy. Another recent study also found that instructing observers to focus on specific features when making an overall matching decision can impact task performance (Megreya & Bindemann, 2018). Participants were told to either utilise the eyebrows, eyes or ears for making a matching decision for a pair of faces. Focusing on the eyebrows improved accuracy, however, viewing the eyes had no impact on matching performance, and matching based on the ears resulted in a decline in overall accuracy. Thus, while training observers to attend specific features can

improve accuracy, focusing on some features can also be detrimental to performance and so this method of training may need further refinement as well as assessment of other facial features.

1.8.2 Training observers with feedback

Another training method that appears promising for improving matching accuracy is the provision of feedback. Errors in real-world face-matching tasks are rarely corrected and consequently individuals are not given the opportunity to learn from their mistakes (Jenkins & Burton, 2011; Jenkins et al., 2011). For instance, an imposter using false identification documents who is allowed to get through passport control, is highly unlikely to alert the passport officer of their mistake. Providing feedback can increase accuracy on both sequential (Hussain, Sekuler, & Bennett, 2009; Meinhardt-Injac, Persike, & Meinhardt, 2010, 2011) and simultaneous matching tasks (White, Kemp, Jenkins, & Burton, 2014). Feedback training provided while the face pair is still onscreen can also procure benefits for further matching tasks where no feedback is provided (White, Kemp, Jenkins, & Burton, 2014). On the other hand, Alenezi and Bindemann (2013) found that providing post-trial feedback only improved match accuracy. However, providing feedback also reduced a performance decline which typically occurs over time on mismatch trials (see, e.g., Alenezi et al., 2015). Similarly, in a more recent study the provision of trial-by-trial feedback was found to maintain but not increase mismatch accuracy over time (Papesh, Heisick, & Warner, 2018). Consequently, and irrespective of whether feedback increases accuracy or reduces the performance decline typically found for mismatch trials, feedback appears to have a positive impact on unfamiliar face-matching accuracy.

While feedback appears to be a promising method of improving performance in unfamiliar face-matching tasks, it cannot be easily implemented in real-world scenarios as it

requires *a priori* knowledge of the correct matching decision. Furthermore, providing trial-by-trial feedback was also found to be detrimental when there was a low-prevalence of mismatch trials, which is typical in secure settings (Papesh et al., 2018). The authors reasoned that this was due to the feedback making the observers more aware of how infrequent the mismatches were and thus, made the mismatches more difficult to detect. In light of these considerations, an alternative method of 'feedback', which utilises clearly-labelled example match and mismatch face-pairs to aid observers with their matching decision, may be more effective. This approach does not necessitate prior knowledge of the correct decision for pairs to be matched and thus, could be utilised in applied settings.

1.9 The structure of this thesis

Although the difficulty of unfamiliar face matching is well established (for reviews see, e.g., Fysh & Bindemann, 2017a; Robertson, Middleton et al., 2015), relatively little is known about how matching decisions are made. Thus, the aim of this thesis is to investigate how matching decisions to unfamiliar faces are made and whether this can provide a route to training. A small number of recent studies have found evidence to suggest that individual features can influence the overall matching decision (see, e.g., Abudarham & Yovel, 2016; Megreya & Bindemann, 2018; Towler et al., 2017). However, these studies demonstrate disagreement in terms of the feature that is most diagnostic. If there is no universal 'critical' feature that drives accuracy, it is possible that successful face matching requires observers to combine matching decisions to individual features to reach an overall decision.

Chapter 2 examined whether matching decisions made to individual facial features (i.e., hair, eyes, nose, mouth) inform the overall decision to the whole stimulus (face) with a series of three experiments. For this purpose, observers were required to match photographs of whole faces, as well as isolated feature pairs created by horizontally slicing the whole

faces into four key feature regions (Experiments 1 and 2). By aggregating and comparing the accuracy of the isolated feature pairs with whole face performance, these experiments sought to determine whether feature decisions are combined to reach a decision for the overall stimulus. For Experiment 3, observers matched whole faces, misaligned whole faces (displaying all facial features, but horizontally offset forcing them to be processed individually rather than as a whole percept), and misaligned part faces (with only two features visible, either the hair and nose or eyes and mouth). Accuracy was contrasted across these three presentation types to examine how the quantity of features available to make a decision and the integration of these features relate to task performance.

Chapter 3 then utilised the results of the previous chapter to develop and assess a novel method for improving matching accuracy. Feedback is a promising method of improving matching performance and has been shown to increase accuracy if provided when a just-classified face pair is still on view (White, Kemp, Jenkins & Burton, 2014) and maintain mismatch accuracy if delivered after a trial is completed (Alenezi & Bindemann, 2013; Papesh et al., 2018). However, this manner of feedback requires a priori knowledge of the correct decision for each pair and thus, is difficult to apply to real-world matching scenarios. The second empirical chapter addressed this shortcoming by providing an alternative form of feedback. As matching performance can be improved with feedback, it is possible that observers do not have adequate criteria for discriminating identity-match and identity-mismatch pairs and that feedback works by improving these criteria. Therefore, exemplars of labelled match and mismatch face pairs were provided alongside target face pairs over three experiments. As there are large individual differences in the accuracy of unfamiliar face matching (see, e.g., Bindemann, Avetisyan et al., 2012; Burton et al., 2010), these experiments specifically focused on how the examples manipulation impacted performance at an individual level. For all experiments, observers' baseline performance was

measured in an initial block of trials. Observers were then divided into two groups. One group was provided with example match and mismatch pairs flanking a target face pair in a second block, while the remaining group completed a repetition of the first block and saw no example pairs (Experiment 4). Generalisability of the examples was also examined when they were no longer displayed, when observers viewed new stimuli from the same set (Experiment 5) and when observers were presented with stimuli from a new set with different characteristics (Experiment 6). By contrasting accuracy for both the example and no-example groups, these experiments aimed to assess whether the provision of examples can improve task accuracy.

The last experimental chapter examined how these face exemplars were utilised in a matching task using eye-tracking (Experiment 7). Eye-movements were measured to determine how observers viewed the examples over the course of the experiment and whether their viewing behaviour related to task improvement. The impact of the nature of the examples provided was also assessed by providing three separate groups of observers with either low-difficulty examples (with little variation between match pairs and more differences between mismatch pairs), high-difficulty examples (with more dissimilar match pairs and mismatch pairs that appeared more similar) or no examples. Accuracy was compared across groups to examine the impact of the nature of the examples on accuracy.

Chapter 2

Matching faces and features:
The whole and the sum of its parts

Introduction

Unfamiliar face matching requires an identity comparison between two simultaneously presented faces unknown to the observer, to determine whether they depict the same person or two different people. A considerable body of psychological research has demonstrated that unfamiliar face matching is highly error prone (for reviews, see, e.g., Fysh & Bindemann, 2017a; Robertson, Middleton, & Burton, 2015). This task is challenging even for experienced professionals, such as passport control officers, who perform this task routinely (White, Kemp, Jenkins, Matheson, & Burton, 2014). Improvements in face matching may be possible with better theoretical understanding of the cognitive processes underlying this task. So far, however, little is known about the process by which face-matching decisions are made.

To understand what underlies performance on tasks requiring unfamiliar face identification, researchers have attempted to ascertain critical facial features that drive accuracy. In the related field of person *recognition*, which requires memory for a previously seen face, the eyes appear to be diagnostic for identification (see, e.g., Gilad, Meng, & Sinha, 2009; Keil, 2009; Tanaka & Farah, 1993; Vinette, Gosselin, & Schyns, 2004). Other studies have also varyingly emphasized the importance of the eyebrows (Peissig, Goode, & Smith, 2009; Sadr, Jarudi, & Sinha, 2003), nose (Hills, Cooper, & Pake, 2013; Hsiao & Cottrell, 2008) and hair (Ellis, 1986; O'Donnell & Bruce, 2001). Although these studies may provide insight into features underlying successful recognition, only limited evidence exists with regards to whether specific features determine accuracy in face *matching*.

To investigate this question, a recent study asked observers to judge discrepancies between specific features of pairs of faces (Abudarham & Yovel, 2016). Features for which observers displayed high perceptual sensitivity for identifying differences were deemed to be the most critical for face matching. Perceptual sensitivity was measured using two methods,

comprising the rating of feature characteristics on a predefined scale (e.g., mouth size), and a feature-matching task where observers compared features for a pair of faces (e.g., which face had the widest jaw). Lip thickness, hair colour and eye colour were found to be most diagnostic of identity, suggesting perhaps reliance on cues that can be fit into concrete categories (e.g., *blue* versus *brown* eyes).

However, the relevance of features for identification appears to be inconsistent across studies and methodologies. In another recent study, for example, observers successively rated the similarity of 11 facial features prior to making an identity-matching decision for a pair of faces (Towler, White, & Kemp, 2017). This approach showed that similarity ratings for the ears, followed by scars and blemishes were most indicative of accurate matching decisions. Instructing participants to focus on a specific feature indicates yet another key feature for making face-matching decisions (Megreya & Bindemann, 2018). In this study, an improvement in matching accuracy was found when observers focused on the eyebrows, but not the eyes, and performance declined when participants concentrated on the ears.

Overall, these studies therefore converge by suggesting distinct facial features can differ in their contribution to face-matching decisions. However, these studies also demonstrate disagreement in terms of *which* features are most diagnostic. One way to reconcile these results is that face-matching decisions are unlikely to be dependent on a universal 'critical' feature, but the features that are informative may depend on the individual, and the photographs of a specific individual, at hand. This reasoning seems sensible given that the same individual can vary substantially in appearance across different photographs (see, e.g., Jenkins, White, Van Montfort, & Burton, 2011), and that people vary in appearance in systematic but idiosyncratic ways (e.g., Burton, Jenkins, & Schweinberger, 2011; Burton, Kramer, Ritchie, & Jenkins, 2016).

If a specific feature is unlikely to drive face-matching decisions, then this raises the possibility that a combination of identity decisions to a set of individual facial features forms the basis of accurate matching decisions for whole faces. In other words, although match-mismatch decisions to pairs of faces must ultimately reflect a judgement that applies to the entire stimulus, these judgements may be proceeded by a series of smaller decisions to individual features that factor into the final matching decision. For example, it is conceivable observers make match-mismatch decisions to facial features, such as the eyes, nose and mouth, which are then combined to arrive at an overall decision as to whether two faces depict the same person or two different people. In such a framework, match-mismatch decisions for pairs of whole faces might be reached through a 'summing' of these smaller judgements, whereby the final decision is based on the overall proportion of individual decisions for distinct features pointing to the same outcome. In support of this reasoning, some evidence already exists to suggest that summing of responses provides insight into face matching accuracy. For example, combining face-matching judgements of small groups of observers using a majority rule to determine a collective decision produces more correct responses than any individual in the group (Balsdon, Summersby, Kemp, & White, 2018; Phillips et al., 2018; White, Burton, Kemp, & Jenkins, 2013). Similarly, allowing participants to work in pairs on identity-matching tasks improves overall performance (Dowsett & Burton, 2015). All of these studies, however, are based on identity judgements of the whole face.

In contrast to previous work, the current study investigated directly whether face matching is based on a series of smaller judgements for individual features and, if so, whether these judgements are summed to reach an overall decision. To this end, observers were presented with pairs of isolated facial features, comprising of hair / forehead, eye, nose and mouth regions. Matching decisions to these feature regions were compared with the

classification of these faces when these were displayed in their entirety. The accuracy with which individual facial features were classified as identity matches and mismatches was then assessed, and the relationship of feature accuracy to matching decisions for the whole face.

Experiment 1

This experiment investigated how decisions to isolated facial features relate to matching decisions for the corresponding whole face. For this purpose, participants made identity-matching decisions for pairs of isolated hair, eye, nose and mouth regions, as well as whole faces. Face matching studies suggest that individual facial features differ in terms of their contribution to an overall matching decision, but there is disagreement in terms of which features are most useful (see, e.g., Abudarham & Yovel, 2016; Megreya & Bindemann, 2018; Towler et al., 2017). Combining decisions to multiple features may therefore give insight for whole face accuracy. If an overall matching decision reflects the sum of judgements to individual features, one would expect that the more feature pairs classified correctly, the more likely the corresponding whole-face decision is to be correct. Applying a strict version of such a summation process leads to clear predictions. For example, for face pairs where all four features are classified correctly, whole face accuracy should be near perfect. By contrast, if only two of the four features are classified correctly, then observers should be equally as likely to respond correctly as incorrectly and thus, whole face performance should be close to chance.

Method

Participants

Twenty-three individuals (4 male, 19 female) from the University of Kent took part in this experiment. Observers had a mean age of 19.13 years (*SD* = 1.25, range: 18-24) and were

given course credit or a small fee for their time. All participants were of Caucasian ethnicity and reported normal or corrected-to-normal vision.

Stimuli

Eighty face pairs from the Glasgow University Face Database (GUFD) were utilised as stimuli for this study (see Burton, White, & McNeill, 2010). These comprised of 40 identity-matches (two different same-day photographs of the same individual) and 40 identity-mismatches (photographs of two different individuals). All faces were displayed in greyscale, with a frontal pose and neutral expression. The faces were cropped to remove extraneous background. The maximum size for a face was 90 x 120 mm, with a maximum gap between faces in a pair of 50 mm.

To create the feature stimuli for this experiment, the faces from these 80 pairs were also divided into four key sections, comprising of the hair, eye, nose and mouth region, so these could be matched in isolation (for an illustration, see Figure 2.1). The original onscreen position of these features was maintained, with the rest of the face replaced with background colour. This procedure produced 400 trials, reflecting hair, eyes, nose, and mouth regions, and whole face for each of the 80 pairs.

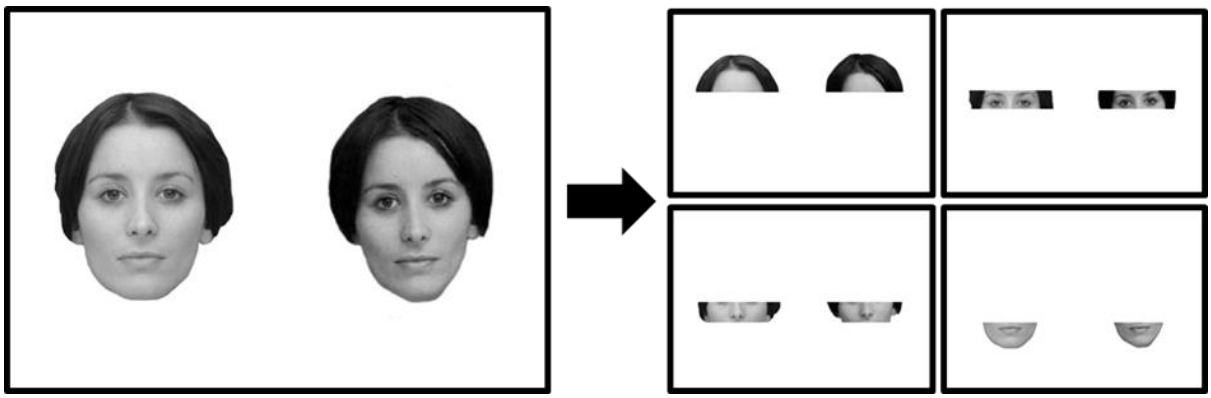

FIGURE 2.1. *Illustration of how whole face stimuli pairs were separated into isolated feature regions (hair, eyes, nose, mouth) for Experiments 1 and 2. The original positions of the features were maintained, and the rest of the face was cropped out.*

Procedure

The experiment was run using 'PsychoPy' software (Peirce, 2007). Participants were shown each of the 80 whole face pairs and the feature pairs derived from each face. The trials were organised into ten blocks, each comprising of 40 trials. All stimulus conditions were intermixed within blocks and presented in a randomized order. Participants could take a short break between blocks if desired. Observers were required to make a match or mismatch decision for each trial by pressing one of two keys on a standard computer keyboard. Participants were instructed to take as much time as necessary to complete the task and that accuracy was preferred over speed.

Results

Accuracy by feature

First, the percentage accuracy of observers' responses was analysed as a function of trial type (match vs. mismatch) and region of interest (ROI: whole face vs. hair vs. eyes vs. nose vs. mouth) to determine which feature produced the highest accuracy. The cross-participant means of this data are illustrated in Figure 2.2. A 2 (trial type) x 5 (ROI) within-subject ANOVA revealed a main effect of trial type, $F(1,22) = 6.04$, $p < .05$, $\eta_p^2 = .22$, and of ROI, $F(4,88) = 48.42$, $p < .001$, $\eta_p^2 = .69$, and an interaction between these factors, $F(4,88) = 6.02$, $p < .001$, $\eta_p^2 = .22$. Analysis of simple main effects indicated higher accuracy for mismatch than match trials for the hair, $F(1,22) = 9.65$, $p < .01$, $\eta_p^2 = .31$, eyes, $F(1,22) = 5.78$, $p < .05$, $\eta_p^2 = .21$ and mouth region, $F(1,22) = 8.45$, $p < .01$, $\eta_p^2 = .28$. Match and mismatch accuracy was comparable for the whole face and nose, both Fs ≤ 0.33, ps $\geq .57$.

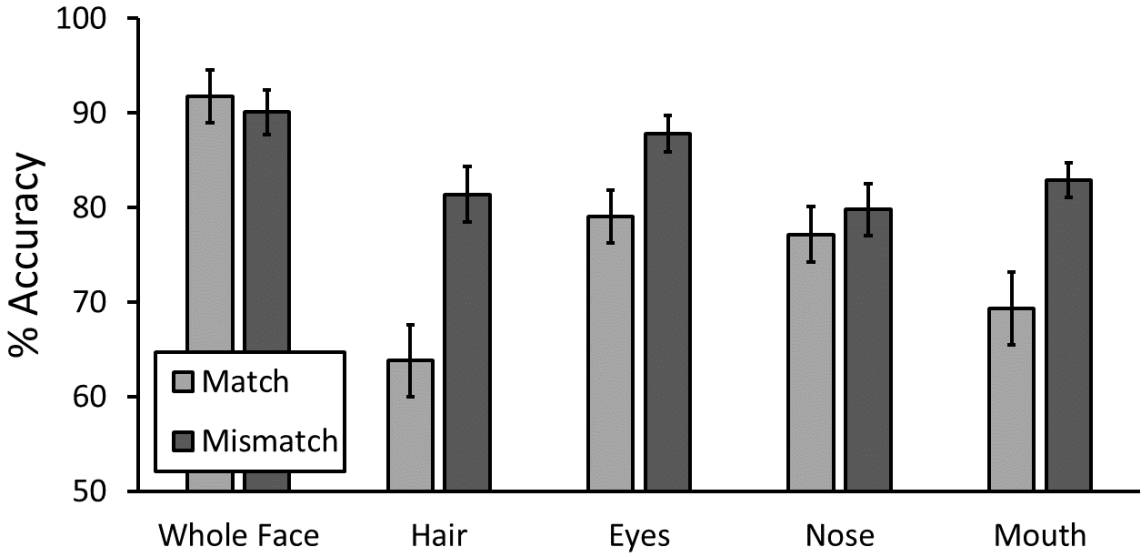

FIGURE 2.2. *Percentage accuracy for the whole face and isolated feature pairs (hair, eyes, nose and mouth) by trial type for Experiment 1.*

In addition, a simple main effect of ROI was found for match trials, $F(4,19) = 14.58$, $p < .001$, $\eta_p^2 = .75$. A series of paired-sample *t*-tests (with alpha corrected to .05/10 = .005 for ten comparisons) revealed higher accuracy for the whole face than all individual features, all $ts \geq 4.59$, $ps \leq .001$. Accuracy was also higher for the eyes than for the hair, $t(22) = 4.10$, $p < .001$, and mouth, $t(22) = 3.59$, $p < .005$, and for the nose than for the hair, $t(22) = 4.77$, $p < .001$, and mouth, $t(22) = 4.05$, $p < .005$. No other comparisons were significant, both $ts \leq 1.82$, $ps \geq .08$.

A simple main effect of ROI was also found for mismatch trials, $F(4,19) = 3.33$, $p < .05$, $\eta_p^2 = .41$. Paired-sample *t*-tests (with alpha corrected to .05/10 = .005 for ten comparisons) revealed higher accuracy for the whole face than the nose, $t(22) = 3.27$, $p < .005$, and mouth regions, $t(22) = 3.32$, $p < .005$. The difference in accuracy was approaching significance for the whole face and hair, $t(22) = 2.87$, $p = .01$, due to higher accuracy for the whole face. Furthermore, the accuracy difference between the eyes and the hair, $t(22) = 2.95$, $p = .01$, the nose, $t(22) = 2.84$, $p = .01$, and the mouth, $t(22) = 2.87$, $p = .01$, were also approaching significance due to higher accuracy for the eyes than the other features. No other

comparisons were significant, all *t*s ≤ 1.69, *p*s ≥ .10. Overall, this data shows that accuracy is generally higher for the whole face than any of the isolated features. Furthermore, none of the individual features demonstrate higher accuracy across *both* trial types than the other features. This suggests that accuracy is not driven by a universal critical feature.

Analysis of *d'* and *criterion* has been omitted here for brevity but is available for completeness in Appendix A for Experiments 1-3.

By-item analysis

To determine how classification of individual facial features relates to identification of the whole face, a by-item analysis was performed. For each facial identity pairing, the percentage accuracy scores for each feature and the whole face were averaged across participants. These scores were then correlated to examine whether accuracy for any of the individual features was associated with whole face accuracy. For match trials, accuracy for the whole face correlated with the eyes only, $r(38) = .71$, $p < .001$. No other correlations were found, all *r*s ≤ .26, *p*s ≥ .11. This indicates that when viewing images of the same person, the overall decision is more likely to be correct if the eyes are also classified correctly. For mismatch trials, accuracy for the whole face correlated moderately with the hair, $r(38) = .39$, $p < .05$, eyes, $r(38) = .35$, $p < .05$, and nose, $r(38) = .39$, $p < .05$, but not the mouth, $r(38) = .14$, $p = .40$. These correlations suggest that multiple facial features contribute to whole face accuracy when viewing images of two different people.

The individual feature pairs were also correlated with each other by trial type, to determine whether they provided independent information. For both match and mismatch trials, accuracy for the eyes correlated with the nose, both *r*s ≥ .33, *p*s ≤ .05. No other correlations for either trial type were found, all *r*s ≤ .28, *p*s ≥ .08. Overall, this data indicates that individual features generally provide independent information for face matching.

Number of features correct

To address the question of main interest, of whether judgements of individual features are combined to reach an overall matching decision, the average percentage of face pairs where observers classified all four, or three, two, one and zero features correctly was calculated. This data is displayed in Figure 2.3. For overall accuracy, this figure shows that three or four of the features belonging to the same face pair were classified correctly on most trials (29.4% and 41.3%, respectively). These percentages converge with the high mean accuracy for whole faces in this task (90.9%). Conversely, there were far fewer occasions on which none (0.6%), or only one (4.6%) or two (14.6%) features of a face pair were classified correctly. A similar pattern was evident when match and mismatch trials were considered separately (see Figure 2.3).

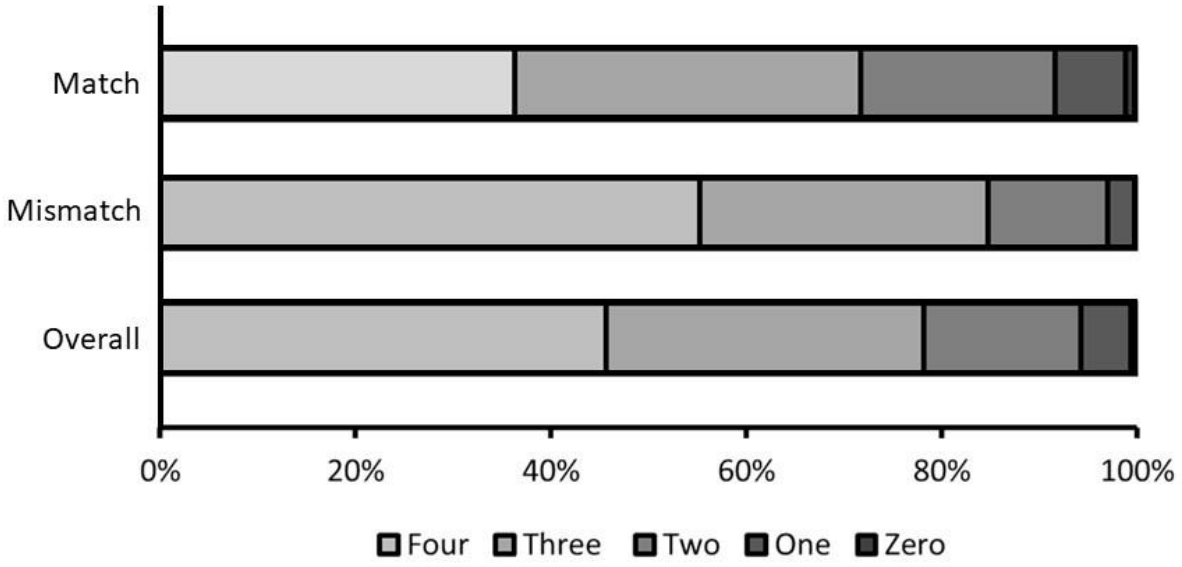

FIGURE 2.3. *Whole face correct trials broken down by the number of corresponding feature pairs correctly matched for Experiment 1. Three or four features correct make up the majority of the responses for both match and mismatch trials.*

A final step of the analysis sought to determine how the accuracy of these classifications, across different features belonging to the same face pairs, relates to the

classification of the respective whole faces. For this purpose, the proportion of correct responses for the whole face was calculated when participants made no, one, two, three or four correct feature decisions to the same identities. Considering the low number of items for which no or only one facial feature was classified correctly (see Figure 2.3), responses were collapsed across match and mismatch trials. This data is illustrated in Figure 2.4 and shows a graded response, whereby the proportion of correct classifications of the whole face pairs increases as more of the corresponding features are also classified correctly. In line with this observation, a one-way within-subject ANOVA of this data revealed a main effect of feature accuracy, $F(4,60) = 45.80$, $p < .001$, $\eta_p^2 = .75$.

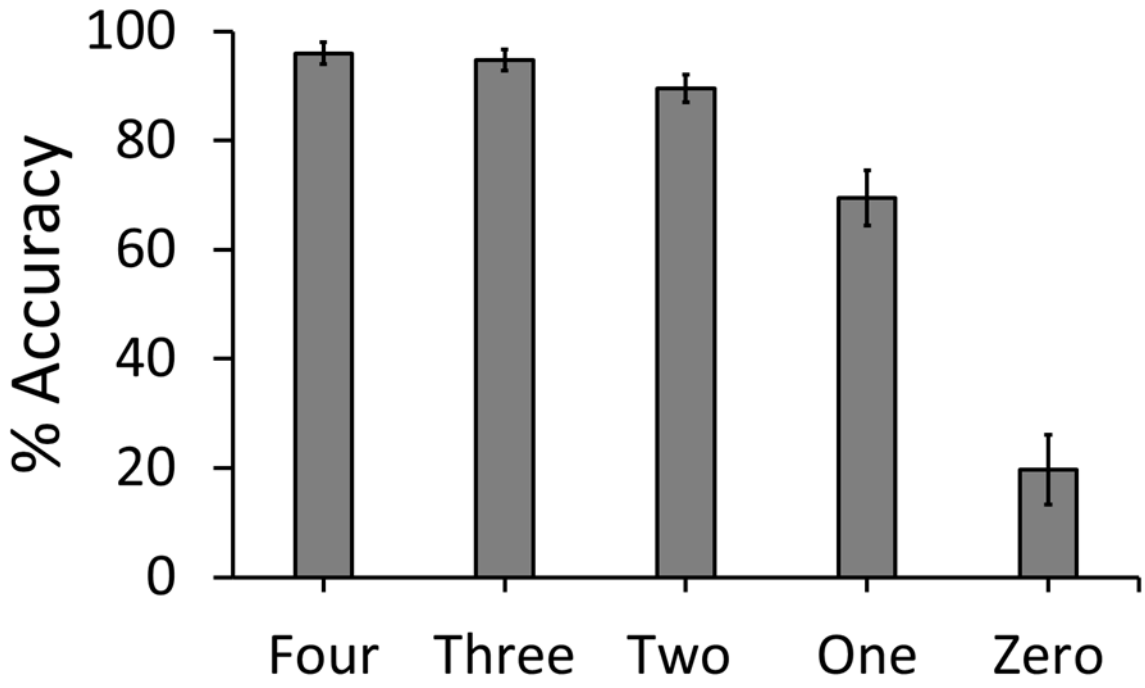

FIGURE 2.4. *Proportional whole face accuracy for each number of isolated feature decisions correct for Experiment 1.*

A series of paired sample *t*-tests (with alpha corrected to .05/10 = .005 for 10 comparisons) revealed whole face pairs for which none of the individual features were classified correctly were less likely to be matched correctly than faces for which all four

features, $t(15) = 8.76$, $p < .001$, three features, $t(15) = 8.37$, $p < .001$, two features, $t(15) = 7.38$, $p < .001$, or only one feature, $t(15) = 5.81$, $p < .001$, were correctly matched. In addition, the proportion of whole face pairs classified correctly was lower when only one feature decision was correct, compared to four, three, or two features correct, all ts ≥ 4.32, ps $\leq .001$. Whole face pairs were also matched correctly on fewer occasions when two features were correct compared to four features, $t(22) = 3.75$, $p < .005$, and was approaching significance for three features, $t(22) = 2.31$, $p = .03$. There was no difference in whole face accuracy when three or four features were correctly classified, $t(22) = 0.87$, $p = .40$. Overall this data indicates the less feature regions correctly classified, the less likely the whole face decision is to also be correct.

Finally, one-sample *t*-tests were conducted to compare proportional whole face accuracy for each number of features correct to chance accuracy (i.e., 50%). This revealed that the proportion of whole face trials classified correctly was above chance when one, two, three or four features were also classified correctly, all ts ≥ 3.84, ps $\leq .01$. However, the proportion of correctly matched whole face trials was below chance when no feature decisions were correct, $t(15) = 3.96$, $p < .01$. Thus, providing that observers classified at least one individual feature correctly, the whole face pairs for the same identities were also likely to have been classified correctly.

Discussion

This experiment investigated how decisions to isolated features may be utilised for reaching an overall matching decision. For this purpose, observers matched pairs of whole faces and isolated feature pairs (hair, eyes, nose and mouth) derived from these faces. Generally, accuracy for the whole face exceeded that of any of the individual features and none of the isolated feature pairs demonstrated consistently higher accuracy, across both

match and mismatch trials, than the other feature pairs. These findings are consistent with the notion put forward here that there is not a universal feature that primarily drives matching decisions. This is in line with previous studies demonstrating disagreement in terms of a critical feature, varyingly emphasising lip thickness (Abudarham & Yovel, 2016), ears (Towler et al., 2017) and eyebrows (Megreya & Bindemann, 2018).

Accuracy for the whole face and isolated features was also correlated to establish feature contributions for the overall matching decision. This analysis indicated that for match trials, observers are more likely to classify the whole face correctly if they have classified the eyes correctly. Thus, the eyes may be important for verifying images of the same person. For mismatch trials, on the other hand, hair, eyes and nose all correlated with the whole face, suggesting several features contribute to overall accuracy. Generally, however, the correlations were moderate in strength, suggesting that a specific feature is not related tightly, across all stimuli, to classification of whole faces.

To address the question of main interest, judgements for isolated features were compared with whole-face decisions to determine whether the number of features correct reflects overall accuracy. Here, more correct feature decisions increased the likelihood of the whole face decision being accurate. This suggests that judgements for individual features may be summed to reach a decision for the whole face. It is interesting to note, however, that even when only one feature decision is correct (i.e., most of the features are misleading and classified incorrectly), the proportion of correct whole face responses remains surprisingly high (around 70%). This may be counter-intuitive as one would predict that if feature judgements are summed to reach an overall decision, the whole face decision should reflect the majority feature judgement and should therefore be incorrect in these instances. Hence, it is possible that feature judgements are not always summed to reach a whole face decision, but that a single feature can also dominate the overall decision.

These conclusions are tempered by an aspect of the experimental design, as a two-alternative forced-choice task was employed. Thus, observers were required to register a match/mismatch decision for individual features even when these may have provided limited or inconclusive information for classification. If this resulted in a substantial number of feature errors, then this might explain those cases in which whole face accuracy was high despite the incorrect classification of most individual features. In the next experiment, this issue is addressed by adding a 'don't know' option, to clarify the findings of Experiment 1 and replicate the key results.

Experiment 2

Experiment 1 found whole face accuracy increased in line with the number of corresponding feature decisions that were correct. However, even when only one out of four features was classified correctly, the whole face decision was much more likely to be correct than incorrect. A possible explanation for this seeming anomaly may reflect the two-alternative forced-choice experimental procedure, which required participants to commit to a matching decision even when a stimulus pairing provided inclusive information. In those cases, observers were therefore obliged to guess, which may have inflated errors in this task. Considering that the individual feature conditions inevitably provided more limited visual content for matching than the whole face, this procedural aspect may have exerted a disproportionate effect on those feature trials. This might explain in part the cases in which several features of the same face are classified incorrectly. In turn, when matching whole faces, accuracy may have been maintained as a result of information from other features compensating for feature pairs that did not provide adequate matching information. If this can account for cases in which the majority of individual features are classified incorrectly but the whole face is classified correctly nonetheless, then these instances should be eliminated by

providing observers with an additional ‘don’t know’ response option. This was examined in Experiment 2.

Method

Participants, stimuli and procedure

Twenty new individuals (4 male, 16 female), with a mean age of 27.20 years (*SD* = 11.34, range: 18-51), took part in this experiment. The participants were given course credit for their time. All observers were of Caucasian ethnicity and reported normal or corrected-to-normal vision.

The stimuli and procedure were the same as for Experiment 1, but for one exception. In addition to the match and mismatch response options given previously, participants were also given the option of eliciting a ‘don’t know’ response via a third response key. Before commencing the experiment, participants were instructed to make use of this ‘don’t know’ option for difficult trials where they were unsure, rather than guessing.

Results

Accuracy by feature

As in Experiment 1, the percentage accuracy of participants’ responses was analysed as a function of trial type (match vs. mismatch) and ROI (whole face vs. hair vs. eyes vs. nose vs. mouth) to determine which feature produced the highest accuracy levels. The cross-participant means of this data are illustrated in Figure 2.5. A 2 (trial type) x 5 (ROI) within-subject ANOVA did not reveal a main effect of trial type, $F(1,19) = 0.56$, $p = .46$, $\eta_p^2 = .03$, but a main effect of ROI, $F(4,76) = 46.01$, $p < .001$, $\eta_p^2 = .71$, and an interaction between trial type and ROI, $F(4,76) = 4.25$, $p < .01$, $\eta_p^2 = .18$. Analysis of simple main effects did not indicate a difference in accuracy for match and mismatch trials for the whole face, eyes, nose

or mouth, all Fs ≤ 1.38, ps ≥ .26. However, the difference in accuracy between trial types was approaching significance for hair, $F(1,19) = 3.91$, $p = .06$, $\eta_p^2 = .17$, due to higher mismatch accuracy.

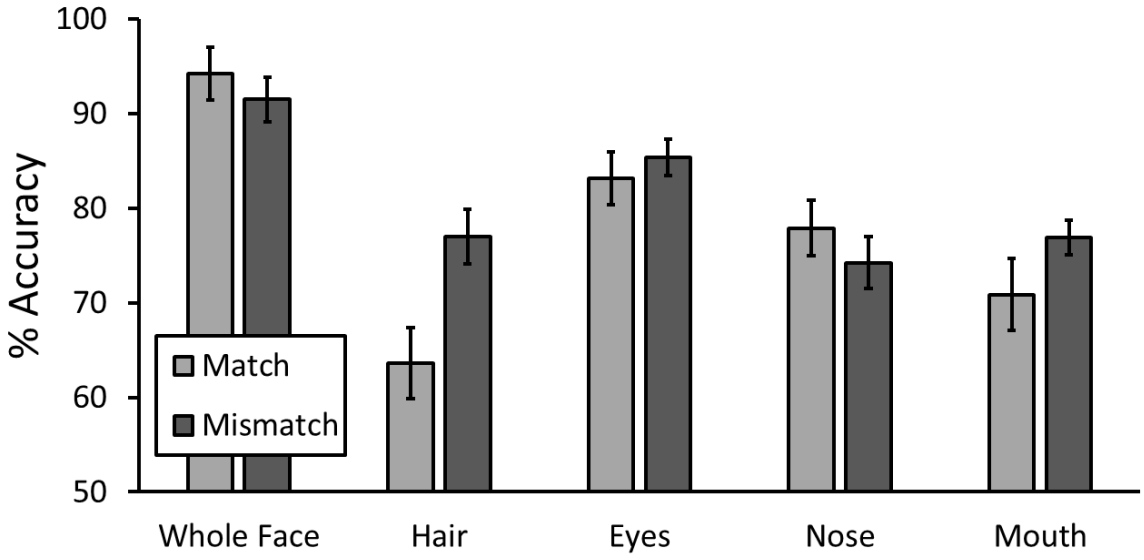

FIGURE 2.5. *Percentage accuracy for the whole face pairs and the feature in isolation pairs (hair, eyes, nose and mouth) by trial type for Experiment 2.*

In addition, a simple main effect of ROI was found for match trials, $F(4,16) = 29.48$, $p < .001$, $\eta_p^2 = .88$. A series of paired-sample t-tests (with alpha corrected to .05/10 = .005 for ten comparisons) revealed higher accuracy for the whole face than for all individual features, all ts ≥ 4.81, ps ≤ .001. Accuracy was also higher for the eyes than for the hair, $t(19) = 6.18$, $p < .001$, and mouth, $t(19) = 4.05$, $p < .005$, and for the nose than for the hair, $t(19) = 3.51$, $p < .005$, and was approaching significance for the mouth, $t(19) = 2.31$, $p < .03$. No other comparisons were significant, both ts ≤ 1.87, ps ≥ .08.

A simple main effect of ROI was also found for mismatch trials, $F(4,16) = 9.43$, $p < .001$, $\eta_p^2 = .70$. Paired-sample t-tests revealed higher accuracy for the whole face than for the hair, nose and mouth, all ts ≥ 4.38, ps ≤ .001. The difference was also approaching significance between the whole face and eyes, $t(19) = 2.88$, $p = .01$, due to higher whole face

accuracy. Accuracy for the eyes was also higher than for the nose, $t(19) = 3.95$, $p < .005$, and was approaching significance for the hair, $t(19) = 2.49$, $p = .02$, and mouth, $t(19) = 3.20$, $p = .01$. No other comparisons were significant, all ts ≤ 1.09, ps $\geq .29$. Overall, this data shows that as in Experiment 1, accuracy is higher for the whole face compared to the isolated features. In addition, accuracy for the eyes is higher than for some of the other isolated features, but not consistently across both trial types.

By-item analysis

To establish whether a relationship exists between accuracy for any of the isolated features and the whole face, percentage accuracy for each feature was averaged across participants by trial and correlated with the means for the whole face pairs for each trial type. For match trials, accuracy for the whole face correlated with the hair, $r(38) = .32$, $p < .05$, eyes, $r(38) = .44$, $p < .01$, nose, $r(38) = .36$, $p < .05$, and mouth, $r(38) = .52$, $p < .01$. Similarly, for mismatch trials, accuracy for the whole face correlated with the hair, $r(38) = .46$, $p < .01$, nose, $r(38) = .46$, $p < .01$, and mouth, $r(38) = .32$, $p < .05$, but not the eyes, $r(38) = .13$, $p = .44$. These correlations suggest multiple features contribute to overall accuracy, thus suggesting that it is not a single feature that drives classification of the whole face but a combination of features.

The isolated feature pairs were also correlated with each other to assess their independence. For match trials, accuracy for the hair correlated with the eyes, $r(38) = .32$, $p < .05$, and accuracy for the nose correlated with the mouth, $r(38) = .46$, $p < .01$, but none of the other features correlated, all rs $\leq .21$, ps $\geq .20$. For mismatch trials, none of the isolated features correlated with each other, all rs $\leq .31$, ps $\geq .05$. These correlations show most of the isolated features provide independent matching information and those that do correlate are not strongly associated.

Number of features correct

In this experiment, participants had an additional response option of 'don't know' as well as match and mismatch. However, this option was utilised on only 6.7% of trials. To determine whether decisions to isolated feature pairs reflect the whole face decision, the proportion of whole face trials correctly classified was calculated when observers matched four, three, two, one and none of the corresponding features correctly. This data is illustrated in Figure 2.6. Similarly to Experiment 1, for the majority of trials, three or four of the features derived from the same face pair were classified correctly (32.1% and 38.5% respectively). Only two, one or zero features were matched correctly on far fewer occasions. This pattern was evident for both match and mismatch trials.

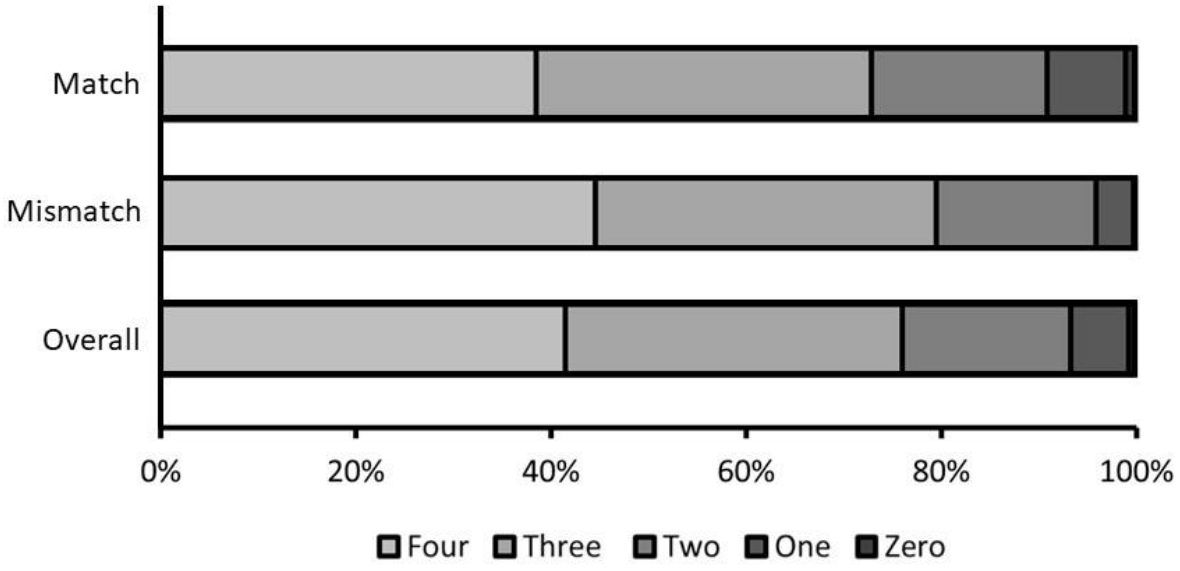

FIGURE 2.6. *Whole face correct trials broken down by the number of corresponding feature pairs correctly matched for Experiment 2. For most of the face pairs, three or four of the corresponding features were classified correctly.*

To determine the effect of feature decision accuracy on whole face performance, a one-way within-subject ANOVA was conducted. ANOVA revealed a main effect of feature accuracy, $F(4,52) = 9.55$, $p < .001$, $\eta_p^2 = .42$ (see Figure 2.7). A series of paired-sample *t*-

tests (with alpha corrected to .05/10 = .005 for 10 comparisons) revealed that having no correct feature decisions reduced the likelihood of the whole face decision for the same identities being correct compared to four, $t(13) = 4.20$, $p < .005$, three, $t(13) = 4.03$, $p < .005$, or two, $t(13) = 3.62$, $p < .005$, correct feature decisions. Similarly, one correct feature decision resulted in a lower proportion of correct whole face responses than four, $t(18) = 3.50$, $p < .005$, and was approaching significance for three, $t(18) = 3.16$, $p = .01$, and two, $t(18) = 2.44$, $p = .03$. The whole face was also less likely to be classified correctly when only half the features (i.e., two) were correct compared to when all the features (i.e., four), $t(19) = 4.60$, $p < .001$, or three, $t(19) = 3.39$, $p < .005$, were correctly matched. No other comparisons were significant, both ts ≤ 1.71, ps $\geq .10$. Overall, this data shows that decisions to whole face pairings were more likely to be correct as more of the features belonging to the same identities were also classified correctly.

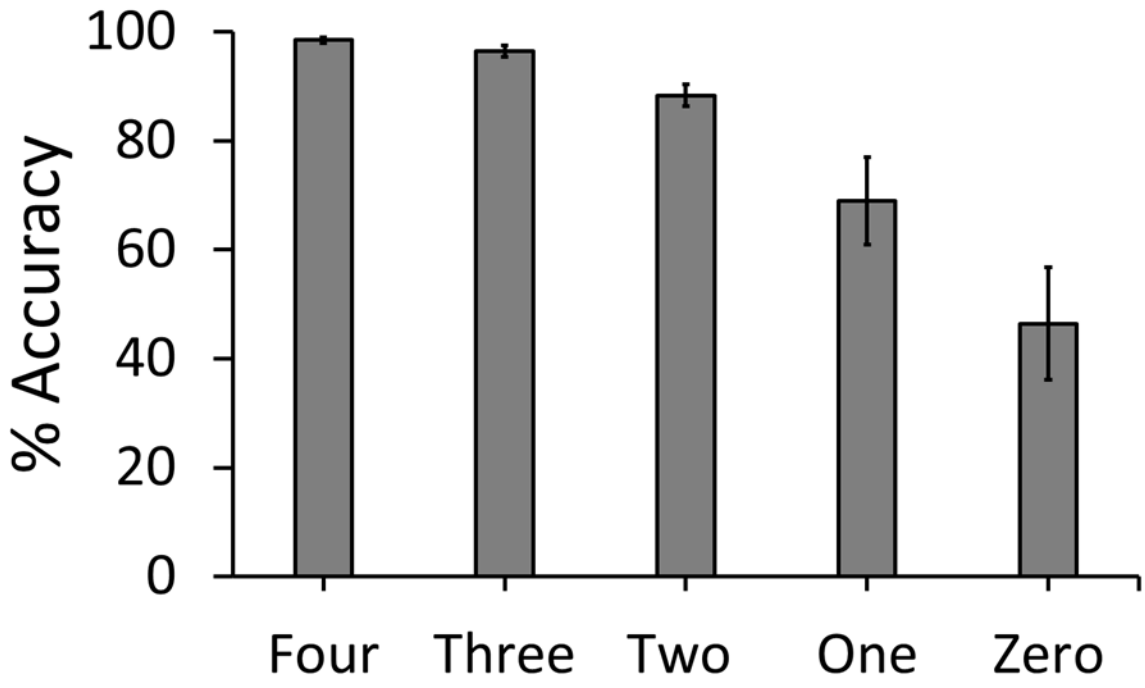

FIGURE 2.7. *Proportional whole face accuracy by number of isolated feature decisions correct for Experiment 2.*

Finally, one sample *t*-tests were also conducted to compare whole face accuracy for each number of features correct to chance performance. Observers performed above chance for whole face pairs when one, two, three or four feature decisions were correct, all ts ≥ 2.29, $p \leq .05$. Whole face accuracy was at chance level when participants classified none of the corresponding features correctly, $t(13) = 0.29$, $p = .78$.

Discussion

This experiment aimed to replicate the findings of Experiment 1 with the addition of a 'don't know' response option, to reduce the need for observers to make a forced choice for facial features that contained limited information for matching decisions. As before, mean accuracy data and correlations indicated that multiple features determine identification of the whole faces, emphasising the importance of understanding how features are combined to reach an overall decision. Similarly to Experiment 1, this experiment demonstrated that the more isolated feature pairs that are classified correctly, the more likely the decision for the corresponding whole face pair is to be correct. This graded pattern for number of features correct suggests that observers may aggregate matching decisions for individual features to reach an overall decision. However, Experiment 2 also corroborated the previous experiment by demonstrating that whole face accuracy remained high even when only one or two feature decisions were correct. This finding suggests that there are cases in which an overall matching decision can be dominated by a single feature, suggesting that information provided by individual features is weighed to reach a decision for the whole face. One possibility is that such key features dominate the overall matching decision by being particularly distinctive, akin to the mechanisms by which distinctive (whole) faces are also better remembered and can be recognised more easily than more average faces (see, e.g., Bartlett,

Hurry, & Thorley, 1984; Schulz, Kaufmann, Walther, & Schweinberger, 2012; Winograd, 1981).

However, the present experiment also found that even when all feature pairs were classified incorrectly (i.e., when there were no distinctive features to aid the overall decision), the whole face was still classified correctly on nearly half of these occasions. The percentage of these trials was low (0.8%) but is surprising considering that the combination of four isolated features should essentially provide the same visual information as the whole face. Despite this, a different decision is made when the same information is displayed together, in whole face pairs, then as individual features. This suggests that the visual context in which the features are viewed, or their integration, is important for unfamiliar face matching. This was investigated directly in Experiment 3.

Experiment 3

Experiments 1 and 2 found that the proportion of whole face correct responses increased with the number of features that were classified correctly in isolation. However, both experiments also showed that even when most feature decisions were incorrect (i.e., only one feature correct), accuracy for the whole face remained at nearly 70%. One possible explanation for this finding is that some features dominate matching decisions in the context of the whole faces, to the point where these can overturn contradictory matching information from several other features.

However, in Experiment 2, observers were also still able to classify the whole face correctly on nearly half of the trials where *all* features were classified incorrectly. This surprising finding suggests that the context in which features are displayed is important, whereby features may be processed differently in the context of a whole face. In the related field of person recognition, a number of studies have demonstrated faces are processed in a

holistic manner, whereby faces are considered in their entirety with all features being processed at once (see, e.g., Davidoff & Donnelly, 1990; Donnelly & Davidoff, 1999; Goffaux & Rossion, 2006; McKone, 2004; Tanaka & Farah, 1993; Tanaka & Sengco, 1997). Evidence of holistic processing in face matching tasks comes, for example, from the composite face effect whereby observers struggle to correctly match top halves of faces depicting the same individual when they are aligned with bottom halves displaying two different individuals (see, e.g., Hole, 1994; Le Grand, Mondloch, Maurer, & Brent, 2004). Face matching is also more impaired when observers are forced to view features individually, compared to when they see the whole face except for the feature they are directly looking at (Van Belle, De Graef, Verfaillie, Busigny, & Rossion, 2010). Thus, an alternative explanation for whole face accuracy being higher than expected when observers classify only a minority of features correctly, is that unfamiliar faces are processed in a holistic manner. Accordingly, featural information may be particularly useful for identification when information from different features is integrated into a complete face percept.

To investigate whether holistic face processing can explain the whole face advantage that was observed in Experiment 1 and 2, this experiment compared identification of whole faces with two new conditions. One condition was comprised of misaligned whole faces, in which the same visual information was presented, but the four isolated features of the preceding experiment were offset horizontally to disrupt holistic processing. The other new condition, termed the misaligned parts condition, was based on the same displays, but only two of the four features were shown. In line with the preceding experiments, performance for the misaligned whole face condition should generally be higher than for misaligned part face displays, by virtue of the fact that a greater number of features are visible. Crucially, however, if face matching is enhanced by the holistic presentation of faces, then whole faces

should outperform the misaligned whole face condition, despite providing identical featural content.

Method

Participants

Twenty-four new individuals (5 male, 19 female) participated in this experiment. Observers had a mean age of 20.71 years (*SD* = 4.55, range: 18-41) and were given course credit or a small fee for their time. All participants were of Caucasian ethnicity and reported normal or corrected-to-normal vision.

Stimuli

One hundred and twenty face pairs (80 from Experiment 1, 40 new pairs) taken from the GUFD were utilised as stimuli for this experiment (see Burton et al., 2010). As in previous experiments, half of the pairs were identity-matches and the other half were identity-mismatches. All pairs were presented in greyscale, depicting a neutral expression and frontal pose, with the background cropped out. The maximum size for a face was 90 x 120 mm and the maximum gap between faces in a pair was 50 mm.

These face pairs were either displayed as they were (whole faces), with the four main features (hair, eyes, nose, mouth) misaligned horizontally, or in the misaligned part face condition, in which only two of the four horizontally-offset features were shown. The hair and nose were offset 20 and 60 pixels to the left respectively from their original positions in the whole face. The eyes and mouth were shifted 70 and 10 pixels to right respectively. In this condition, either the hair and nose feature regions, or the eye and mouth regions were retained. This process created 480 trials (120 whole face, 120 misaligned whole face, and 240 misaligned part face). However, each participant only viewed half of the part face pairs

(either the hair and nose or the eyes and mouth in 120 trials; counterbalanced across observers) and so completed 360 trials over the course of the experiment. Example stimuli are illustrated in Figure 2.8.

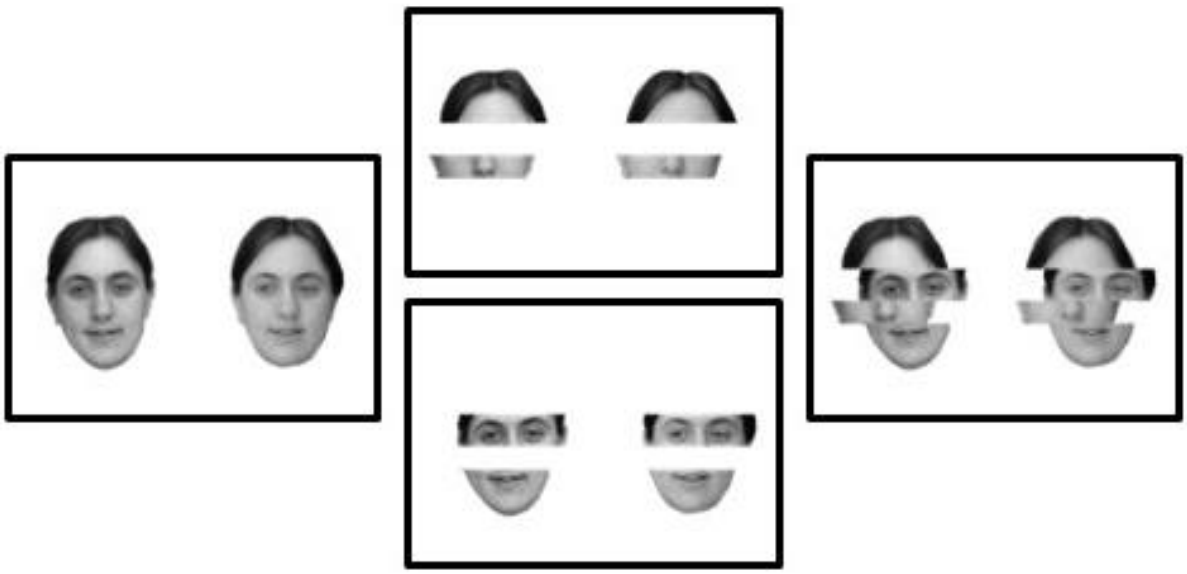

FIGURE 2.8. *Example whole face (left), misaligned part face (centre) and misaligned whole face pairs (right) used as stimuli in Experiment 3. Participants only see one set of the misaligned part face pairs (either hair and nose or eyes and mouth).*

Procedure

The experiment was run using 'PsychoPy' software (Peirce, 2007). Each participant completed 360 trials (120 each of the whole face, misaligned whole face, and misaligned part face condition), which were organised into nine blocks of 40 trials. All stimulus types were intermixed within blocks and displayed in a randomised order. Participants were able to take a short break between blocks if needed. Observers were required to make a match or mismatch decision for each stimulus by pressing one of two response keys on a standard computer keyboard. Participants were told that accuracy was preferred over speed and to take as much time as necessary to complete each trial. Once a decision was made, a blank screen was shown for 500 ms, and then the next stimuli appeared on screen.

Results

The percentage accuracy of participants' responses was analysed as a function of trial type (match vs. mismatch) and face type (whole face vs. misaligned whole face vs. misaligned part face). The cross-subject means of this data are illustrated in Figure 2.9. A 2 (trial type) x 3 (face type) within-subject ANOVA did not reveal a main effect of trial type, $F(1,23) = 0.16$, $p = .70$, $\eta_p^2 = .01$, but a main effect of face type, $F(2,46) = 6.93$, $p < .01$, $\eta_p^2 = .23$, and an interaction between factors, $F(2,46) = 4.81$, $p < .05$, $\eta_p^2 = .17$. Analysis of simple main effects showed that match and mismatch accuracy was comparable in the misaligned whole face condition, $F(1,23) = 0.18$, $p = .68$, $\eta_p^2 = .01$, and the misaligned part face condition, $F(1,23) = 0.80$, $p = .38$, $\eta_p^2 = .38$, but match accuracy was higher than mismatch accuracy for whole faces, $F(1,23) = 8.24$, $p < .01$, $\eta_p^2 = .26$. A simple main effect of face type was also found for match trials, $F(2,22) = 9.39$, $p < .01$, $\eta_p^2 = .46$. A series of paired-sample *t*-tests (with alpha corrected to .05/3 = .017 for three comparisons) revealed higher accuracy for whole faces than misaligned whole faces, $t(23) = 3.83$, $p < .017$, and misaligned part faces, $t(23) = 4.05$, $p < .001$, whereas accuracy was similar for the two misaligned conditions, $t(23) = 0.58$, $p = .57$. By contrast, a simple main effect of face type was not found for mismatch trials, $F(2,22) = 0.03$, $p = .97$, $\eta_p^2 = .00$. These results therefore indicate that face-matching accuracy is best when features are presented as an integrated percept, but that this is evident only for match but not mismatch trials.

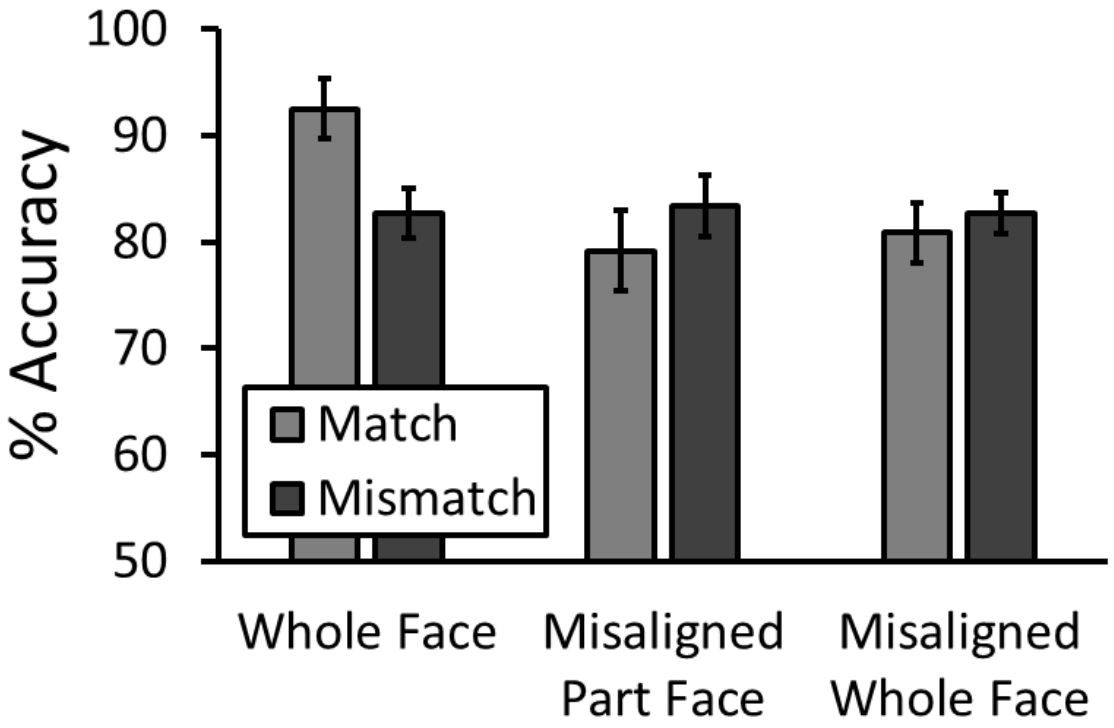

FIGURE 2.9. *Percentage accuracy for the whole face, part face and split face pairs by trial type for Experiment 3.*

To determine if there was a relationship between match and mismatch accuracy, performance for each face type was correlated for these two trials types. Match and mismatch performance was not related for the whole face, $r(22) = -.11$, $p = .62$, misaligned whole face, $r(22) = -.15$, $p = .50$, or misaligned part face, $r(22) = -.10$, $p = .65$. This finding suggests that match and mismatch performance for this task may be driven by dissociable factors. Accuracy for the three different face types was also correlated by trial type. For match trials, performance for the whole face did not correlate with accuracy for the misaligned whole face, $r(22) = .28$, $p = .19$, or the misaligned part face, $r(22) = .30$, $p = .16$. However, performance for the misaligned whole face was correlated with the misaligned part face, $r(22) = .59$, $p < .01$. Similarly, for mismatch trials, whole face accuracy was not associated with accuracy for the misaligned whole face, $r(22) = .36$, $p = .08$, or misaligned part face, $r(22) = .25$, $p = .24$. However, there was a relationship between performance for misaligned whole faces and

misaligned part faces, $r(22) = .65$, $p < .01$. Taken together, these findings indicate that misaligned faces may be processed differently from whole faces.

Discussion

Experiment 3 expanded on the previous two experiments by comparing accuracy of whole faces, misaligned whole faces and misaligned part faces, to determine how the number of features available and the integration of these features relates to overall matching accuracy. For both match and mismatch trials, performance for misaligned whole faces was akin to misaligned part faces. This finding suggests that the additional visual information (features) available in the whole face pairs does not account for the increased accuracy for the whole face pairs compared to the isolated feature pairs seen in Experiments 1 and 2.

For match trials, accuracy was higher for the whole face pairs than the misaligned whole or part face pairs. These results indicate that viewing all features is not sufficient to maximise accuracy, but rather that the integration of these features into a coherent stimulus enhances observers' ability to make correct matching decisions. This effect is likened to the holistic processing advantage for faces that has been reported extensively in the face *recognition* literature (see, e.g., Donnelly & Davidoff, 1999; Goffaux & Rossion, 2006; McKone, 2004; Tanaka & Sengco, 1997). A similar reliance on holistic processing has also been found in face matching tasks using the composite effect and gaze-contingent paradigms (see, e.g., Hole, 1994; Le Grand et al., 2004; Van Belle et al., 2010). Furthermore, performance for misaligned whole faces increased in line with performance for misaligned part faces for both trial types, which suggests they are processed in the same way. However, accuracy for the two misaligned pair types was not associated with whole face accuracy. This finding provides further evidence that observers may use an alternative (holistic) mechanism to process integrated whole faces and so may account for the increase in match accuracy for

the whole face pairs. However, an advantage of viewing an integrated whole face was not found for mismatch trials. Performance for match and mismatch stimuli has been found to be dissociable (see, e.g., Kokje, Bindemann, & Megreya, 2018; Megreya & Burton, 2006b, 2007). Thus, it is possible that these two trial types rely on different mechanisms and that holistic processing is more critical for telling people together than telling people apart.

General Discussion

Unfamiliar face matching is a highly error-prone task (for reviews see, e.g., Fysh & Bindemann, 2017a; Robertson, Middleton et al., 2015) and yet, the decision-making process behind this task is not well understood. One possibility is that matching decisions to pairs of whole faces reflect a series of 'smaller' decisions to individual features, which are then combined to reach an overall decision. The current study investigated this possibility by comparing matching of individual features with that of the whole face. Specifically, this study sought to explore whether the likelihood that the whole face is classified correctly is related to the proportion of its constituent features that are classified in the same way.

In Experiments 1 and 2, a graded response pattern was observed whereby accuracy for whole face pairs improved in line with the number of correct feature decisions. Whole face accuracy was highest when three or four (out of a possible four) feature decisions were classified correctly. This suggests a summing of feature information to achieve an overall matching decision. However, if *all* matching decisions for features are combined in this way to reach a whole face decision, then one would expect overall accuracy to be near chance level when only half of the features are classified correctly. Classifying two features as a match and the other two as a mismatch should provide a conflict such that when viewing the whole face, observers are equally as likely to respond correctly as they are incorrectly. Contrary to this reasoning, observers can still maintain high levels of accuracy for pairs of

whole faces even if only one or two features are matched correctly, which suggests that individual features can correctly determine the overall matching decision. Notably, the cases in which three features are classified correctly suggest that the reverse of this scenario, whereby a single feature decision leads to incorrect classification of whole faces, is rarely found.

So why are observers able to maintain high levels of accuracy for the whole face even if only a minority of the features have been classified correctly? One possibility is that for some faces the decision process may reflect a *weighing* of feature decisions as opposed to a *summing* of these decisions. Accordingly, if one feature provides especially compelling matching information, it may dominate the decision process. For example, it is conceivable the hair, eyes and mouth of two different faces look highly similar, but the noses appear notably different, leading observers to decide these faces ultimately depict different people. On the Matching Familiar Figures Test (MFFT), which correlates with unfamiliar face matching performance (see Megreya & Burton, 2006b), observers must typically rely on a difference in a single feature to discriminate between line drawings of an object (e.g., a chimney on a ferry boat). Thus, to reach the correct decision on the MFFT, participants must base their decision on one feature, which provides conflicting matching information to the rest of the stimulus. Unfamiliar face matching might rely on a similar weighting of decisions for individual features before an overall matching decision is made.

One visual characteristic that might support such a weighing process is distinctiveness, whereby an usual variant of a facial feature may provide more compelling matching information than the other features of a face. Distinctive faces are better remembered and more easily recognised than average looking faces (e.g., Bartlett et al., 1984; Schulz et al., 2012; Winograd, 1981). Furthermore, using caricature to increase the distinctiveness of faces has been found to improve face matching accuracy (McIntyre,

Hancock, Kittler, & Langton, 2013). Thus, it is possible distinctive faces are also easier to match. A decision for one distinctive feature could override decisions to other features, making it possible for observers to still classify the whole face correctly. However, as neither Experiment 1 or 2 indicated a universal feature which drives accuracy, it is possible the distinctive (dominant) feature varies from face to face. This converges with studies that suggest individual variation in facial appearance is highly idiosyncratic (see, e.g., Burton et al., 2011; Burton et al., 2016; Jenkins et al., 2011).

However, in Experiments 1 and 2 accuracy for whole faces also exceeded any of the individual feature pairs. Furthermore, Experiment 2 indicated that even when all of the isolated feature pairs were classified incorrectly, the correct decision for the corresponding whole face could still be reached on half of these occasions. One possibility is that features are integrated into a holistic face percept, which facilitates facial identification. Experiment 3 addressed and explored this possibility directly by contrasting accuracy for whole face pairs in which all features were aligned to form a holistic stimulus or misaligned to disrupt such processing. In this experiment, aligned faces outperformed misaligned whole face pairs on match trials. This converges with the findings of the first two experiments that whole face accuracy is higher than that of any of the individual features and suggests that, while individual features can strongly influence the overall decision, integration of features into a coherent face percept ultimately gives the best possible performance. However, a similar whole face advantage was not found for mismatch trials. Previous research has demonstrated that accuracy for match and mismatch trials may be driven by dissociable factors (see, e.g., Megreya & Burton, 2006b, 2007). This suggests that holistic processing may be more important for pairs of faces which depict the same individual, than pairs which incorporate two different people.

An alternative explanation for high whole face accuracy when only one or two features are classified correctly, is that there is simply more information in the whole face pairs (four features as opposed to one). However, accuracy for whole faces was also reliably higher than accuracy for misaligned faces (Experiment 3). The misaligned face stimuli were designed to force observers to process the face pairs feature by feature as opposed to in a more holistic manner. Therefore, as integration of features appears to be important for successful matching, it is possible that judging the similarity of features in isolation is different to evaluating them in the context of a whole face. In the related field of person *recognition*, a number of studies have demonstrated faces are processed in a holistic manner (see, e.g., Goffaux & Rossion, 2006; Le Grand et al., 2004; Tanaka & Farah, 1993; Tanaka & Sengco, 1997). Asking participants to make decisions to individual features may therefore be forcing them to process the faces atypically and make decisions that are not normally part of the whole face decision process.

In summary, this chapter demonstrated classifying more individual features correctly increases the likelihood of the overall decision being accurate (Experiments 1 and 2). There did not appear to be a single universal feature that drove accuracy. However, a matching decision for single feature can disproportionately influence the overall (whole face) matching decision. Furthermore, as face variation is idiosyncratic (see, e.g., Burton et al., 2011; Burton et al., 2016; Jenkins et al., 2011) it is likely that the dominant feature varies from face to face. Experiment 3 found whole face accuracy was better than misaligned whole face and misaligned part face performance, which converges with whole face accuracy exceeding that of any of the individual features for Experiments 1 and 2. This indicates that, while a decision for one feature can strongly influence the overall decision, an integration of features into a holistic percept helps to maximise performance. However, this whole face advantage was

only found for match trials, and thus it is possible that match trials require a greater reliance on holistic processing than mismatch trials.

Chapter 3

Examples improve face-matching accuracy in low-performing individuals

Introduction

Chapter 2 investigated the decision process that underlies unfamiliar face matching. The previous chapter demonstrated that judgements for individual facial features can strongly influence the overall matching decision (Experiments 1 and 2). However, accuracy is ultimately best when the face is processed as an integrated percept (Experiment 3). This Chapter will address a further question of interest, whether unfamiliar face matching performance can be improved. Despite the ubiquity of this task in security settings, such as passport control, a substantial body of psychological research demonstrates that unfamiliar-face matching is prone to error (for reviews, see, e.g., Fysh & Bindemann, 2017a; Jenkins & Burton, 2011; Robertson, Middleton, & Burton, 2015).

Under idealised conditions, in which observers are asked to compare same-day high-quality photographs of faces, errors are made on 10-20% of trials (see, e.g., Bindemann, Avetisyan, & Blackwell, 2010; Burton, White, & McNeill, 2010; Megreya & Burton, 2006b). These error rates are already considered problematic for large-scale security operations, where a small percentage of errors can result in a large number of cases that give rise to incorrect decisions (Dhir, Singh, Kumar, & Singh, 2010; Jenkins & Burton, 2008b). Accuracy declines further under conditions that are likely to present in applied settings, such as when to-be-compared face photographs were taken many months apart, as is typically the case with a passport photograph and its bearer (Megreya, Sandford, & Burton, 2013), when faces have to be matched over extended time periods (Alenezi & Bindemann, 2013; Alenezi, Bindemann, Fysh, & Johnston, 2015), when operatives are under time pressure to perform this task (Bindemann, Fysh, Cross, & Watts, 2016; Fysh & Bindemann, 2017b; Özbek & Bindemann, 2011; Wirth & Carbon, 2017), and during human supervision of automated facial recognition decisions (Fysh & Bindemann, 2018b; White, Dunn, Schmid, & Kemp, 2015).

These findings highlight that face matching is generally challenging, but are based on measures of mean performance, across groups of participants. In addition, substantial individual differences also exist in this task, which are such that some people perform close-to-chance when others achieve perfect accuracy (see, e.g., Bindemann, Avetisyan, & Rakow, 2012; Burton et al., 2010; Estudillo & Bindemann, 2014; Fysh & Bindemann, 2018a). This range in performance is important for demonstrating that security could be enhanced by selecting individuals with a specific aptitude for face processing (see, e.g., Bobak, Dowsett, & Bate, 2016; Bobak, Hancock, & Bate, 2016; White, Kemp, Jenkins, Matheson, & Burton, 2014). These individual differences indicate also that many face-matching errors do not arise from data limits, whereby stimuli carry insufficient information to allow accurate identifications to be made, but from the failure of some observers to correctly use the available information within some stimuli (see, e.g., Fysh & Bindemann, 2017a; Jenkins & Burton, 2011). In turn, the observation that some people can successfully match the same stimuli is important for indicating that improvements in accuracy for other people, who do not perform to the same level, are in principle possible.

To date, limited research still exists on methods to improve a person's face-matching accuracy, and not all methods procure benefits. Training observers to classify face shapes, for example, does not improve face-matching accuracy (Towler, White, & Kemp, 2014). However, one method that has been shown to improve unfamiliar-face matching is the application of feedback. Real-word scenarios provide very limited scope to correct face-matching errors. Consequently, observers rarely have the opportunity to recognise, and learn from, their own face-matching mistakes (Jenkins & Burton, 2011). Providing such feedback immediately after a face-matching trial can help to maintain accuracy in subsequent trials of this task (Alenezi & Bindemann, 2013), and feedback can *improve* subsequent performance when this is provided whilst a just-classified face pair is still in view (White, Kemp, Jenkins,

& Burton, 2014). However, such trial-by-trial feedback cannot be easily implemented in applied settings, as the accuracy of matching decisions is not known at the point of an identification.

In this study, an alternative form of 'feedback' that could be provided in applied settings was therefore investigated, to determine if this confers improvements in face-matching accuracy. Our approach is based on providing example face-pairs of identity-matches and mismatches, which are clearly labelled as such, to the left and right of a centrally-presented target face-pair. The rationale for this manipulation is that face-matching errors may arise because observers do not have clearly defined *criteria* for distinguishing same- and different-identity face pairs. The observation that trial-by-trial feedback improves accuracy supports this reasoning and suggests that the feedback benefit arises by helping observers to refine their face-matching criteria.

In contrast to the trial-by-trial feedback manipulations of previous studies (Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014), the examples manipulation that is investigated here has greater potential to be implemented in applied settings because it does not require prior knowledge of the nature of the target face pair. To determine if such a benefit is found, observers' face-matching accuracy was first assessed without examples, to obtain a baseline measure of their performance. Examples were then provided in a second block of trials to look for improvement in performance. Observers in this condition were also compared with another group, who were not provided with examples in the second block, on a between-subjects basis.

This group-level comparison provides a useful contrast to assess the *general* impact of examples on face matching performance. However, considering the broad differences that exist in face-matching accuracy between observers (e.g., Bindemann, Avetisyan et al., 2012; Bobak, Pampoulov, & Bate, 2016; Burton et al., 2010; White, Kemp, Jenkins, Matheson, et

al., 2014; for a review, see Lander, Bruce, & Bindemann, 2018), this study was also interested in how examples influenced accuracy at an individual level, by comparing any changes in performance with a person's baseline performance. For example, one might expect that observers with high ability have limited scope for improvement and are therefore unlikely to benefit from examples. On the other hand, it is also possible that observers at the other end of the face-matching ability spectrum lack the capacity to improve. An analysis of individual improvement in the examples condition, compared to baseline performance, should reveal this.

If an improvement in face matching accuracy with the provision of examples is found, then it is also important to determine whether this is transferable to conditions in which examples are no longer seen. To address this question, in Experiments 5 and 6 participants were given two additional blocks that were presented after the examples were withdrawn. One of these blocks comprised a repetition of the target face pairs from the examples block, but these were now shown without the example stimuli. The other block presented new target face pairs, which had not been encountered before in the experiment, but were taken from the same stimulus set (Experiment 5). In addition, this study sought to assess whether any example-advantage would generalise to a completely different new set of face stimuli. For this purpose, a second group was also given a repetition of the target face pairs from the examples block (without the example stimuli), followed by a block of trials from a different face-matching test (Experiment 6). The contrast of these conditions should reveal whether any improvements in performance with examples generalise to previously unseen faces from the stimulus set from which the target and example pairs were drawn, and to face pairs from a different stimulus set.

Experiment 4

This experiment investigated whether the provision of clearly-labelled match and mismatch examples, presented to the left and right, improves classification of centrally-presented target face pairs. Participants first completed a block without such examples to establish a measure of baseline performance. Examples were then provided in a second block in an attempt to improve face-matching accuracy. In addition, a second group of observers completed the experiment without such examples, to determine improvements in accuracy on a between-subject basis.

Method

Participants

Sixty students (57 female, 3 male) from the University of Kent, with a mean age of 21.5 years (*SD* = 6.2; range: 18-49), took part in this experiment. The participants were given course credit for their time. All participants were of Caucasian ethnicity and reported normal or corrected-to-normal vision. All experiments reported here were approved by the Ethics Committee of the School of Psychology at the University of Kent and conducted according to BPS ethical guidelines.

Stimuli

Eighty face pairs from the Glasgow University Face Database (GUFD) were employed as stimuli in this study (see Burton et al., 2010). These were comprised of 40 identity-matches, in which two different same-day photographs of the same person were shown, and 40 identity-mismatches, depicting two different individuals in each pair. All the faces were depicted in greyscale, a frontal pose, and with a neutral expression, and were cropped to remove extraneous background. The maximum size for a face was 43 x 54 mm,

while the maximum gap between faces in a pair was 25 mm. Each face pair was shown beneath the question "Match or Mismatch?".

In the experiment, 40 of these face pairs (20 matches, 20 mismatches) were employed as the stimuli in Block 1, and were then repeated as the centrally-presented target stimuli in Block 2. Repeating the stimuli in this way ensures that any changes in *individual* performance cannot be attributed to variation in stimulus content across blocks. In the experimental condition, another 40 face pairs were presented as example stimuli to the left and right of the target pairs in Block 2. Two example face pairs were provided with each target stimulus and were clearly labelled as identity-matches and mismatches. These example face pairs were randomly selected, but each pair occurred with equal frequency during the experiment. In addition, the sex of the example faces always matched that of the target face pair. For an illustration of a stimulus array, see Figure 3.1.

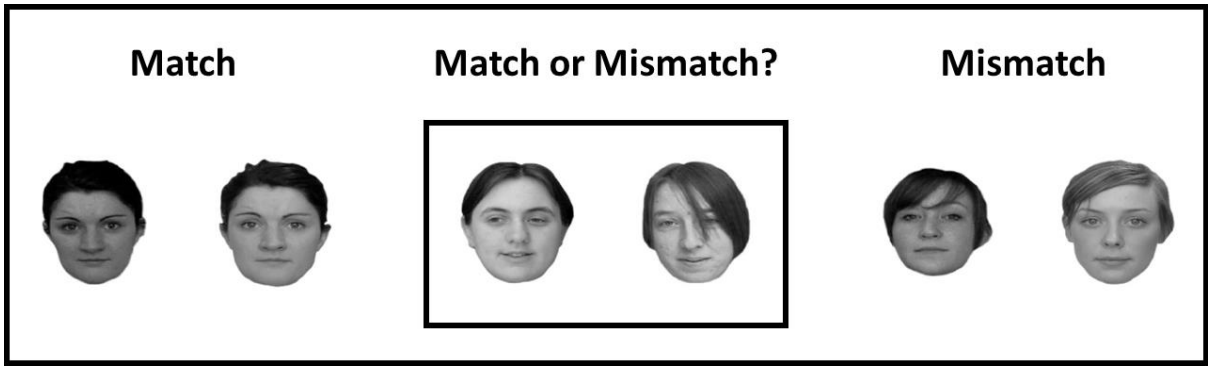

FIGURE 3.1. *Illustration of a stimulus array from Block 2 of the experimental condition, comprising a centrally-presented pair of target faces, and labelled example match and mismatch pairs. In the no-examples condition, the match and mismatch stimuli to the left and right of the target pair were not shown.*

Procedure

The experiment was run using 'PsychoPy' software (Peirce, 2007). Participants were shown two blocks, each containing 20 match and 20 mismatch trials displayed in a randomly intermixed order. Half of the participants (N = 30) were allocated to the no-examples

condition, where the same face pairs were shown in both of these blocks. The remaining participants (N = 30) were assigned to the examples condition, where Block 1 was identical to the no-examples condition, but the target face pairs were flanked by example match and mismatch face pairs in Block 2. All participants were instructed to classify the centrally-presented face pairs as identity-matches or mismatches as accurately as possible, by pressing one of two response keys on a standard computer keyboard. In addition, participants in the experimental condition were given additional instructions prior to Block 2, which explained the presence of the examples and encouraged observers to make use of these to aid their identification decisions. After each trial of Block 2, these participants were also asked to indicate whether they had made use of the examples to aid their last decision, by pressing one of two response keys on the computer keyboard.

Results

Group-level accuracy

Participants' responses indicated that examples were utilised on 26.3% (*SD* = 17.9) of trials, demonstrating that these stimuli were used by observers in an attempt to enhance performance. To determine the effect of examples on face matching, the percentage accuracy of observers' responses was analysed as a function of condition (examples vs. no-examples), block (Block 1 vs. Block 2), and trial type (match vs. mismatch). The cross-subjects means of these data are illustrated in Figure 3.2. A 2 (condition) x 2 (block) x 2 (trial type) mixed-factor ANOVA revealed a three-way interaction between these factors, $F(1,58) = 9.93$, $p < .01$, $\eta_p^2 = .15$. To explore this interaction, separate 2 (block) x 2 (trial type) ANOVAs were conducted for the examples and no-examples conditions.

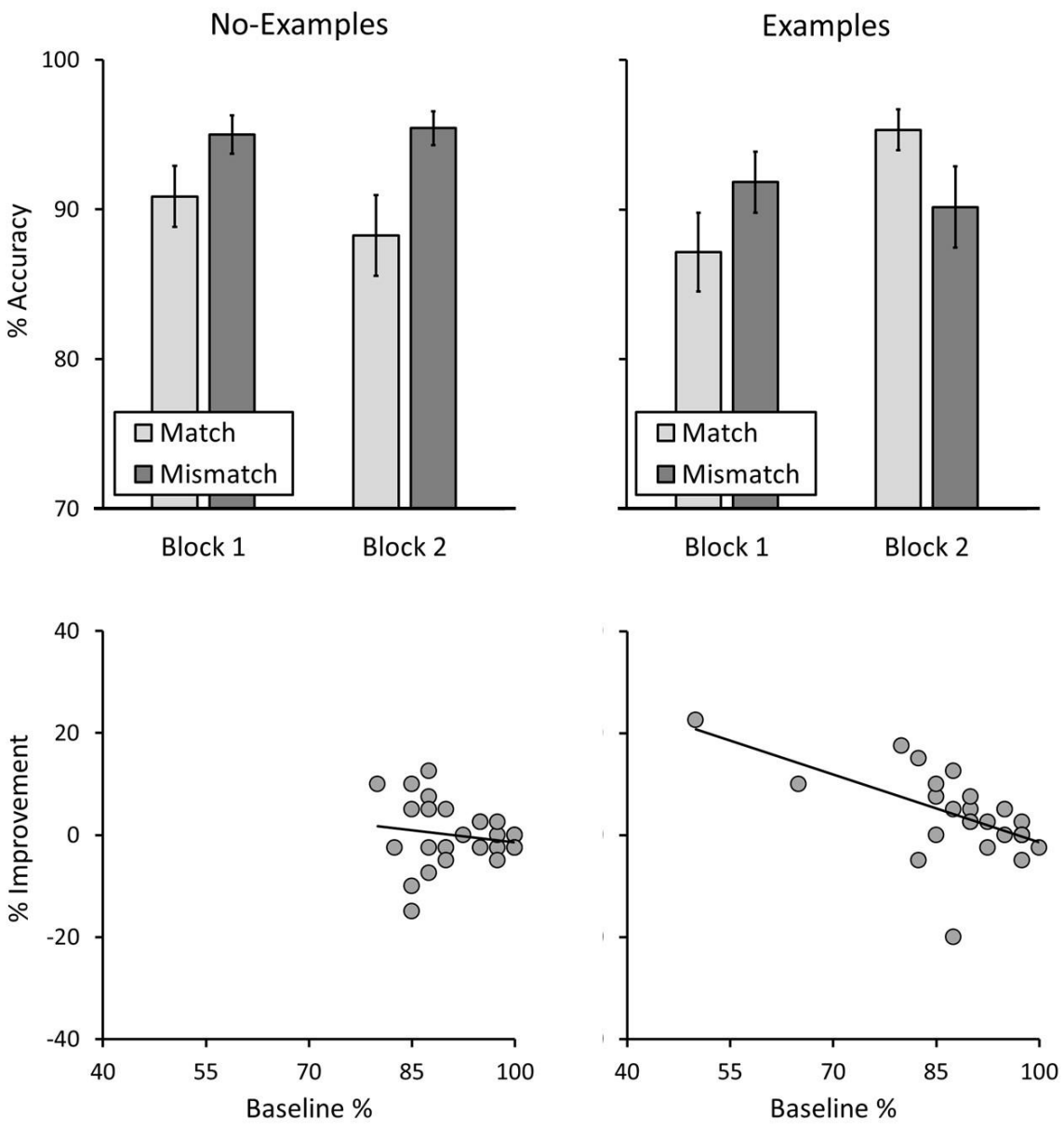

FIGURE 3.2. *Percentage accuracy by trial type (error bars show the standard error of the mean) and overall baseline accuracy correlated with percentage improvement for the no-examples and examples conditions in Experiment 4.*

For the examples condition, no main effect of trial type was found, $F(1,29) = 0.01$, $p = .92$, $\eta_p^2 = .00$, but a main effect of block, $F(1,29) = 5.17$, $p < .05$, $\eta_p^2 = .15$, and an interaction between these factors, $F(1,29) = 15.65$, $p < .001$, $\eta_p^2 = .35$. Analysis of simple main effects showed that accuracy was similar for match and mismatch trials in Block 1, $F(1,29) = 2.94$, $p = .10$, $\eta_p^2 = .09$, and Block 2, $F(1,29) = 3.11$, $p = .09$, $\eta_p^2 = .10$. Mismatch accuracy was also comparable across both blocks, $F(1,29) = 0.96$, $p = .34$, $\eta_p^2 = .03$.

However, an increase in match accuracy was observed from Block 1 to Block 2, $F(1,29) = 15.58$, $p < .001$, $\eta_p^2 = .35$, suggesting an improvement in performance for this trial type.

For the no-examples condition, ANOVA indicated a main effect of trial type, $F(1,29) = 6.67$, $p < .05$, $\eta_p^2 = .19$, due to higher accuracy on mismatch trials. ANOVA did not reveal a main effect of block, $F(1,29) = 0.02$, $p = .88$, $\eta_p^2 = .00$, or an interaction between these factors, $F(1,29) = 0.28$, $p = .60$, $\eta_p^2 = .01$.

Note that analysis of *d'* and *criterion* are omitted here for brevity, but these data are provided for completeness in Appendix B for Experiments 4-6.

Individual differences

The question of main interest was how examples affected performance at an individual level. To analyse individual differences, observers' percentage accuracy in Block 1 was subtracted from Block 2 to provide a measure of change in performance. This score was then correlated with Block 1 to determine whether any improvements in accuracy were related to individual differences in baseline performance. This data is also illustrated in Figure 3.2 and shows that there was no correlation between baseline accuracy and change in performance in the no-examples condition, $r(28) = -.16$, $p = .39$. By contrast, change correlated negatively with baseline accuracy in the examples condition, $r(28) = -.60$, $p < .01$. This indicates that observers who performed with lower accuracy at baseline were more likely to improve with the provision of example face pairs.

These correlations were also conducted separately for match and mismatch trials. For the examples condition, baseline accuracy correlated negatively with change for identity-matches, $r(28) = -.86$, $p < .001$, but not identity-mismatches, $r(28) = .04$, $p = .84$. This demonstrates, once again, that poor-performing observers were more likely to improve under the provision of example face pairs, but indicates that this effect was driven by identity-match

trials. By contrast, a correlation between baseline accuracy and change was not found for identity-matches in the no-examples condition, $r(28) = -.26$, $p = .17$. However, baseline accuracy correlated negatively with change for identity-mismatches, $r(28) = -.45$, $p < .05$.

Discussion

This experiment assessed whether the provision of examples can improve unfamiliar face-matching accuracy, by presenting match and mismatch face pairs either side of central target pairs. At a group level, examples elicited an increase in performance of 8%, but this was only present for identity-matches. These findings were qualified by an analysis of individual differences, which showed that observers did not benefit equally from the provision of examples. Those who already performed with higher accuracy at baseline showed more limited improvement thereafter. By contrast, the observers who displayed lower accuracy at baseline were more likely to improve with the provision of examples. This improvement was observed in overall accuracy but was driven by performance on identity-match trials. Thus, the provision of examples enhances face-matching accuracy in low-performing individuals, but performance gains appear to be limited to decisions confirming that two different face photographs of an individual, depict the same person. However, a similar association between baseline accuracy and performance was also found for the mismatch trials of the no-examples condition, which casts some doubt on the examples effect. A second experiment was therefore conducted to explore these findings.

Experiment 5

In Experiment 4, example stimuli displayed concurrently alongside a target face pair helped improve identity-matching accuracy. However, this improvement was dependent on observers' baseline performance and appeared to be driven mainly by identity-match trials.

The aim of Experiment 5 was to provide a replication to establish the robustness of these effects. In addition, for this manipulation to be useful for applied settings, it is important to assess whether the examples-advantage transfers to different stimuli which have not been previously viewed in conjunction with the examples. Therefore, this study also sought to explore the examples-advantage further, by assessing the generalisability of this effect. Specifically, this study investigated whether any enhancement in accuracy persists only whilst examples are on display, or whether such an effect remains present after examples are removed. If so, then the question arises also of whether such an effect is present only for repetition of target face pairs that were seen previously with examples, or whether it generalises to new, previously-unseen target face pairs. To address these questions, Experiment 5 replicated the design of Experiment 4 but with two additional blocks that were presented after the examples were withdrawn. One of these blocks comprised of a repetition of the target face pairs from the examples block, but these were now shown without the example stimuli. The other block presented new target face pairs, which had not been encountered before in the experiment.

Method

Participants

Sixty-two new individuals (45 female, 17 male) with a mean age of 22.5 years (*SD* = 7.4, range: 18-57), took part in this study. Two of these participants were excluded from analysis for showing limited task engagement, exhibited by repeatedly pressing the same response key (i.e., on more than 25% of consecutive trials in a block) irrespective of trial content *and* producing these responses at very short speeds (i.e., of less than one second). Participants were given course credit or paid a small fee in exchange for their time. All participants were of Caucasian ethnicity and reported normal or corrected-to-normal vision.

Stimuli and procedure

One hundred and twenty face pairs (80 from Experiment 4, 40 new pairs) from the GFMT (see Burton et al., 2010) served as stimuli for this study. As in Experiment 4, half of these stimuli were identity-match and half were identity-mismatch pairs. Experiment 5 progressed in the same way as Experiment 4, except that participants were required to complete two additional blocks after the baseline (Block 1) and the experimental block (Block 2; in which examples or no examples were provided on a between-subject basis). The first additional block was a repetition of the stimuli from Block 1, to determine whether any increases in face-matching accuracy from viewing examples were retained when these were no longer present (i.e., Block Old Faces). The second additional block contained 40 previously unseen face pairs (20 matches and 20 mismatches) from the GFMT and was included to assess whether any performance gains generalised to such new stimuli (Block New Faces). The order of these two new blocks was counterbalanced across participants. Trial order was randomised in all blocks.

Results

Group-level accuracy

Participants' responses indicated that examples were utilised on 25.0% (*SD* = 15.0) of trials. The percentage accuracy of observers' responses was analysed as a function of condition (examples vs. no-examples), trial type (match vs. mismatch), and block (Block 1 vs. Block 2 vs. Block Old Faces vs. Block New Faces). The cross-subjects means of this data are illustrated in Figure 3.3. A 2 (condition) x 2 (trial type) x 4 (block) mixed-factor ANOVA revealed a three-way interaction between factors, $F(3,174) = 5.01$, $p < .01$, $\eta_p^2 = .08$. To analyse this interaction, separate 2 (trial type) x 4 (block) ANOVAs were conducted for the examples and no-examples conditions.

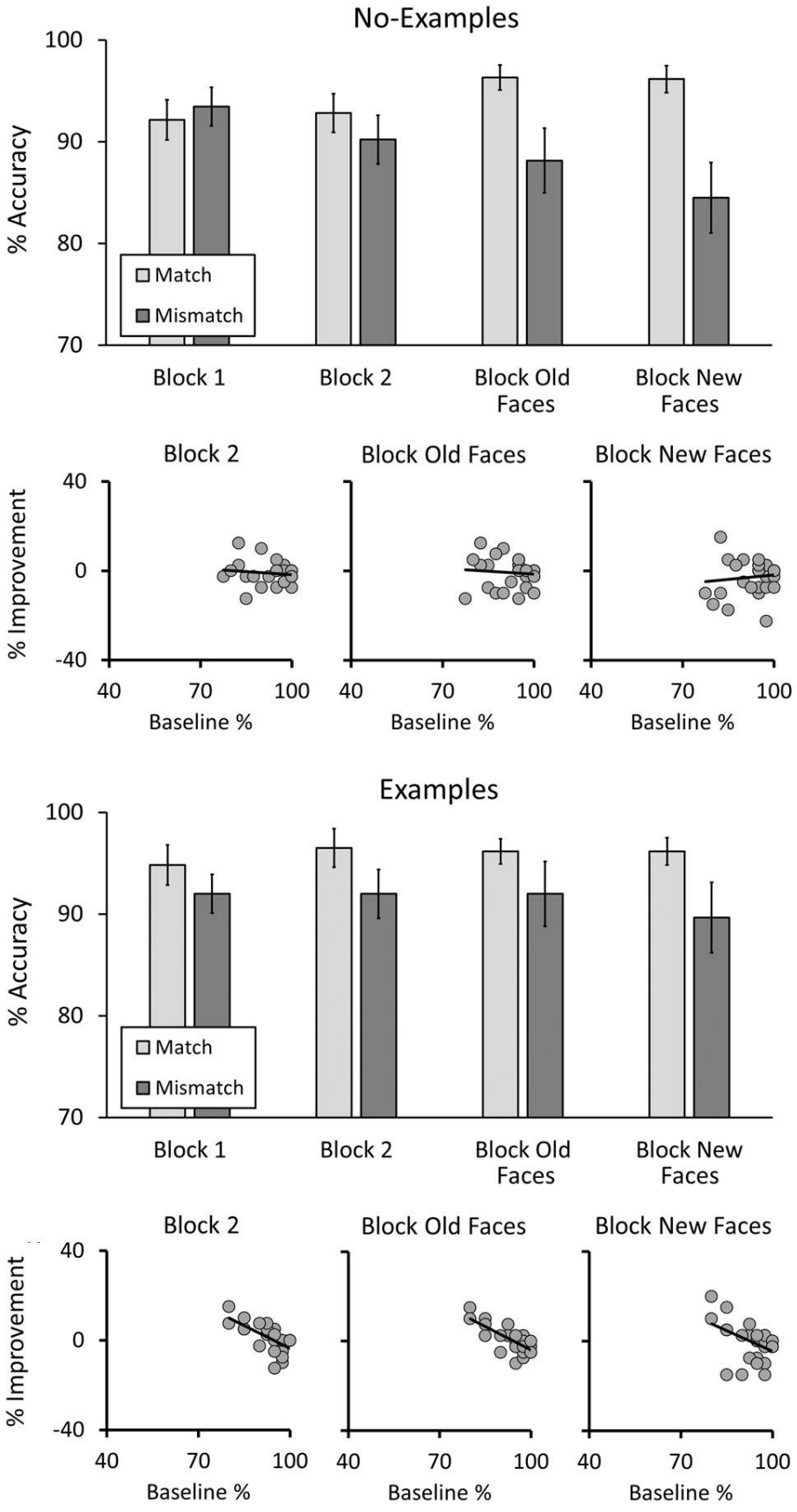

FIGURE 3.3. *Percentage accuracy by trial type (error bars show the standard error of the mean) and overall baseline accuracy correlated with percentage improvement for the no-examples and examples conditions across blocks in Experiment 5.*

For the examples condition, this analysis revealed a main effect of trial type, $F(1,29) = 8.15$, $p < .01$, $\eta_p^2 = .22$, due to higher accuracy on match trials. ANOVA did not reveal a main effect of block, $F(3,87) = 0.62$, $p = .61$, $\eta_p^2 = .02$, or an interaction between factors, $F(3,87) = 0.87$, $p = .46$, $\eta_p^2 = .03$.

In the no-examples condition, no main effect of block, $F(3,87) = 2.03$, $p = .12$, $\eta_p^2 = .07$, or of trial type, $F(1,29) = 2.68$, $p = .11$, $\eta_p^2 = .08$, was found. However, ANOVA revealed an interaction between factors, $F(3,87) = 17.75$, $p < .001$, $\eta_p^2 = .38$. Analysis of simple main effects indicated that there was no difference in accuracy for match and mismatch trials in Block 1, $F(1,29) = 0.53$, $p = .47$, $\eta_p^2 = .02$, and Block 2, $F(1,29) = 0.54$, $p = .47$, $\eta_p^2 = .02$. However, match accuracy was higher than mismatch accuracy in Block Old Faces, $F(1,29) = 5.19$, $p < .05$, $\eta_p^2 = .15$, and Block New Faces, $F(1,29) = 11.37$, $p < .01$, $\eta_p^2 = .28$. In addition, a simple main effect of block was observed for match trials, $F(3,27) = 5.30$, $p < .01$, $\eta_p^2 = .37$. Paired-sample *t*-tests (with alpha corrected to .05/6 = .008 for six comparisons) revealed that match accuracy was lower for Block 1 compared to Block Old Faces, $t(29) = 3.32$, $p < .008$, and Block New Faces, $t(29) = 3.28$, $p < .008$. Match accuracy was also lower for Block 2 than for Block Old Faces, $t(29) = 3.47$, $p < .008$. No other comparisons reached significance, all $ts \leq 2.63$, $ps \geq .01$. A simple main effect of block was also observed for mismatch trials, $F(3,27) = 7.39$, $p < .01$, $\eta_p^2 = .45$, due to higher mismatch accuracy in Block 1 compared to Block 2, $t(29) = 2.94$, $p < .008$, Block Old Faces, $t(29) = 3.43$, $p < .008$, and Block New Faces, $t(29) = 4.23$, $p < .001$. Mismatch accuracy was also higher for Block 2 than Block Old Faces, $t(29) = 3.16$, $p < .008$. No other comparisons reached significance, both $ts \leq 2.33$, $ps \geq .03$.

Individual differences

Once again, in the main step of the analyses, individual differences in performance were assessed by correlating baseline performance (Block 1) with change in accuracy from Block 1 to Block 2 and the subsequent blocks (Block Old Faces / Block New Faces minus Block 1). This data is illustrated in Figure 3.3 and shows that change in Block 2 correlated negatively with baseline accuracy in the examples condition, $r(28) = -.67$, $p < .001$. This indicates that lower-performing observers at baseline improved more with the provision of examples than higher-performing individuals. Similar correlations were also observed for Block Old Faces, $r(28) = -.73$, $p < .001$, and Block New Faces, $r(28) = -.45$, $p < .05$, which suggests that improvements in accuracy were maintained after the removal of the example stimuli and generalised to previously-unseen target face pairs. By contrast, correlations between baseline accuracy and change were not observed in Block 2, $r(28) = -.13$, $p = .51$, Block Old Faces, $r(28) = -.09$, $p = .64$, or Block New Faces, $r(28) = .11$, $p = .56$, of the no-examples condition.

These correlations were also conducted separately for match and mismatch trials. In the examples condition, baseline accuracy for match trials correlated with change for Block 2, $r(28) = -.77$, $p < .001$, and Block Old Faces, $r(28) = -.72$, $p < .001$, and was approaching significance for Block New Faces, $r(28) = -.34$, $p = .06$. Similarly, mismatch accuracy at baseline correlated with change for Block 2, $r(28) = -.74$, $p < .001$, Block Old Faces, $r(28) = -.68$, $p < .001$, and Block New Faces, $r(28) = -.56$, $p < .01$. These results converge with the correlations for overall accuracy to show that lower-performing observers benefitted more from the examples, and this effect persisted when examples were removed and carried over to new match and mismatch face pairs. However, in the no-examples condition, baseline accuracy for match trials also correlated with change for Block 2, $r(28) = -.40$, $p < .05$, Block Old Faces, $r(28) = -.79$, $p < .001$, and Block New Faces, $r(28) = -.74$, $p < .001$. In contrast,

baseline accuracy for mismatch trials did not correlate with change for Block 2, $r(28) = .18$, $p = .34$, or Block New Faces, $r(28) = .35$, $p = .06$, and correlated positively with Block Old Faces, $r(28) = .52$, $p < .01$.

Discussion

Experiment 4 found an improvement in face-matching accuracy with the provision of example stimuli for individuals who displayed lower accuracy at baseline. The current experiment aimed to replicate these findings. In addition, this experiment examined whether the benefits of examples could be retained after their removal, and whether these would generalise to previously unseen face pairs. At a group level, examples did not produce an improvement in performance. However, mismatch accuracy declined, and match accuracy increased in the no-examples condition over the course of the experiment. This pattern converges with reports of a bias to classify face pairs as identity matches, which grows over the course of face matching experiments (Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann et al., 2016; Fysh & Bindemann, 2017b; Papesh, Heisick, & Warner, 2018). This contrast between the examples and the control condition suggest that provision of examples might serve to prevent development of the match bias, similar to the effect of feedback on face matching (see Alenezi & Bindemann, 2013; Papesh et al., 2018). This finding will be explored further in the General Discussion.

More importantly, analysis of individual differences in this experiment showed that, similarly to Experiment 4, examples also helped to improve individuals who performed with low accuracy at baseline. This improvement was found for overall accuracy when examples were present (Block 2), when examples were then removed but the same target stimuli were repeated (Block Old Faces), and for new stimuli taken from the same stimulus set (Block New Faces). This pattern persisted when overall accuracy was broken down into match and

mismatch trials. For mismatch trials, for example, correlational analysis revealed accuracy improvements during the examples block (Block 2) compared to baseline (Block 1), and when examples were then removed (Block Old Faces), and when new stimuli were shown (Block New Faces). Similar correlations were observed for match trials for the examples block (Block 2) and after the examples were removed (Block Old Faces), whilst the correlation between baseline and Block New Faces was approaching significance.

However, not all aspects of the results were clear-cut, as correlations with baseline performance were also found for match trials with the experimental block (Block 2) and for repeated and new stimuli (Block Old / New Faces) in the no-examples group. These correlations likely reflect the increase in match accuracy that was observed at a group level in this condition, which was most pronounced for observers with lower baseline accuracy as these have more scope for such an improvement. At the same time, it is noted that such correlations were not observed with mismatches despite a corresponding decrease in accuracy for this trial type across blocks. The issue of these match-trial correlations in this experiment is complicated by a comparison with Experiment 4, in which baseline accuracy correlated negatively only with improvement for identity-matches in the examples condition, but such a correlation was also observed for identity-mismatches in the no-examples condition.

Taken together, these data show firstly that improvement correlations were consistently found in the examples conditions of Experiment 4 and 5, but the expression of these effects varied somewhat. Secondly, some seemingly similar correlations are observed in the no-examples condition, albeit in fewer measures and blocks, but these effects are also expressed inconsistently across experiments. In an attempt to clarify these results further, a third experiment was conducted.

Experiment 6

In Experiment 4, example match and mismatch pairs that flanked a centrally displayed target pair were found to improve matching accuracy. However, the improvement was dependent on observers' baseline accuracy, with lower-performing individuals benefitting from the examples most. In Experiment 5, the benefits of examples for poor-performing individuals were found to persist after the examples were removed again, and for previously unseen face pairs taken from the same stimuli set. However, the expression of this improvement effect was somewhat inconsistent across both experiments, and a seemingly similar pattern was observed in some measures of the no-examples conditions. The aim of Experiment 6 was, therefore, to provide a further replication of the improvement correlations that were observed with examples in Experiments 4 and 5, and also to replicate the finding that the example-advantage persists once these are removed. Furthermore, this study aimed to assess whether any example-advantage would generalise to a completely different new set of face stimuli. For this purpose, this experiment replicated the design of Experiment 5, but the block of previously unseen stimuli from the GFMT (Block New Faces) was replaced with face pairs from the Kent Face Matching Test (KFMT; see Fysh & Bindemann, 2018a).

Method

Participants

Sixty-two new individuals (40 female, 22 male), with a mean age of 20.4 years (*SD* = 2.1, range: 18-29), took part in this study. Two participants demonstrated limited task engagement by repeatedly pressing the same response key (i.e., on more than 25% of consecutive trials in a block) regardless of trial content *and* produced these responses within one second of viewing the stimulus and thus, were excluded from the analysis. Participants

were given course credit or a small fee for their time. All participants were of Caucasian ethnicity and reported normal or corrected-to-normal vision.

Stimuli and procedure

The design was identical to Experiment 5, comprising a baseline block of 40 matching trials (Block 1), followed by a block in which the same face pairs were flanked by examples (Block 2). This was followed by a third block containing identical stimuli to Block 1 (called Block Old Faces in Experiment 5, but referred to here as Block GFMT). In addition, a further block of 40 trials was included comprising 20 match and 20 mismatch stimuli from the KFMT (see Fysh & Bindemann, 2018a). In contrast to the GFMT face pairs, the KFMT presents a relatively uncontrolled image from student photo-ID alongside a portrait that was recorded under controlled conditions (see Figure 3.4). These different photographs were taken months apart for each identity and vary in terms of, for example, hairstyle, expression, and so forth (for full details, see Fysh & Bindemann, 2018a). In contrast to the face pairs from the GFMT, the KFMT stimuli therefore capture greater variation in appearance within identities. In this experiment, the KFMT portraits were presented at a size of 63 x 65 mm, whereas the photo-ID photographs measured 32 x 36 mm. Both photos were displayed on a blank white canvas, 65 mm apart. In the experiment, Block 1 and Block 2 were always shown first. These were followed by Block GFMT and Block KFMT, but the order of these latter blocks was counterbalanced across participants.

FIGURE 3.4. *Illustration of match (left) and mismatch (right) stimuli from Block KFMT.*

Results

Group-level accuracy

Participants' responses indicated that examples were utilised on 27.2% (*SD* = 22.4) of trials. The percentage accuracy of observers' responses was analysed as a function of condition (examples vs. no-examples), block (Block 1 vs. Block 2 vs. Block GFMT vs. Block KFMT), and trial type (match vs. mismatch). The cross-subjects means of this data are illustrated in Figure 3.5. A 2 (condition) x 2 (trial type) x 4 (block) mixed-factor ANOVA of this data did not reveal a main effect of condition, $F(1,58) = 0.05$, $p = .83$, $\eta_p^2 = .00$, or an interaction between condition and trial type, $F(1,58) = 0.34$, $p = .56$, $\eta_p^2 = .01$, or condition and block, $F(3,174) = 0.88$, $p = .45$, $\eta_p^2 = .02$. Furthermore, a three-way interaction between condition, trial type and block was not found, $F(3,174) = 0.53$, $p = .67$, $\eta_p^2 = .01$. However, when the data was collapsed across both conditions, ANOVA showed main effects of block, $F(3,174) = 534.78$, $p < .001$, $\eta_p^2 = .90$, and trial type, $F(1,58) = 4.73$, $p < .05$, $\eta_p^2 = .08$, and an interaction between these factors, $F(3,174) = 5.70$, $p < .01$, $\eta_p^2 = .09$.

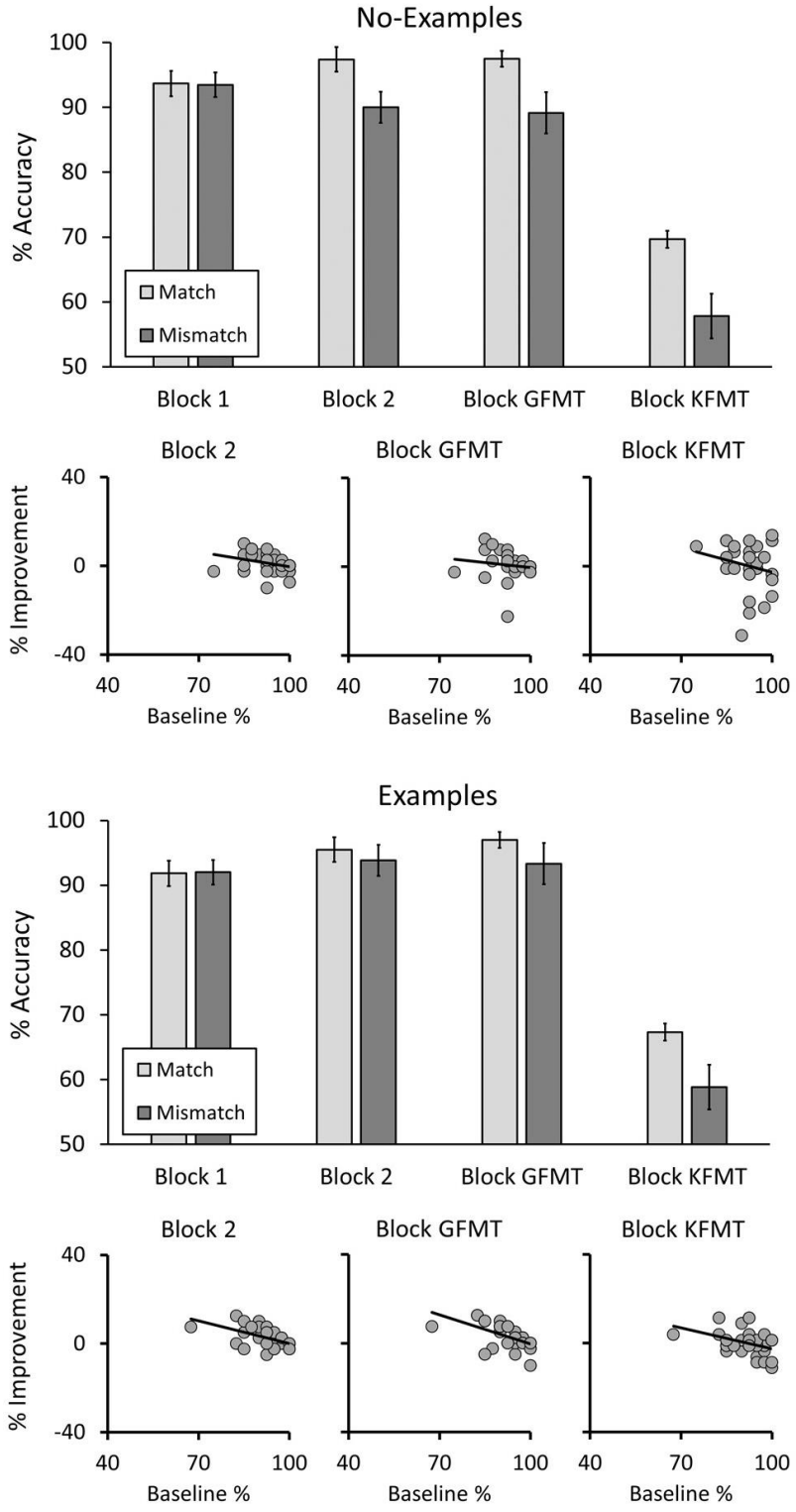

FIGURE 3.5. *Percentage accuracy by trial type (error bars show the standard error of the mean) and overall baseline accuracy correlated with percentage improvement for the no-examples and examples conditions across blocks in Experiment 6. Note that accuracy on the KFMT is consistently lower than the GFMT (see Fysh & Bindemann, 2018a). Therefore, to make the KFMT score compatible with the GFMT scores for the correlational figures, the mean difference between baseline and KFMT accuracy was calculated (28.8%) and then added to each KFMT score.*

Analysis of simple main effects did not reveal a main effect of trial type for Block 1, $F(1,58) = 0.23$, $p = .64$, $\eta_p^2 = .00$. However, the effect of trial type was approaching significance for Block 2, $F(1,58) = 3.69$, $p = .06$, $\eta_p^2 = .06$, and was reliable for Block GFMT, $F(1,58) = 12.03$, $p < .01$, $\eta_p^2 = .17$, and Block KFMT, $F(1,58) = 5.38$, $p < .05$, $\eta_p^2 = .09$, due to higher accuracy on match than mismatch trials.

Analysis of simple main effects also revealed a main effect of block for match trials, $F(3,56) = 60.38$, $p < .001$, $\eta_p^2 = .76$. A series of paired-sample *t*-tests (with alpha corrected to .05/6 = .008 for six comparisons) revealed that match accuracy for Block 1 was lower than for Block 2, $t(59) = 4.51$, $p < .001$, and Block GFMT, $t(59) = 5.36$, $p < .001$, whereas no difference in match accuracy was found between Block 2 and Block GFMT, $t(59) = 2.08$, $p = .04$. These comparisons indicate that match accuracy for face pairs from the GFMT increased after the first block of this experiment. In addition, accuracy for Block KFMT was lower than for Block 1, $t(59) = 11.53$, $p < .001$, Block 2, $t(59) = 13.67$, $p < .001$, and Block GFMT, $t(59) = 13.41$, $p < .001$, reflecting the established greater difficulty of face pairs from the KFMT than the GFMT (see Fysh & Bindemann, 2018a).

A simple main effect of block was also found for mismatch trials, $F(3,56) = 89.48$, $p < .001$, $\eta_p^2 = .83$. Paired-sample *t*-tests (with alpha corrected to .05/6 = .008 for six comparisons) revealed that mismatch accuracy for Block KFMT was lower than for Block 1, $t(59) = 15.24$, $p < .001$, Block 2, $t(59) = 16.79$, $p < .001$, and Block GFMT, $t(59) = 15.06$, $p < .001$. No other comparisons were significant, all ts ≤ 1.31, ps $\geq .20$.

Individual differences

As in previous experiments, the differences in accuracy between Block 1 and all subsequent blocks were calculated to provide a measure of change in performance across blocks. These change scores were then correlated with baseline (Block 1) accuracy (see Figure 3.5). For overall accuracy in the examples condition, baseline performance correlated negatively with change for Block 2 and Block GFMT, $r(28) = -.51$, $p < .01$ and $r(28) = -.55$, $p < .01$, respectively. This indicates that examples most improved the individuals who performed with lower accuracy at baseline. This correlation was also found for Block KFMT, $r(28) = -.40$, $p < .05$, indicating generalization of the examples effect onto face pairs from a different stimulus set.

Once again, these correlations were also conducted separately for match and mismatch trials. In the examples condition, baseline accuracy for match and mismatch trials correlated with change for Block 2, $r(28) = -.60$, $p < .01$ and $r(28) = -.53$, $p < .01$, and for Block GFMT, $r(28) = -.81$, $p < .001$ and $r(28) = -.60$, $p < .001$, again indicating improvement in accuracy with examples which was most pronounced for lower-performing individuals at baseline. However, these correlations for match and mismatch trials were not reliable for Block KFMT, $r(28) = .03$, $p = .89$ and $r(28) = -.21$, $p = .28$, which indicates that generalizability of the examples effect onto stimuli from a different face set was not robust here when this was assessed by trial type.

In the no-examples condition, overall accuracy at baseline did not correlate with change in Block 2, $r(28) = -.30$, $p = .11$, Block GFMT, $r(28) = -.14$, $p = .46$, or Block KFMT, $r(28) = -.20$, $p = .28$. Similarly, a breakdown of this data shows that baseline accuracy on mismatch trials did not correlate with changes in accuracy for Block 2, $r(28) = -.15$, $p = .42$, Block GFMT, $r(28) = -.13$, $p = .51$, or Block KFMT, $r(28) = .04$, $p = .82$. However, such

correlations were observed for match trials on Block 2, $r(28) = -.87$, $p < .001$, and Block GFMT, $r(28) = -.89$, $p < .001$, but not Block KFMT, $r(28) = -.05$, $p = .79$.

Discussion

This experiment aimed to consolidate the findings of Experiments 4 and 5, which suggest that the provision of example stimuli during face matching improves identification accuracy, particularly in lower-performing individuals. In terms of correlations using individuals' overall accuracy, Experiment 6 replicates the examples improvement of the preceding experiments and shows also that the beneficial effect of the examples is retained when the example pairs are no longer displayed. In addition, Experiment 6 also demonstrates that such generalisation extends to an entirely different stimulus set, comprising face pairs from the KFMT (Fysh & Bindemann, 2018a). However, unlike the examples improvement with GFMT stimuli, generalization of improvement onto KFMT stimuli was observed only in overall accuracy, but not when match and mismatch trials were considered separately. Together, these findings consistently indicate the existence of generalization of the examples improvement onto previously unseen stimuli but suggest also that this effect is less robust with stimuli from a different source to the examples, with differing characteristics.

As in Experiment 5, these conclusions are tempered by an inspection of the no-examples group, which also revealed negative correlations between baseline performance and accuracy on all subsequent blocks with GFMT faces. In contrast to the examples condition, these correlations were observed only with match trials and occurred in the context of an increase in match responses at group level, which increased during the experiment. It is therefore possible that the correlations in the no-examples condition arise from a match bias that develops over the course of face matching experiments (see, e.g., Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann et al., 2016; Fysh & Bindemann, 2017b; Papesh et al.,

2018). However, the correlation on match trials of the no-examples condition, and the group-level increase in match responses across blocks, were not observed consistently across all three experiments. It is possible that this arises from the limited sample sizes for the current experiments in conjunction with the large individual differences that are typically observed in face matching experiments (see, e.g., Bindemann, Avetisyan et al., 2012; Bobak, Hancock et al., 2016; Burton et al., 2010). Therefore, an additional analysis was conducted to explore the robustness of these findings, by combining the data of all experiments.

Comparison across experiments

Combining data from all three experiments, the effect of examples on face matching accuracy was first examined at a group level, based on the cross-subject means of observers mean percentage accuracy. A 2 (condition: examples vs. no-examples) x 2 (block: Block 1 vs. Block 2), x 2 (trial type: match vs. mismatch) ANOVA of this data revealed an interaction of block and condition, $F(1,178) = 7.08$, $p < .01$, $\eta_p^2 = .04$. Analysis of simple main effects showed that overall accuracy was similar for the examples condition and the no-examples condition in Block 1 (91.6 % and 92.6%), $F(1,178) = 0.81$, $p = .37$, $\eta_p^2 = .01$, and Block 2 (93.9% and 92.6%), $F(1,178) = 1.58$, $p = .21$, $\eta_p^2 = .01$, and was also comparable across blocks in the no-examples condition (92.6% and 92.6%), $F(1,178) = 0.00$, $p = 1.00$, $\eta_p^2 = .00$. However, accuracy increased from Block 1 to Block 2 in the examples condition (91.6% and 93.9%), $F(1,178) = 14.16$, $p < .001$, $\eta_p^2 = .07$.

ANOVA also revealed a main effect of block, $F(1,178) = 7.08$, $p < .01$, $\eta_p^2 = .04$, and an interaction between block and trial type, $F(1,178) = 17.37$, $p < .001$, $\eta_p^2 = .09$. This interaction reflects that match and mismatch accuracy was comparable for Block 1 (91.1% and 93.1%), $F(1,178) = 2.86$, $p = .09$, $\eta_p^2 = .02$, and mismatch accuracy was also similar across both blocks (93.1% and 92.2%), $F(1,178) = 1.91$, $p = .17$, $\eta_p^2 = .01$. In contrast, match

accuracy increased from Block 1 to Block 2 (91.1% and 94.3%), $F(1,178) = 21.52$, $p < .001$, $\eta_p^2 = .11$, and the difference in accuracy between match and mismatch trials in Block 2 was also approaching significance for Block 2 (94.3% and 92.2%), $F(1,178) = 3.37$, $p = .07$, $\eta_p^2 = .02$, due to higher accuracy on match trials. Overall, these analyses therefore demonstrate that two separable effects occurred across all experiments at a group level. The first presents an improvement in face-matching accuracy with the provision of examples. The second suggests that the proportion of correct match responses increased across blocks over the course of the experiments. None of the remaining main effects or interactions were significant, all $Fs \leq 2.19$, $ps \geq .14$, $\eta_p^2 \leq .01$.

The next step of this analysis sought to explore the role of individual differences in face matching, by repeating the correlational analysis for the full sample of participants, pooled across all three experiments here. This analysis was motivated by variation in the pattern of significant correlations that were observed across experiments. A summary of these correlations is provided in Table 3.1. Firstly, this data reveals some notable consistencies. For example, in each of the three experiments, improvement correlations were *always* observed for overall accuracy in Block 2 of the examples condition. At the same time, such correlations were *never* observed in Block 2 of the no-examples condition. This data therefore provides converging evidence that the provision of example stimuli improves face-matching accuracy, particularly in lower-performing individuals. In addition, similar correlations in overall accuracy were also observed in the examples condition when GFMT face pairs were repeated without examples in Experiment 5 (i.e., in Block Old Faces) and Experiment 6 (Block GFMT), when previously unseen faces from the GFMT were presented in Experiment 5 (Block New Faces), and when faces from a different stimulus set were presented in Experiment 6 (Block KFMT). Again, the same correlations were always absent

in the corresponding blocks of the no-examples condition, thus further strengthening the conclusion that examples improved face-matching accuracy.

Whilst the overall accuracy data presents a clear picture, the pattern of effects is more complex when the data is broken down into match and mismatch trials (see Table 3.1). This shows, for example, that an improvement correlation was observed in Block 2 of the examples condition of Experiment 4 for match trials, but not for mismatch trials. In contrast, these correlations were significant for both trial types in Experiment 5 and 6. In addition, some correlations were also observed in the no-examples condition, such as for identity-mismatches in Block 2 of Experiment 4 and for matches in Block 2 of Experiments 5 and 6. At present, it is not possible to explain this variation in effects. However, face matching is characterised by very broad individual differences in accuracy among observers (see, e.g., Burton et al., 2010; Fysh & Bindemann, 2018a; White, Kemp, Jenkins, Matheson, et al., 2014), as well as inconsistent responding on a block-by-block basis when the same individuals are tested repeatedly (e.g., Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann, Avetisyan et al., 2012, Bindemann et al., 2016; Fysh & Bindemann, 2017b). These inter- and intra-individual differences may underlie the variation in correlations that was observed here, particularly considering that sample size was limited for each experiment (with N = 30 per condition).

Experiment 1		Block 2
Examples	Overall	*r* (28) = -.60 **
	Matches	*r* (28) = -.86 ***
	Mismatches	*r* (28) = .04
No-Examples	Overall	*r* (28) = -.16
	Matches	*r* (28) = -.26
	Mismatches	*r* (28) = -.45 *

Experiment 2		Block 2	Block Old Faces	Block New Faces
Examples	Overall	*r* (28) = -.67 ***	*r* (28) = -.73 ***	*r* (28) = -.45 *
	Matches	*r* (28) = -.77 ***	*r* (28) = -.72 ***	*r* (28) = -.34
	Mismatches	*r* (28) = -.74 ***	*r* (28) = -.68 ***	*r* (28) = -.56 **
No-Examples	Overall	*r* (28) = -.13	*r* (28) = -.09	*r* (28) = .11
	Matches	*r* (28) = -.40 *	*r* (28) = -.79 ***	*r* (28) = -.74 ***
	Mismatches	*r* (28) = .18	*r* (28) = .52 **	*r* (28) = .35

Experiment 3		Block 2	Block GFMT	Block KFMT
Examples	Overall	*r* (28) = -.51 **	*r* (28) = -.55 **	*r* (28) = -.40 *
	Matches	*r* (28) = -.60 **	*r* (28) = -.81 **	*r* (28) = .03
	Mismatches	*r* (28) = -.53 **	*r* (28) = -.60 ***	*r* (28) = -.21
No-Examples	Overall	*r* (28) = -.30	*r* (28) = -.14	*r* (28) = -.20
	Matches	*r* (28) = -.87 ***	*r* (28) = -.89 ***	*r* (28) = -.05
	Mismatches	*r* (28) = -.15	*r* (28) = -.13	*r* (28) = .04

Overall		Block 2			
		All	*W/o outliers*	*Lower accuracy*	*Higher accuracy*
Examples	Overall	*r* (88) = -.60 ***	*r* (85) = -.53 ***	*r* (42) = -.45 **	*r* (28) = .06
	Matches	*r* (88) = -.78 ***	*r* (85) = -.86 ***	*r* (42) = -.75 ***	*r* (28) = -.48 **
	Mismatches	*r* (88) = -.39 ***	*r* (85) = -.330 **	*r* (42) = -.30 *	*r* (28) = .09
No-Examples	Overall	*r* (88) = -.19	*r* (88) = -.19	*r* (43) = -.01	*r* (43) = -.28
	Matches	*r* (88) = -.38 ***	*r* (88) = -.38 ***	*r* (43) = -.31 *	*r* (43) = -.43 **
	Mismatches	*r* (88) = -.10	*r* (88) = -.10	*r* (43) = -.12	*r* (43) = -.03

* = $p < .05$, ** = $p < .01$, *** = $p < .001$

TABLE 3.1. *Summary of correlations for Experiments 4 to 6, and for the combined data for these experiments, with and without potential outliers (all and w/o outliers), and for median-split data to show correlations for the worst (lower accuracy) and best performers (higher accuracy) at baseline.*

To explore this issue further, the data from Block 2 was collated across experiments (see Table 3.1 and Figure 3.6). For the examples condition, this analysis reveals clear improvement correlations for overall accuracy as well as for match and mismatch trials. Inspection of Figure 3.6 suggests the presence of three potential outliers (i.e., accuracy lower

than 70%), but the pattern of correlations remains the same when these data points are removed (see Table 3.1). This study also sought to explore further whether these correlations are genuinely driven by improvement in the lower-performing observers at baseline, or whether these could be attributed, at least in part, to high performing observers. This is plausible considering that the best-performing observers were at or near-ceiling. Thus, these observers can essentially only maintain their baseline accuracy level in Block 2, or drop below this level, which could potentially underpin the negative correlations that were observed here. Crucially, however, such a pattern would contradict the conclusion that examples *improved* performance. To investigate this possibility, participants were sorted by their baseline accuracy and conducted a median split on these data. The correlations were then repeated for observers with the lower baseline accuracy and the higher baseline accuracy. The outcome of this analysis is also displayed in Table 3.1 and shows clearly that the improvement correlations were driven by observers with lower baseline accuracy. Taken together, the analyses presented here therefore provide clear evidence that examples improved face-matching accuracy in the current experiments, particularly in lower-performing individuals.

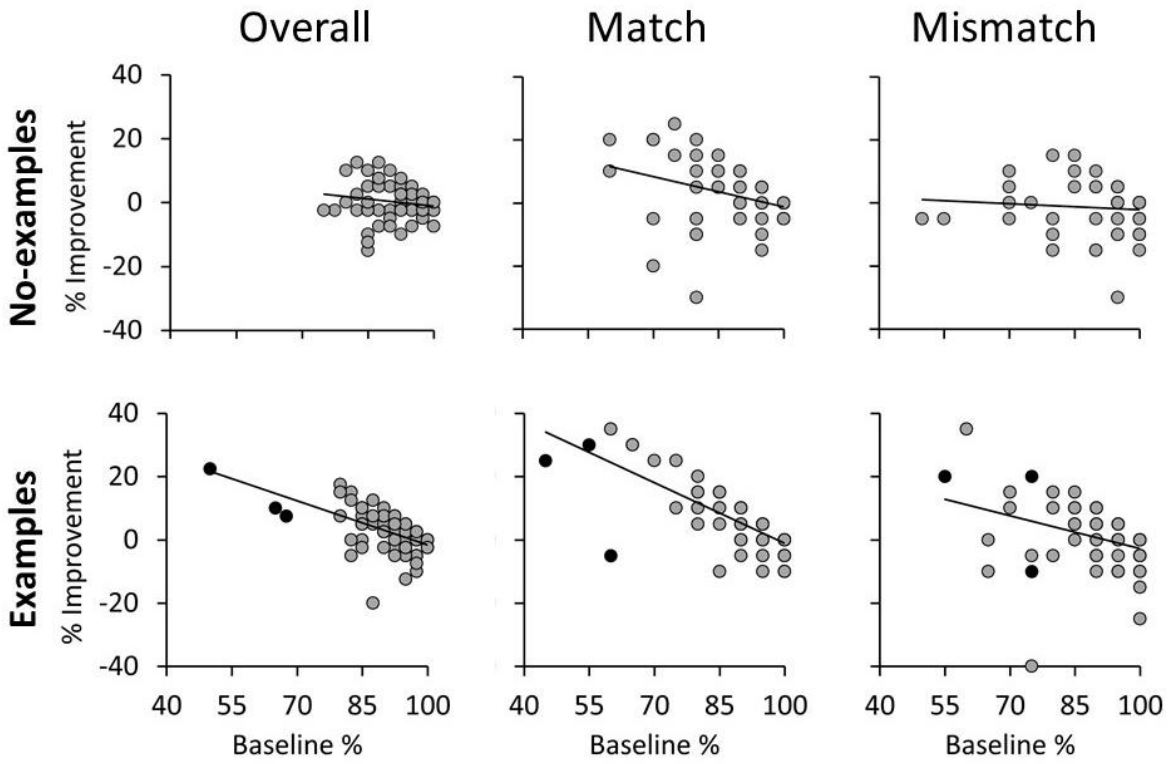

FIGURE 3.6. *Baseline accuracy correlated with improvement from Block 1 to Block 2 collapsed across all three experiments. Black markers denote potential outliers (observers who scored less than 70% overall at baseline).*

The pooling of data across experiments also serves to clarify the variation in significant and non-significant correlations that were observed in the no-examples condition across individual experiments. Analysis of the pooled data reveals consistently negative correlations between baseline performance and change in accuracy on subsequent blocks for match trials. Contrary to the examples condition, however, such correlations were not present for mismatch trials and overall accuracy. In the context of the block by trial type interaction at a group level, which was driven by an increase in accuracy during the experiment on match trials, the correlations on match trials may be attributed to a tendency to record increasingly more correct match responses over the course of the experiment. There is logically more scope for the lower-performing observers from the baseline block to record such an increase in Block 2, thus leading to the negative correlations that were observed in the no-examples condition here. A similar increase in match responses over the course of experiments has now

been observed in several studies (Alenezi & Bindemann, 2013; Bindemann et al., 2016; Fysh & Bindemann, 2017b; Papesh et al., 2018), though the cause of this effect remains unclear (see Alenezi et al., 2015).

General Discussion

Unfamiliar face matching is difficult and error-prone (see, e.g., Bindemann, Avetisyan et al., 2012; Burton et al., 2010; Henderson, Bruce, & Burton, 2001; Megreya & Burton, 2006b), even for experienced professionals who perform this task routinely (e.g., White, Kemp, Jenkins, Matheson et al., 2014; White, Philips, Hahn, Hill, & O'Toole, 2015). Therefore, this study examined whether matching accuracy can be improved by the provision of example face pairs, which were displayed either side of a target face pair and clearly-labelled as identity-matches or mismatches. Examples improved performance at a group level when data was pooled across all three experiments presented here. Correlational analyses revealed that examples also improved individual performance, but particularly in observers who were least accurate at the beginning of the experiment. This improvement was observed consistently in overall accuracy in each of the three experiments reported here, and also in match and mismatch trials when data was collated across all experiments to boost sample size for correlational analysis. In addition, this improvement generalised to new face pairs when these were taken from the same stimulus set as the examples (the GFMT, in Experiment 5) and, to a lesser extent, to face pairs from a different stimulus set (the KFMT in Experiment 6).

How might examples help to improve face-matching performance? One possibility is that observers have limited a priori criteria for knowing what constitutes an identity-match or mismatch. This idea is compelling considering that faces display considerable within-person variation in appearance (Jenkins, White, Van Montfort, & Burton, 2011). As a consequence,

identification rates can vary for different face pairs of the same person, even when these are acquired on the same day but with different image-capture methods (Bindemann & Sandford, 2011). Thus, face matching could be characterised as a rather 'fluid' task, in which different criteria might be required to make an identification depending on the images at hand, even when these depict the same person. In turn, this implies that one source of *poor* performance on this task could be the criteria that observers have available to solve it.

A range of evidence might support such an account. Firstly, poor performance in unfamiliar-face matching indicates that many individuals have a limited idea of what constitutes a match or mismatch (see, e.g., Bindemann, Avetisyan et al., 2010; Megreya & Burton, 2008). However, *some* individuals also perform exceptionally well at matching tasks (see, e.g., Bindemann, Avetisyan et al., 2012; Bobak, Dowsett et al., 2016; Bobak, Hancock et al., 2016; Bobak, Pampoulov et al., 2016; Burton et al., 2010; Estudillo & Bindemann, 2014; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). This indicates a possible resource limit problem, whereby to-be-matched face pairs contain the necessary visual information to make an identification, but some individuals simply do not know how to utilize this information effectively (see, e.g., Fysh & Bindemann, 2017a; Jenkins & Burton, 2011).

Secondly, many observers appear to have limited insight into their own ability to process unfamiliar faces (see, e.g., Bindemann, Attard, & Johnson, 2014; Bobak, Mileva, & Hancock, 2018; Palermo et al., 2017), which also makes it likely that they possess inadequate criteria for identification. This notion receives further support from the phenomenon of choice blindness, which shows that individuals will unwittingly justify a matching decision for face pairs, even if they did, in fact, originally make a different identification decision to the same stimuli (Sauerland et al., 2016). This phenomenon should simply not present if people had definitive, stable criteria for categorizing identity-match and mismatch face pairs.

In turn, providing criteria of some kind seems to help improve this task. For example, trial-by-trial feedback for an observer's responses can improve their face-matching accuracy (White, Kemp, Jenkins, & Burton, 2014), but only if this feedback is delivered whilst the stimuli remain on display (c.f., Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014). This suggests that such feedback might work by enhancing participants' face-matching criteria. However, such feedback can only be administered if the nature of a face pair (i.e., as an identity-match or mismatch) is already known, which is not typically the case in applied settings at the point of identification. Thus, it is possible that the provision of examples might enhance face-matching accuracy in a similar manner, by providing information about what constitutes an identity-match or mismatch, but with the advantage that this is not dependent on prior knowledge of the nature of a target face pair.

The current experiments show that the examples-advantage generalises to face pairs from the same set (i.e., the GFMT, Burton et al., 2010) even after the examples are no longer presented. This indicates that the face-matching criteria that are acquired from the examples are internalised by observers and can continue to improve performance after their removal. At the same time, such generalisation was more limited for face pairs from a different stimulus set (the KFMT; see Fysh & Bindemann, 2018a). This is a potential limitation if one were to consider the provision of examples for improving face-matching performance in applied settings, such as passport control, where observers would inevitably encounter faces from a broad range of types and ethnicities. It is noted, however, that other methods currently under investigation for improving face-matching accuracy, such as feature comparison strategies (Towler, White, & Kemp, 2017) and feature instructions (Megreya & Bindemann, 2018), also show limited generalisation to other stimulus sets. One possible explanation for this finding is that the criteria that are required for identity-matching vary across different face sets. In support of this reasoning, it is already known that different visual features carry identity

information in faces of different races (see, e.g., Hills, Cooper, & Pake, 2013; Hills & Pake, 2013). If the same applies to different face sets of the same race, then this could explain why reduced generalisation is found for the KFMT faces in Experiment 6. To this point, it is noted that the GFMT and KFMT differ in construction considerably, whereby one test offers same-day face pairs that are optimised for identity-matching, whereas the other comprises more variable stimuli that were taken months apart (c.f., Burton et al., 2010; Fysh & Bindemann, 2018a).

In conclusion, the current chapter shows that the provision of examples improves face-matching accuracy, particularly in lower-performing individuals. This examples-advantage persists after these are removed from view and generalises to previously unseen face pairs that are drawn from the same and a different stimulus set. Therefore, these findings suggest that examples aid performance by providing criteria to distinguish identity-matches and mismatches that observers would otherwise have to deduce by their own judgement during face matching.

Chapter 4

Understanding the examples advantage: An eye-tracking investigation

Introduction

Chapter 3 introduced examples as a method of improving unfamiliar face matching accuracy by providing observers with a kind of simultaneous feedback, without the need for knowledge of the correct decision for a given pair. Provision of example matches and mismatches improved identification accuracy for target face pairs (Experiment 4), especially for observers who initially performed poorly at this task. This examples-advantage was maintained after examples were removed, generalised to new face pairs from the same stimulus set (Experiment 5), and demonstrated generalisation to face pairs from a different stimuli set (Experiment 6). However, although observers improved with the provision of examples, *self-reported* example usage was low across experiments (approximately 25% of trials) and did not show a clear relationship with accuracy on a trial-by-trial basis. Furthermore, self-reported example usage was not associated with individual improvement in face-matching performance ($r(88) = .05$, $p = .65$).

Observers have limited insight into their internal cognitive processes (Nisbett & Wilson, 1977; Wilson & Dunn, 2004). For example, participants demonstrate poor awareness of their eye-movements (see, e.g., Clarke, Mahon, Irvine, & Hunt, 2017; Mahon, Clarke, & Hunt, 2018) and struggle to accurately report where they have looked previously (see, e.g., Kok, Aizenman, Võ, & Wolfe, 2017; Võ, Aizenman, & Wolfe, 2016). If observers show a similar lack of insight for their viewing and processing of the example pairs, it is likely that the self-report measure of Chapter 3 did not accurately capture observers' example usage. Furthermore, the task instructions implied that examples should be utilised for particularly difficult trials, hence participants may display demand characteristics by reporting the use of examples when they found a trial particularly challenging, regardless of their *actual* usage. Another possibility is that the definition of what constituted a 'use' of example pairs may have varied across observers. For instance, some participants may have classed the mere

viewing of examples as usage, whilst for others, a use may have constituted actively using information provided by the examples in their target identification. Thus, the question arises of whether we can gain a clearer understanding of how examples impact face-matching performance with a more direct measure of example usage. In this chapter, eye-tracking is employed to provide such a measure.

Eye-tracking gives a direct measure of observers' looking behaviour and can be used to monitor ongoing cognitive processing (see, e.g., Henderson, 2003, 2007; Rayner, 1998). Eye-tracking has been used in numerous studies of face processing (e.g., Fletcher, Butavicius, & Lee, 2008; Heisz & Shore, 2008; Luria & Strauss, 1978; Smilek, Birmingham, Cameron, Bischof, & Kingstone, 2006; Walker-Smith, Gale, & Findlay, 2013). Studying eye movements has given insights into face perception (e.g., Bindemann, Scheepers, & Burton, 2009; Blais, Jack, Scheepers, Fiset, & Caldara, 2008), face learning (e.g., Estudillo & Bindemann, 2017; Millen, Hope, Hillstrom, & Vrij, 2017), face detection (e.g., Bindemann, Scheepers, Ferguson, & Burton, 2010; Crouzet, Kirchner, & Thorpe, 2010) and face memory (e.g., Althoff & Cohen, 1999; Henderson, Williams, & Falk, 2005). Eye-tracking has also informed face viewing strategies, such as the order in which facial features are viewed (e.g., Bindemann et al., 2009), how pairs of faces are fixated when a matching decision is required (e.g., Özbek & Bindemann, 2011), how a target is selected in a line-up (e.g., Mansour & Flowe, 2010) and where observers look in scenes with people present (e.g., Birmingham, Bischof, & Kingstone, 2008a, 2008b, 2009). Thus, it is likely that eye-tracking can also be used to determine how observers' view and utilise examples in an unfamiliar face matching task and so give a more sensitive measure of how examples are used over the course of the experiment.

A second question arising from the experiments reported in Chapter 3 is how much does the *nature* of the examples provided influence the improvement effect found? In the

previous chapter, it was suggested that examples may improve accuracy by helping individuals to solidify the criteria they use to make matching decisions. Observers appear to have limited insight into their own face processing capabilities (Bindemann, Attard, & Johnston, 2014; Bobak, Mileva, & Hancock, 2018; Bobak, Pampoulov, & Bate, 2016; Palermo et al., 2017, but see Livingston & Shah, 2017; Ventura, Livingston, & Shah, 2018), and therefore may have inadequate matching criteria for discriminating match and mismatch trials (see, e.g., Lander, Bruce, & Bindemann, 2018). Thus, example pairs that help shape these criteria are likely to be the most effective for this task.

In the preceding experiments, the difficulty of example face pairs was not manipulated systematically. However, there is good reason to assume that this may impact on the task, as faces can display considerable within-person variation (Jenkins, White, Van Montfort, & Burton, 2011). Consequently, matching accuracy of images of the same person can be highly variable, even if the photos are taken on the same day but using different cameras (Bindemann & Sandford, 2011). Moreover, viewing more varied images of individuals has been shown to improve learning of unfamiliar faces (e.g., Burton, 2013; Burton, Kramer, Ritchie, & Jenkins, 2016; Ritchie & Burton, 2017). It is therefore likely that example match pairs with greater variation between images of the same individual will be more useful for this experiment. In turn, for unfamiliar face matching tasks, mismatch pairs are typically selected by finding individuals within the stimuli set who look similar to each other (see, e.g., Burton, White, & McNeill, 2010; Fysh & Bindemann, 2018a). This practice reflects how in a real-life matching scenario, impostors will aim to make themselves as close in appearance to their falsely obtained photo IDs as possible. However, variation exists in the extent to which such mismatches can be achieved (see, e.g., by-item mismatch data in Fysh & Bindemann, 2018a). Mismatch pairs that are of high similarity should also be more useful for refining observers' existing matching criteria, by illustrating more ways in which the faces of

different people can appear very similar, as well as more subtle differences between persons, than mismatches that are of lesser difficulty.

In addition to the self-report measure of example usage employed in the previous chapter, this study utilised eye-tracking to assess how examples are used to help improve unfamiliar face matching. Eye-tracking is a more sensitive measure than self-report so should help to clarify how examples are utilised for this task. Moreover, viewing time appears to relate to matching accuracy, whereby more prolonged viewing improves matching decisions (see, e.g., Özbek & Bindemann, 2011; Fysh & Bindemann, 2017b). Thus, it is expected that the duration of the example fixations should reflect accuracy such that the more time spent looking at examples, the greater the likelihood of task improvement. To further explore whether the nature of examples provided impacts unfamiliar face matching improvement, observers were provided with either low-difficulty examples (match pairs with little variation, greater dissimilarity between mismatch pairs) or high-difficulty examples (increased variation between match pairs, mismatch pairs consisting of more similar looking individuals). A further group of observers did not view the example pairs to provide an accuracy baseline.

Experiment 7

Method

Participants

Ninety individuals (71 female, 19 male) from the University of Kent, with a mean age of 20.5 years (*SD* = 3.8; range: 18-39), took part in this experiment. The participants were given course credit or a small fee for their time. All participants were of Caucasian ethnicity and reported normal or corrected-to-normal vision. The experiment reported here was

approved by the Ethics Committee of the School of Psychology at the University of Kent and conducted according to BPS guidelines.

Stimuli

Sixty-four face pairs from the Glasgow University Face Database (GUFD) provided stimuli for this study (see Burton et al., 2010). These consisted of 32 match pairs (two different same-day photographs of the same person) and 32 mismatch pairs (two different individuals in each pair). All the faces were presented in greyscale, with a front pose and neutral expression. The faces were cropped to remove extraneous background. The maximum size for a face was 43 x 54 mm, with a maximum gap between faces in a pair of 25 mm.

In the experiment, 40 pairs (20 matches, 20 mismatches) were utilised as the target pairs for observers to match. Each target pair was shown beneath the question "Match or Mismatch?". These pairs were centrally presented and repeated across three blocks. The remaining pairs were divided in to 'low' and 'high' difficulty (based on average observer performance in the previous chapter) to serve as example face pairs. The mean accuracy for the low and high difficulty examples was 96.9% and 85.1% respectively. An independent samples *t*-test confirmed performance was higher for the low-difficulty than the high-difficulty pairs, $t(22) = 7.54$, $p < .001$. These face pairs were then employed as example stimuli in Block 2 and flanked the target pair for the two experimental conditions. Two examples were provided for each target, one matching and one mismatching pair and were clearly labelled as such. The example pairs for each target were selected randomly, but the sex of the example faces was matched to the target face for all trials. For an illustration of stimulus arrays, see Figure 4.1.

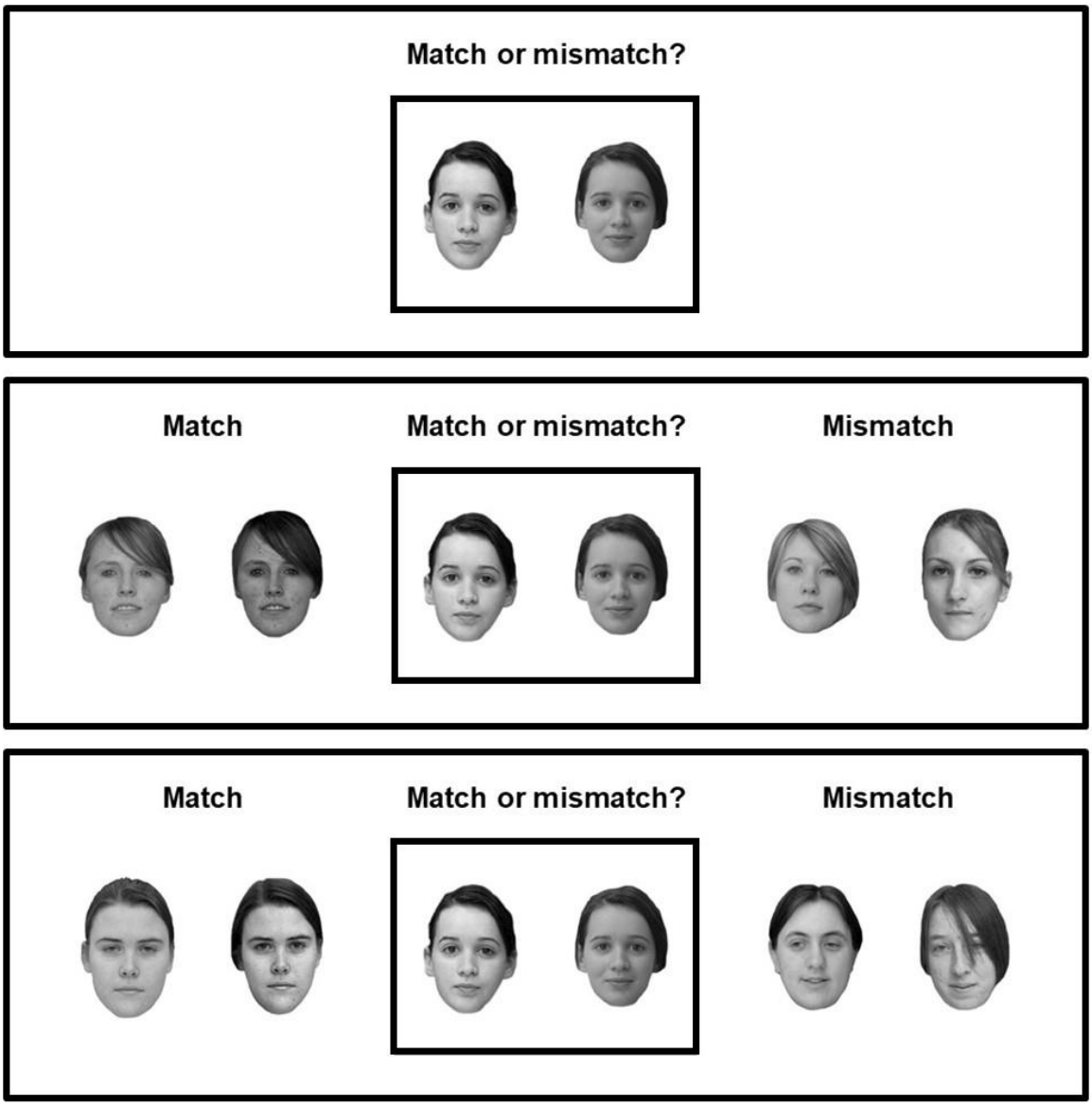

FIGURE 4.1. *Example stimulus arrays for Blocks 1 and 3 (top row) for all conditions and Block 2 for the low-difficulty examples (middle row) and high-difficulty examples (bottom row) for Experiment 7. In the no-examples condition, Block 2 was the same as Blocks 1 and 3 (top row).*

Procedure

SR-Research Experiment Builder software (VERSION 1.1.0) was utilised to display the stimuli for this experiment. The stimuli were presented on a 21-inch colour monitor with a screen resolution of 1024 x 768. An SR-Research Eyelink 1000 was employed to track eye movements, using a sampling rate of 1000 Hz, a spatial resolution < 0.01˚ and gaze position

accuracy < 0.5˚. Observers were positioned 60 cm from the monitor, with a chinrest to maintain the distance and head position. Only the left eye was tracked for each participant and was calibrated using the standard Eyelink nine-point fixation procedure. Hence, observers viewed a set of nine fixation targets, with a second sequence of nine targets to validate the calibration. If this process revealed a measurement error of greater than 1˚ of visual angle, the calibration procedure was repeated. A fixation cross was shown at the beginning of each trial to allow for drift correction, followed by a grey screen for a duration of one second. The stimuli were then displayed until the observer confirmed their matching decision by pressing one of two response keys on a standard computer keyboard. All participants completed a total of three blocks, each containing 40 trials (20 match, 20 mismatch). In the second block of the example conditions, each stimuli screen followed with a screen asking participants to use one of two different response keys to indicate whether they had utilised the examples.

Results

Group-level accuracy

In the first step of analysis, a 2 (trial type: match vs. mismatch) x 3 (example type: no-examples vs. low-difficulty vs. high-difficulty) x 3 (block: Block 1 vs. Block 2 vs. Block 3) mixed-factor ANOVA was conducted to determine whether the nature of the examples presented can impact task performance. The cross-subject means of this data are displayed in Figure 4.2. ANOVA revealed the main effect of trial type was approaching significance, $F(1,87) = 4.06$, $p = .05$, $\eta_p^2 = .05$, due to higher accuracy on match trials. There was also a main effect of block, $F(2,174) = 6.85$, $p < .01$, $\eta_p^2 = .07$, and an interaction between these two factors, $F(2,174) = 11.27$, $p < .001$, $\eta_p^2 = .12$.

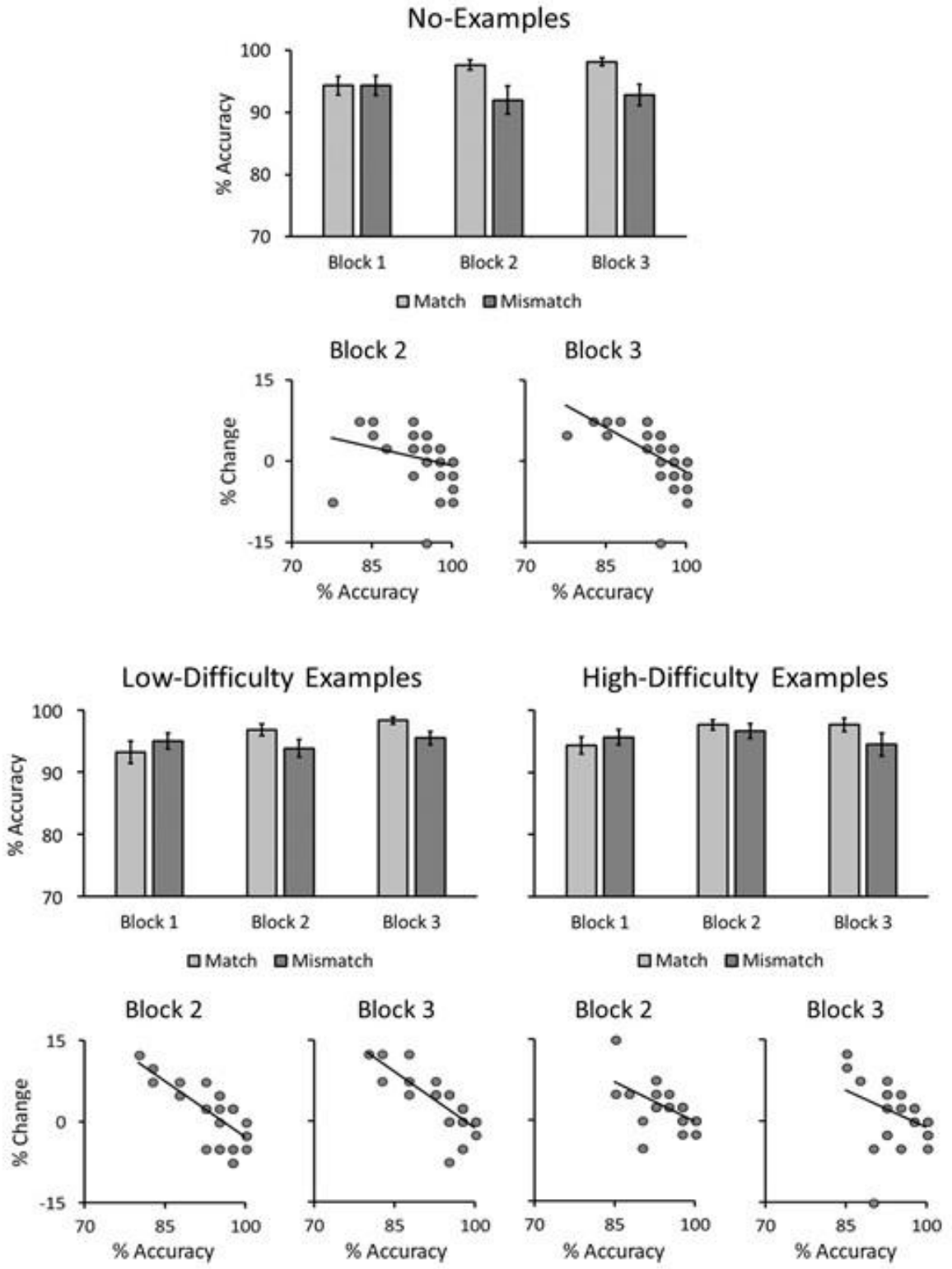

FIGURE 4.2. *Percentage accuracy across blocks by trial type (top row) and percentage change in accuracy between blocks correlated with baseline accuracy (bottom row) for no-examples, low-difficulty examples and high-difficulty examples for Experiment 7.*

Analysis of simple main effects did not reveal a main effect of trial type for Block 1, $F(1,87) = 0.64$, $p = .43$, $\eta_p^2 = .01$. However, there were reliable main effects of trial type for Block 2, $F(1,87) = 7.94$, $p < .01$, $\eta_p^2 = .08$, and Block 3, $F(1,87) = 13.06$, $p < .01$, $\eta_p^2 = .13$, due to higher accuracy on match trials compared to mismatch trials. There was also a main effect of block for match trials, $F(2,86) = 12.43$, $p < .001$, $\eta_p^2 = .22$. A series of paired-sample *t*-tests with alpha corrected to .05/3 = .017 for three comparisons) revealed match accuracy for Block 1 was lower than for Block 2, $t(89) = 4.51$, $p < .001$, and Block 3, $t(89) = 5.04$, $p < .001$. No difference in match accuracy was found between Block 2 and Block 3, $t(89) = 1.59$, $p = .12$. These comparisons indicate that match accuracy increased for all conditions after Block 1 of this experiment. A main effect of block was not found for mismatch trials, $F(2,86) = 0.62$, $p = .54$, $\eta_p^2 = .01$.

ANOVA did not reveal a main effect of example type, $F(2,87) = 0.59$, $p = .55$, $\eta_p^2 = .01$, or an interaction between example type and block, $F(2,87) = 0.75$, $p = .48$, $\eta_p^2 = .02$, or trial type, $F(4,174) = 1.70$, $p = .15$, $\eta_p^2 = .04$. The three-way interaction between these factors was also not significant, $F(4,174) = 0.46$, $p = .76$, $\eta_p^2 = .01$.

Individual differences

To assess individual differences, a measure of change in performance was calculated by subtracting observers' percentage accuracy on Block 1 from Block 2 and Block 3 respectively. These scores were then correlated with Block 1 accuracy to determine if improvements in accuracy were associated with baseline performance[1]. This data is illustrated in Table 4.1. Negative correlations were consistently observed between baseline accuracy and change in performance in the low-difficulty and high-difficulty examples conditions, but also in the no-examples condition. However, there were a few exceptions. No

[1] In the previous chapter, three outliers (i.e., with a baseline accuracy of <70%) were identified. However, for this experiment, the minimum overall accuracy observed at baseline was 77.5%. Therefore, no data points were identified as outliers in this experiment as all individuals obtained high levels of accuracy and exhibited normative performance for the GFMT (see Burton et al., 2010).

correlation was found for overall accuracy or mismatch accuracy in Block 2 for the no-examples condition. In addition, no correlation was found for mismatch accuracy in Block 2 or Block 3 for the high-difficulty examples condition. Thus, in general, observers who perform worse at baseline perform better in Blocks 2 and 3 across all three conditions.

Experiment 7		Block 2	Block 3
No-Examples	Overall	$r(28) = -.25$	$r(28) = -.59$ **
	Matches	$r(28) = -.87$ ***	$r(28) = -.90$ ***
	Mismatches	$r(28) = .02$	$r(28) = -.46$ *
Low-Difficulty Examples	Overall	$r(28) = -.71$ ***	$r(28) = -.78$ ***
	Matches	$r(28) = -.86$ ***	$r(28) = -.96$ ***
	Mismatches	$r(28) = -.42$ *	$r(28) = -.65$ ***
High-Difficulty Examples	Overall	$r(28) = -.59$ **	$r(28) = -.39$ *
	Matches	$r(28) = -.83$ ***	$r(28) = -.71$ ***
	Mismatches	$r(28) = -.35$	$r(28) = -.19$

* = $p < .05$, ** = $p < .01$, *** = $p < .001$

TABLE 4.1. *Summary of correlations for Experiment 7. Baseline (Block 1) accuracy was correlated with change in performance (Block 2 / 3 performance minus Block 1 performance) when observers were or were not given examples (Block 2) and in the subsequent block were examples were removed if provided previously (Block 3).*

Example Usage

This experiment also aimed to assess how observers use examples in an unfamiliar face matching task and thus, participants' eye-movements were analysed. Prior to analysis, eye-movements were filtered to amalgamate fixations of less than 80 ms with the prior or subsequent fixation if it was within half a degree of visual angle. If not, these short fixations were discounted. As processing is unlikely to stop during an eye-blink, when these occurred, their duration was added to the immediately previous fixation.

For this experiment example usage data was collected using two methods; self-report (as in Chapter 3) and observers' fixations on the example stimuli. Participants reported utilising examples on 28.3 % (*SD* = 12.7) and 28.3% (*SD* = 14.5) of trials for low-difficulty and high-difficulty examples respectively. In contrast, participants fixated the low-difficulty example pairs on 73.1% (*SD* = 20.0) of trials and looked at the high-difficulty example pairs on 80.8% (*SD* = 22.9) of trials. A 2 (example difficulty: low-difficulty vs. high-difficulty) x 2 (measure: fixation vs. self-report) revealed a main effect of measure, $F(1,58) = 305.11$, $p < .001$, $\eta_p^2 = .84$, due to more fixation to example pairs than reported use of them (see Figure 4.3). Observers therefore look at examples on far more trials than they report utilising them to make their matching decisions. There was no main effect of condition, $F(1,58) = 1.08$, $p = .30$, $\eta_p^2 = .02$, and no interaction between these variables, $F(1,58) = 1.86$, $p = .18$, $\eta_p^2 = .03$.

Fixation and self-report of example usage was also correlated to determine the agreement between these two measures (see Figure 4.3). For low-difficulty examples, fixation to the example pairs did not correlate with report of example usage, $r(28) = .16$, $p = .39$. However, for high-difficulty examples, fixation and report of example usage were positively correlated, $r(28) = .43$, $p < .05$. This suggests observers felt that they learned more from the high-difficulty examples as the more observers look at the examples, the more they report using them. As increased fixation to low-difficulty examples did not predict observers reported usage, it suggests that observers perceived these to be less useful for improving classification of the target pairs.

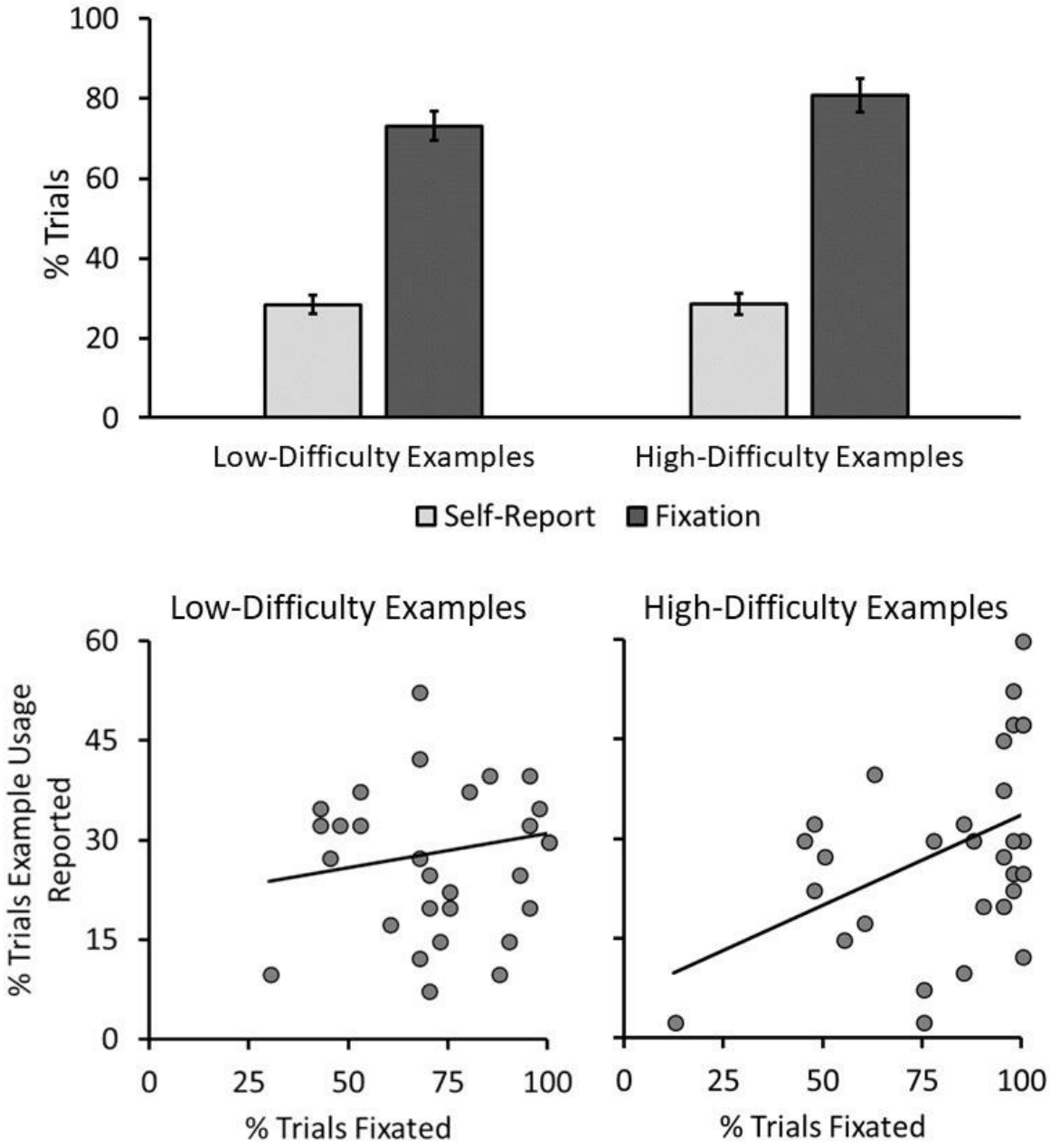

FIGURE 4.3. *Percentage of trials where participants fixated on and reported utilising examples for the low-difficulty and high-difficulty example observers for Experiment 7. These two variables were also correlated for each condition to determine if the number of trials fixated predicts report of example usage.*

Example usage and Improvement

To determine if reported example usage is related to improved task performance, change in performance was correlated with self-reported example usage (see Figure 4.4). For the low-difficulty examples, self-reported example usage did not correlate with change in

performance for Block 2, $r(28) = -.02$, $p = .91$, or Block 3, $r(28) = .03$, $p = .86$. Similarly, for the high-difficulty examples, reported example usage did not correlate with change in performance for Block 2, $r(28) = .18$, $p = .35$, or Block 3, $r(28) = -.07$, $p = .72$. These findings corroborate those of the previous chapter that self-report is not a reliable measure of example usage on this task.

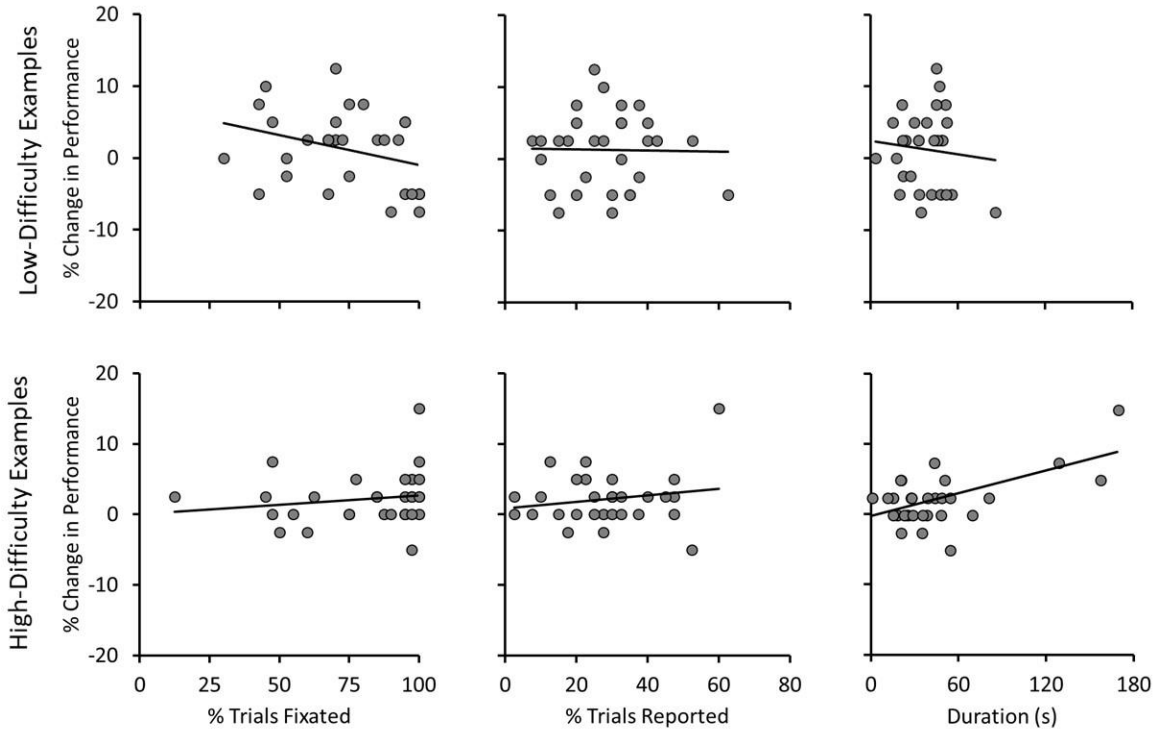

FIGURE 4.4. *Correlation between percentage change in performance from Block 1 to Block 2 with the percentage of trials fixated, the percentage of trials where observers reported viewing examples and the duration of example viewing for the low-difficulty and high-difficulty examples conditions for Experiment 7.*

Change in performance was also correlated with fixation to example pairs to establish whether looking at example pairs on more trials is associated with improved task performance (see Figure 4.4). For the low-difficulty examples, fixation did not correlate with change in performance for Block 2, $r(28) = -.32$, $p = .09$, or Block 3, $r(28) = -.11$, $p = .56$. Likewise, for the high-difficulty examples, fixation did not correlate with change in

performance for either Block 2, $r(28) = .16$, $p = .40$, or Block 3, $r(28) = .09$, $p = .64$. This suggests that observers learning from examples does not relate directly to the number of times these are viewed.

Fixation Duration

In this experiment, fixation only captured whether or not observers looked at the examples and did not discriminate between merely glancing at the examples or carefully studying them. The time spent looking at examples may therefore provide a better measure of example usage than simply assessing whether or not examples were looked at. A paired sample *t*-test indicated that time spent viewing the example pairs was comparable for the low-difficulty and high-difficulty example condition, $t(58) = 0.94$, $p = .35$. However, whereas duration did not correlate with change in performance for low-difficulty examples in Block 2, $r(28) = -.10$, $p = .62$, or Block 3, $r(28) = -.16$, $p = .41$, duration was positively correlated with change in performance for the high-difficulty examples for Block 2, $r(28) = .59$, $p < .01$, but not Block 3, $r(28) = .21$, $p = .26$ (see Figure 4.4). This indicates that the high-difficulty examples procured a benefit for target classification on trials in which these were viewed for longer.

Discussion

In the previous chapter, the provision of example match and mismatch face pairs was found to improve unfamiliar face matching accuracy, especially for individuals who were low-performing at baseline. However, self-report of example usage was low across experiments (approximately 25% of trials). Furthermore, increased example usage did not predict task improvement, which suggests that self-reported example usage is inaccurate. Previous research has demonstrated that observers have limited insight into their own

cognitive processes (Nisbett & Wilson, 1977; Wilson & Dunn, 2004). Thus, the present study examined whether eye-tracking provided better insight into how observers utilise example pairs to improve their task performance. In addition, this study aimed to investigate how the nature of the examples provided relates to task performance, by affording participants no-examples, or by providing low- and high-difficulty examples.

For both the low-difficulty and high-difficulty examples conditions, observers fixated on the examples on substantially more trials than they reported using them. This is in line with previous studies that suggest observers struggle to accurately report their cognitive processes (see, e.g., Clarke et al., 2017; Mahon et al., 2018). Although observers fixated on the examples and reported using the examples on a similar percentage of trials for both example types, these two variables were only correlated for the high-difficulty examples. This suggests that observers feel they are using the high-difficulty examples more for discriminating between match and mismatch trials when they look at them.

However, self-reported example usage did not inform task improvement for either the low- or high-difficulty examples. Previous studies have found that observers have limited insight into their own face processing ability (see, e.g., Bindemann et al., 2014; Bobak, Pampoulov et al., 2016; Palermo et al., 2017, but see Ventura et al., 2018). Thus, it is possible that observers are also unaware of when the examples are needed to make a decision and when the examples have influenced their decision. Participants were asked to report whether or not they had utilised the examples for their decision after each trial, however, their interpretation of the word ‘use’ may be subjective. For instance, some observers may report utilising the example pairs if they simply look at them during the trials, whereas others may only report using the examples if they have used the pair in their decision-making process (e.g., as a contrast to the target pair). Furthermore, the task instructions implied that the examples should be utilised on difficult trials. Therefore, observers may have reported using

the examples if they found a trial challenging, regardless of whether they actually used the example to help reach their decision.

To explore these questions, eye movements to the examples were also recorded. This data indicates that the percentage of trials where examples were fixated also did not relate to improvement. However, the fixation measure only determines whether the example pair was looked at (and does not discriminate between observers merely glancing at the examples or carefully studying them). Therefore, it was also necessary to consider the duration of these fixations. For the high-difficulty examples, more time viewing the examples increased the likelihood of improvement. However, no such association was found for the low-difficulty examples condition. Accurate matching decisions for unfamiliar face pairs can be made within two seconds of viewing (see, e.g., Bindemann, Fysh, Cross, & Watts, 2016; Fysh & Bindemann, 2017b; Özbek & Bindemann, 2011). Thus, it is possible that as the low-difficulty examples consist of pairs that are clear matches and mismatches, looking at them for longer time periods does not increase learning. The high-difficulty pairs are likely to require more processing, as the match pairs have more variation between faces and the mismatch pairs are greater in similarity and therefore, need to be studied for longer for learning to take place. Taken together, these results suggest that the nature of the examples provided impacts the way the examples are processed and used in an unfamiliar matching task.

This experiment also investigated how the nature of the examples provided impacts task accuracy. At a group level, the provision of examples (low-difficulty or high-difficulty) did not improve task performance. However, in the previous chapter the difference in group level accuracy was numerically small and required a larger sample size to detect. Furthermore, observers were found to not benefit equally from the provision of examples. Correlational analyses for Blocks 2 and 3 revealed a similar pattern to the previous chapter, indicating an overall improvement for low-performing individuals for both groups provided

with examples. When broken down into match and mismatch trials, a match improvement was found for both examples groups, but a mismatch improvement for poor-performing individuals was only found for the low-difficulty condition. However, a match improvement was observed not only for the two examples groups but also for the no-examples group. This match effect is in line with the previous chapter and is likely to be a result of an increase in match responses over time that has been observed in numerous studies (see, e.g., Alenezi, Bindemann, Fysh, & Johnston, 2015; Bindemann et al., 2016; Fysh & Bindemann, 2017b; Papesh, Heisick, & Warner, 2018). However, in contrast to the previous chapter, change in performance was also associated with overall baseline accuracy and mismatch accuracy in Block 3 for the no-examples group. This implies that in the final block, low-performing individuals were also able to improve in the *absence* of examples. Therefore, this experiment did not provide clear evidence that examples improve performance at a group or individual level.

So why might some observers have demonstrated task improvement without the help of examples? One potential explanation is that this experiment differed in fundamental ways to those reported in the previous chapter. For example, eye-tracking requires the experimenter to initiate every trial due to the drift correction, which slowed observers down completing the task so may have increased their accuracy. Furthermore, the eye-tracking set-up requires the experimenter to be present in the room and sitting next to the participant as they complete the task. Previous studies have indicated that food and money can act as motivational incentives that improve unfamiliar face matching accuracy (e.g., Bobak, Dowsett, & Bate, 2016; Moore & Johnston, 2013). Thus, the presence of the experimenter may have acted as a similar motivation to perform better at the task. Moreover, the proximity of the experimenter as well as the eye-tracking nature of the task may have led participants to feel they were being closely monitored as they completed the experiment. The perception of being watched has

been shown to impact behaviour in a number of different scenarios (see, e.g., Bateson, Callow, Holmes, Redmond Roche, & Nettle, 2013; Bateson, Nettle, & Roberts, 2006; Fathi, Bateson, & Nettle, 2014; Nettle, Nott, & Bateson, 2012). For example, telling observers they are being monitored has been shown to reduce processing speed and ultimately increase accuracy on visual search tasks (Miyazaki, 2013). Hence, the perception of being monitored by the experimenter during the task may have increased task performance.

These explanations are in line with higher overall accuracy and a reduced range of scores at baseline for the present study (M = 94.47, SD = 5.20, range: 77.5-100) compared to in the previous chapter (M = 92.10, SD = 7.22, range: 50-100; $t(268) = 2.78$, $p < .01$). The GFMT is a relatively easy face matching task with a high normative performance of around 80-90% (Burton et al., 2010). While this was not a problem in the studies reported in the previous chapter, this in combination with the increases in accuracy that likely resulted from fundamental task differences (such as the proximity of the experimenter), meant there was very little room for improvement with the provision of examples as task performance was near ceiling. Consequently, performance was far more likely to decline over the course of the task than improve (see, e.g., Alenezi & Bindemann, 2013; Fysh & Bindemann, 2017b; Papesh et al., 2018). This decline is evident in the correlations between baseline accuracy and change in accuracy across the three conditions, which appear to be driven by higher performing individuals getting worse over time as opposed to lower performing individuals improving (see Figure 4.2).

For future research, it would be interesting to manipulate the eye-tracking set up so that the experimenter does not need to be in the room with the participants while they complete the task. This may help to address the near ceiling level performance found at baseline which likely resulted from the perception of being watched or being more motivated to be accurate due to the presence of the experimenter. Alternatively, the use of a more

challenging stimuli set such as the Kent Face Matching Test (KFMT, Fysh & Bindemann, 2018a) or the Models Face Matching Test (MFMT, Dowsett & Burton, 2015), would provide more room for improvement and thus, would allow a better insight into if and how the provision of examples improves accuracy for unfamiliar face matching.

In summary, this experiment provides insight into how examples are utilised to make decisions in an unfamiliar face matching task. In line with previous research, this study suggests that participants struggle to accurately report cognitive processing, as participants fixate on examples on substantially more trials than they report using them. Observers appear to feel that high-difficulty examples are more useful. However, these examples require more processing (i.e., a longer fixation duration) to facilitate learning and procure a benefit. However, in contrast to the experiments reported in the previous chapter, the present study did not provide clear evidence that the provision of examples can improve performance at a group or individual level. This was likely due to a combination of using relatively easy GFMT stimuli and the nature of the eye-tracking setup (i.e., the close proximity of the experimenter to the participant) motivating observers to be more accurate on the task across all three conditions. Consequently, more challenging stimuli such as the KFMT (Fysh & Bindemann, 2018a) or the MFMT (Dowsett & Burton, 2015) or an alternative eye-tracking set up (i.e., where the experimenter is not in the room with the participant) may be required to clarify the benefit of the provision of examples for unfamiliar face matching.

Chapter 5

Summary, conclusions and future research

This thesis investigated how matching decisions are reached for pairs of unfamiliar faces and assessed a novel training method for improving accuracy in this task. A large body of psychological research has demonstrated that unfamiliar-face matching is a surprisingly error-prone task (for reviews, see, e.g., Fysh & Bindemann, 2017a; Robertson, Middleton, & Burton, 2015). However, while the difficulty of this task is well established, much less is known about the decision-making process that underlies it. Improved understanding of how observers perform this task may help to inform training, which could in turn increase matching performance. To this end, a small number of recent studies have attempted to establish whether there are 'critical' features that drive task accuracy (see, e.g., Abudarham & Yovel, 2016; Megreya & Bindemann, 2018; Towler, White, & Kemp, 2017). These studies converge to suggest that specific individual facial features can influence an overall matching decision for a pair of faces. However, these studies also demonstrate disagreement in terms of which feature is most important, variously emphasising lip thickness, hair and eye colour (Abudarham & Yovel, 2016), the eyebrows (Megreya & Bindemann, 2018) or ears (Towler et al., 2017). It is therefore possible that the most important features for reaching face-matching decisions vary across different face images and individuals. This reasoning is supported by the notion that face photographs of a single individual can display substantial variation (see, e.g., Jenkins, White, Van Montfort, & Burton, 2011).

If there is no universal facial feature that drives accuracy, then it is possible that matching requires a combination of judgements for different features to reach a decision for the whole face. In other words, while an overall face-matching decision must be based on the assessment of the entire stimulus, it is possible that a number of smaller evaluations are first made for individual facial features. Chapter 2 investigated this possibility with a series of three experiments. For this purpose, observers were required to make match or mismatch decisions to pairs of whole faces as well as isolated feature regions (Experiment 1) or could

also respond ‘don’t know’ if unsure of the correct response (Experiment 2). The feature regions comprised of the hair / forehead, eyes, nose and mouth. By comparing whole-face accuracy for identity pairs for which different numbers of individual features were classified correctly, Chapter 2 sought to determine whether overall matching decisions reflect judgements to individual facial features, and if so, how these smaller feature assessments are utilised or combined to reach the final decision for the whole stimulus.

These initial experiments demonstrated a graded response pattern such that accuracy for the whole-face pairs increased in line with the number of correct feature decisions. Performance was best for the whole face when three or four of the corresponding feature regions had been classified correctly. This suggests that judgements made to individual features might be *summed* to reach a decision for the whole face. However, accuracy for the whole-face pairs was also high, at nearly 90%, when only two of the four features under investigation here were classified correctly. If half of the feature regions were classified as a match and the other half as a mismatch, one might expect this to cause sufficient conflict such that observers should be equally likely to classify the whole face as a match or mismatch. Thus, one would expect accuracy to be closer to chance. Moreover, whole-face accuracy also remained relatively high, at around 70%, when only one feature was classified correctly. Thus, despite three of four features being classified incorrectly, observers still reached the correct decision for the whole face on the majority of these occasions.

These findings are in line with previous research that demonstrates that individual facial features can strongly influence the overall matching decision (see, e.g., Abudarham & Yovel, 2016; Megreya & Bindemann, 2018; Towler et al., 2017). These studies have attempted to identify a universal critical feature that underlies face matching accuracy. However, Experiments 1 and 2 did not reveal any single facial feature that was matched consistently more accurately than the others. This finding may suggest that there is not a

universal feature which underpins matching accuracy, but instead that the 'critical' feature varies from face to face.

So how can high levels of accuracy be maintained for the whole face even when most of the facial features are misleading and point to the incorrect matching decision? One possibility is that for some faces, feature decisions may need to be *weighted* rather than summed. In these cases, some facial features may dominate the decision process if they provide particularly compelling information. For instance, it is conceivable that for a pair of faces, the hair, nose and mouth look similar, but the eyes appear distinctly different, leading an observer to conclude that the pair is a mismatch. This reasoning receives some support from the Matching Familiar Figures Test (MFFT), where observers must rely on differences in a single aspect of a line drawing (e.g., the direction of a flag on a ferry boat) to determine which item in a line-up is an exact match to a target object (Megreya & Burton, 2006b). MFFT performance correlates moderately with unfamiliar face matching accuracy, suggesting that a similar reliance on a single feature may sometimes inform the correct decision for the whole face.

In Experiments 1 and 2, accuracy for the whole face pairs surpassed that of any of the isolated feature regions. This may have been a result of the additional matching information (facial features) available in the whole face pairs, and so may provide an alternative explanation of why whole face accuracy remained high when only one or two features were classified correctly. However, even when observers classified none of the feature regions correctly in Experiment 2, they were still able to reach the correct overall decision on nearly half of these occasions. This suggests there is an advantage, over and beyond observing all features of a face in isolation, of viewing faces as an integrated whole. Therefore, the extra featural information available in the whole face pairs may not adequately explain the increased accuracy for this stimulus type.

To investigate this further, observers were required to match pairs consisting of whole faces, misaligned whole faces (with all features horizontally offset to disrupt holistic processing) and misaligned part faces (which either displayed the hair and nose regions or the eye and mouth regions) in Experiment 3. For both match and mismatch pairs, performance for misaligned whole face pairs was comparable to misaligned part faces. Therefore, the additional featural information available is unlikely to account for the increased accuracy for whole faces observed in Experiments 1 and 2. For match trials, performance was also better for whole face pairs than misaligned whole or part face pairs. This suggests that viewing all facial features at the same time is not sufficient to increase matching accuracy, but that these features need to be integrated into a whole face to maximise performance. This effect may be similar to the holistic processing advantage described in the face *recognition* literature (see, e.g., Goffaux & Rossion, 2006; McKone, 2004; Tanaka & Sengco, 1997). Thus, ultimately whether face processing is required for matching or recognition, the context in which facial features are viewed is important. However, there was no such advantage of viewing the whole face for mismatch trials. Previous research has shown that performance for match and mismatch stimuli is dissociable (see, e.g., Kokje, Bindemann, & Megreya, 2018; Megreya & Burton, 2006b, 2007). Hence, it is possible these trial types rely on different mechanisms, whereby holistic processing is more important for processing face pairs consisting of the same person.

The experiments reported in Chapter 2 suggest that observers can adapt the strategy they use for face matching based on the specific face pair at hand. For example, when the majority of feature judgements point to the same decision, observers appear to sum these judgements to reach an overall decision. However, when feature judgements are conflicting, observers can utilise information from a compelling feature to reach the correct overall decision. Judgements for particularly compelling facial features may be weighted such that

they can dominate a matching decision, even if all other facial features point to the incorrect decision (Experiments 1 and 2). As there was no single feature that determined task accuracy in these experiments, it is likely that the most compelling or most useful feature varied from face to face. This is supported by the notion that faces vary in an idiosyncratic manner, so images of the same individual can appear very different from each other (Jenkins et al., 2011). Despite this, Experiment 3 demonstrated that the integration of facial features, in a whole face, is required to maximise performance. Therefore, a combination of featural and holistic processing is likely required to reach the correct matching decision. However, as the whole face advantage was only found for match trials, it is possible that holistic processing is more important for "telling people together" than for "telling them apart".

These findings raise some interesting questions for future research. The experiments in Chapter 2 suggest that observers may employ different matching strategies for different face pairs. This finding could be explored further by using an eye-tracking paradigm to investigate how different facial features are viewed during this task. Analysing eye-movements for different items may give further insight into whether people employ different matching strategies for different faces. For example, if observers sum feature judgements to reach a decision for a particular pair, it is possible that all features are viewed initially and then re-scanned just before an overall decision is made. If, on the other hand, a matching decision is dominated by one compelling feature, then observers may view this feature disproportionately, and last, just before a decision is made.

Individual performance for face-matching tasks is highly variable (for a review see, Lander, Bruce, & Bindemann, 2018). These differences are such that some individuals excel at face matching tasks (see, e.g., Bobak, Dowsett, & Bate, 2016; Norell et al., 2015; White, Phillips, Hahn, Hill, & O'Toole, 2015), whereas, others perform close to chance (see e.g., Burton, White, & McNeill, 2010; Fysh & Bindemann, 2018a). Thus, it would be interesting

to also explore the decision-making processes and viewing strategies of both low- and high-performing individuals with eye-tracking. Using this method, it may be possible to identify strategies that are most effective for face matching in order to improve accuracy for poor-performing individuals. For example, it may be possible to examine whether the high-performing individuals notice details that other observers miss, which allow them to be more accurate (see Figure 5.1). As images of the same individual can be highly variable (Jenkins et al., 2011), it is possible that different matching strategies are required to reach the correct decision for different pairs of faces. Furthermore, the most important identity-related information may need to be derived from different features according to the specific face pair at hand. Thus, analysing eye-movements for individual items as well as individual participants may give greater insight into how observers approach unfamiliar face matching tasks and why performance across participants is highly variable.

FIGURE 5.1. *These two example faces taken from the KFMT (Fysh & Bindemann, 2018a) appear very different but are in fact the same individual. However, a characteristic pattern of moles is visible in both images which helps to identify them as the same person. It is possible that higher-performing individuals are more likely to notice such details than other observers, helping them to be more accurate.*

Experiment 3 indicates that holistic processing appears to be more important for pairs that depict the same individual than pairs consisting of two different people. It is possible that holistic processing is more important for match trials, because for images of the same person, the features should all point to the same decision (i.e., to the faces depicted being the same identity). By contrast, while mismatch trials are constructed to incorporate two individuals who look highly similar, these ultimately do not depict the same person, so concurrent dissimilarities between faces must also be present. Thus, featural processing may be more important for mismatch pairs, as comparing individual features is more likely to allow the detection of differences. Again, it would be interesting to investigate these matching strategies further with eye-tracking. Comparing eye-movements for whole faces and misaligned whole faces for these two trial types may give further insight into how observers approach this task.

Observers may use different strategies to process the whole face and misaligned whole face pairs. For example, it is possible that when presented with whole face pairs, observers view one face in its entirety and then view the other face, before making feature comparisons. However, for misaligned faces where holistic processing of the entire face is more difficult, they may only compare specific features across the faces. Thus, comparing viewing strategies for these two pair types may demonstrate whether performance was lower for the misaligned faces because observers are only able to use one strategy. Furthermore, when eye-movements for identity-match and identity-mismatch pairs are considered separately, this may reveal if observers spend longer comparing features for mismatch trials, which could indicate that featural processing is more important for these stimuli.

A potential limitation of Experiment 3 was that the stimuli remained on screen until the observers made a matching decision. Consequently, observers had time to use featural processing to carefully compare the face sections for the misaligned whole and part face

stimuli. This may explain why accuracy for the misaligned conditions was comparable to whole face accuracy for mismatch trials. Hole (1994) found that subjects were more successful on a composite task which required the matching of the top half of upright or upside-down chimeric faces, when the faces were presented for 2 s compared to 80 ms. Viewing the faces for 2 s allowed observers to engage in feature comparison across the two faces thus, reducing errors. However, the 80 ms condition forced observers to process the composites holistically and therefore reduced accuracy. It would therefore be interesting to replicate Experiment 3 with the faces only displayed for a short duration, to force holistic processing of all stimuli types. It is likely that this manipulation would significantly reduce accuracy for the misaligned stimuli and thus may reveal a whole face advantage for both match and mismatch trials.

Whilst Chapter 2 provides insight into the decision-making process for unfamiliar face matching, another topic that is currently of growing interest to researchers is whether it is possible to increase performance on face-matching tasks. Experiments 1-3 suggest that training individuals to improve matching by relying on a specific facial feature or strategy may be difficult. Accordingly, training using facial features has demonstrated mixed success. For example, training observers to use face shape does not work (Towler, White, & Kemp, 2014). Other featural training methods have been more successful, such as training observers to compare different facial features before making an overall decision (Towler et al., 2017) and focusing on a specific feature such as the eyebrows (Megreya & Bindemann, 2018). However, while both of these approaches have been shown to improve accuracy, generalisation to new faces has been more limited.

Therefore, this thesis also investigated a novel means of improving unfamiliar face-matching accuracy. Previous methods for increasing matching performance, can be divided into stimulus-based and observer-based approaches. Stimulus-based approaches focus on

providing observers with improved face representations for matching, such as averaged faces or caricatures (see, e.g., Burton, Jenkins, Hancock, & White, 2005; McIntyre, Hancock, Kittler, & Langton, 2013; Robertson, Kramer, & Burton, 2015). However, such stimulus-based approaches can be time-consuming or impractical to apply to photographic ID. Construction of average faces, for example, requires multiple images of an individual, and this process becomes more effective as more photographs are incorporated (Burton et al., 2005).

Observer-based approaches, on the other hand, seek to improve the accuracy of individuals conducting face matching, using methods such as combining the performance of multiple observers or by providing training (see, e.g., Balsdon, Summersby, Kemp, & White, 2018; Megreya & Bindemann, 2018; Towler et al., 2017; White, Burton, Kemp, & Jenkins, 2013). One such method of training is the provision of feedback for a person's face-matching accuracy. In real-world matching scenarios, observers typically do not have the opportunity to learn from their errors (Jenkins & Burton, 2011). Training strategies that allow observers to develop their matching criteria (see, e.g., Lander et al., 2018), such as training with feedback, can improve task performance. For example, matching accuracy increases if feedback is provided while a just-classified stimulus is still on view (White, Kemp, Jenkins, & Burton, 2014). Similarly, providing feedback after a trial can help to maintain mismatch accuracy (Alenezi & Bindemann, 2013; Papesh, Heisick, & Warner, 2018) which typically declines over the course of a matching task (see, e.g., Alenezi, Bindemann, Fysh, & Johnston, 2015). The improvement that feedback produces on task performance suggests that individuals may have limited criteria for discerning between a match and mismatch (see, e.g., Lander et al., 2018). Therefore, feedback may work by providing observers with a platform where they can deduce matching criteria they need to successfully complete the task. However, providing feedback in this manner requires knowledge of the nature of the face

pairs (i.e., whether these are matches or mismatches) at the point of identification, which is not possible in real-world scenarios.

To address this shortcoming, Chapter 3 investigated an alternative form of feedback which does not require prior knowledge of the correct matching decision for a face pair. This method utilised match and mismatch examples as a method of improving match accuracy. Across three experiments, participants first completed a block of target face pairs to establish baseline accuracy, half then completed a second block where the target faces were flanked by example pairs in the examples condition, while the remainder saw the target faces only in the no-examples condition. The examples displayed either side of the target face pairs constituted clearly-labelled match and mismatch pairs, which observers could use to help inform their decision for the central pair (Experiment 4). By comparing the accuracy of the examples and no-examples groups, this chapter investigated the general effect of examples on face matching. However, as there are also substantial individual differences in matching accuracy (see, e.g., Bindemann, Avetisyan, & Rakow, 2012; Burton et al., 2010; Estudillo & Bindemann, 2014; Fysh & Bindemann, 2018a), the impact of examples was primarily assessed at an individual level. In addition, the experiments in this chapter also assessed whether any improvement with examples is maintained after these are removed again, and whether improvements with examples generalised to new stimuli. For this purpose, observers also completed a third block of stimuli consisting of target pairs only (Experiments 5 and 6). A further block of stimuli that comprised either previously unseen stimuli from the same set (Experiment 5) or new stimuli with different characteristics (Experiment 6) was also included to determine whether the observers had learnt something about the face-matching task which could then be applied to different stimuli.

When data was pooled across experiments, examples were found to increase matching performance at a group level. However, this effect was numerically small (less than 3%) and

inconsistent across experiments. In contrast, analysis of individual differences revealed more clearly that examples improved performance and showed that this was the case particularly for individuals whose accuracy was comparatively low at baseline. This advantage was found consistently across all three experiments, and for match and mismatch performance as well as for overall accuracy. A similar association was found for individuals in the no-examples group, but this was limited to match trials only. In the context of the block by trial type interaction that was observed in the cross-experiment data at a group level, and which was driven by an increase in accuracy during the experiments on match trials, the match trial correlations in the no-examples condition also appear to reflect a tendency to make increasingly more match responses over the course of the experiment. This pattern has now been reported in a number of studies (see, e.g., Alenezi & Bindemann, 2013; Alenezi et al., 2015; Fysh & Bindemann, 2017b).

So how might examples improve task performance? Examples may work by helping to refine observers' criteria for dissociating between match and mismatch trials. Face matching is highly error-prone, which suggests that some individuals may not have sufficient criteria to easily discriminate between match and mismatch trials (see, e.g., Lander et al., 2018). If observers had concrete criteria for what constitutes a match or mismatch face pair, then it is possible that matching errors might be reduced. Further evidence that might support some observers having limited criteria for effectively completing this task is that some individuals can excel at this task (see, e.g., Bobak, Hancock, & Bate, 2016; Norell et al., 2015; Phillips et al., 2018; White, Phillips et al., 2015). Thus, it is possible poor performance is a result of a resource-limit problem, where face pairs encompass all the visual information needed to make the correct identification, but some observers are unable to apply this information effectively.

In addition, individuals seem to have limited awareness of their own unfamiliar face processing ability (see, e.g., Bindemann, Attard, & Johnston, 2014; Bobak, Mileva, & Hancock, 2018; Palermo et al., 2017), which suggests they may not have stable criteria for completing the task. Moreover, the phenomenon of choice blindness has shown that individuals will justify a matching decision for a pair of faces even if this contradicts their original judgement (Sauerland et al., 2016). If observers had clear criteria for dissociating identity-matches and mismatches, choice blindness should not occur. In turn, providing observers with some sort of criteria, such as trial-by-trial feedback can improve accuracy (White, Kemp, Jenkins, & Burton, 2014) or reduce the decline in mismatch accuracy over time (see, e.g., Alenezi & Bindemann, 2013; Papesh et al., 2018). Thus, this feedback may work by helping to refine observers' matching criteria, which may in turn increase accuracy on this task.

Examples may provide observers with a platform by which they can deduce their own matching criteria, thus allowing them to improve their matching accuracy. Further evidence for this notion comes from a recent study that demonstrated feedback can improve accuracy if it is provided while a just classified stimulus is still in view (White, Kemp, Jenkins, & Burton, 2014). Allowing observers to view a target pair in conjunction with feedback is likely to develop their matching criteria and thus, may account for their increase in accuracy. Furthermore, pairs of observers can also outperform individuals in face-matching tasks (Dowsett & Burton, 2015). Discussing the reasoning for a matching decision with another observer may strengthen the criteria of both individuals in a similar manner to the provision of feedback. Therefore, in contrast to other training approaches such as feature comparison (Towler et al., 2017) and focusing on specific features (Megreya & Bindemann, 2018), the examples method allows observers to develop their own matching strategies and adapt their strategy based on the face pair at hand.

Chapter 3 also found that the examples-advantage was maintained after examples were removed and generalised to previously unseen stimuli from the same GFMT set (Experiment 5). As the examples still procure a benefit when they are no longer in view, it suggests that observers learn from the examples and may internalise the criteria so that their performance can continue to improve without the examples. Despite this, the generalisation to pairs from the new KFMT stimulus set, which has different characteristics, was more limited (Experiment 6). This finding may reduce the applied value of examples for security settings, such as passport control, where operators are likely to encounter a diverse range of faces. However, other training strategies such as feature comparison (Towler et al., 2017) and feature instructions (Megreya & Bindemann, 2018) have also demonstrated limited generalisation to new stimuli sets. It is therefore possible that different face stimuli require different criteria to be applied to successfully discriminate between identity-match and mismatch pairs.

The more limited generalisation to the KFMT stimuli suggests that the benefits of the examples may be *stimuli specific* (i.e., exposure to GFMT examples specifically improves GFMT performance). It is possible that examples improve performance by helping to shape the criteria observers use to complete the matching task. Viewing examples may provide a platform that allows observers to deduce what level of variation is likely to occur between images of the same person and the degree of similarity possible between images of two different people. The GFMT stimuli were produced by taking images of individuals using two different cameras a few minutes apart for match pairs and selecting similar looking individuals from within the set for mismatch pairs (see Burton, White & McNeill, 2010). Exposure to example pairs taken from the GFMT may therefore lead observers to expect matching pairs to show very little variation between images. This observation may change the criteria they use to discriminate match and mismatch pairs. However, match pairs in the

KFMT are compiled differently, as images of the same individual are taken at least three months apart. Hence, there is more within-person variability in images of the same person for the KFMT.

If observers' matching criteria were influenced by the GFMT examples that suggested matches should exhibit very little variation, their criteria were likely to be less effective for the KFMT which may explain the more limited generalisation seen in Experiment 6. It is therefore possible that in order to obtain an improvement on the KFMT, observers would need to view KFMT examples that provide an indication of the variability within that particular stimuli set. Therefore, replicating these experiments using only KFMT stimuli is likely to further clarify whether the examples advantage is stimuli specific. Furthermore, it would be interesting to examine whether training using KFMT examples would demonstrate similarly limited generalisation to GFMT target face pairs. However, it is also important to note that within-person variation can be larger than between-person variation (Jenkins et al., 2011). The KFMT is specifically designed to incorporate within-person variation into match pairs. Thus, it is possible that variation may be greater between match pairs for the KFMT than mismatch pairs, which is likely to make it more difficult for observers to develop effective criteria for differentiating match and mismatch trials for the this task.

The notion that the examples may produce stimuli-specific benefits, has important applications for training individuals who are required to perform identity comparisons routinely in applied settings. Experiments 5 and 6 suggest that observers need to be trained on examples specific to those they are likely to encounter. For example, passport officers may need to see examples of different race faces and images that have been taken months or years apart which they are likely to encounter on a daily basis. Both of these factors can make face matching more difficult (see, e.g., Fysh & Bindemann, 2018a; Hills, Cooper & Pake, 2013; Hills & Pake, 2013; Megreya, Sandford, & Burton, 2013). Training in this way should allow

observers to determine the level of variation they are likely to encounter for match pairs and the degree of similarity possible for mismatch pairs and may ultimately improve their accuracy for unfamiliar face matching.

Although the provision of examples was found to improve unfamiliar face matching for low-performing individuals, reported example usage was low (approximately 25% of trials) across Experiments 4, 5 and 6. Furthermore, self-reported example usage was not associated with task improvement. Therefore, the final experimental chapter sought to investigate the discrepancy between self-reported example usage and task improvement that was observed in Chapter 3. Individuals demonstrate limited awareness into their own cognitive processes (Nisbett & Wilson, 1977; Wilson & Dunn, 2004). Hence, it is likely that observers' example usage was not accurately captured by the self-report measure utilised in Chapter 3. Eye-tracking has been utilised in a wide range of studies to give greater insight into face processing (see, e.g., Fletcher, Butavicius, & Lee, 2008; Heisz & Shore, 2008; Smilek, Birmingham, Cameron, Bischof, & Kingstone, 2006; Walker-Smith, Gale, & Findlay, 2013). Thus, Chapter 4 employed eye-tracking to provide a more sensitive measure of how observers view and utilise examples for unfamiliar-face matching (Experiment 7).

A further question arising from Chapter 3 concerned whether the nature of examples provided impacted task improvement. Examples may help to improve the matching criteria that observers apply to conduct unfamiliar face matching tasks. However, faces demonstrate considerable within-person variation (Jenkins et al., 2011), so it is likely some example pairs are more beneficial than others. Affording observers with more variable images of individuals can improve face learning (e.g., Burton, 2013; Burton, Kramer, Ritchie, & Jenkins, 2016; Ritchie & Burton, 2017). Thus, example pairs that incorporate more variation between images of the same individual may be more useful for face matching. On the other hand, mismatch pairs for unfamiliar face-matching tasks are usually selected by choosing the

individuals who look the most alike from the stimuli set (see, e.g., Burton et al., 2010; Fysh & Bindemann, 2018a). This process is representative of imposters attempting to make themselves appear as similar as possible to their falsely-obtained photo-IDs in real-life matching scenarios. Thus, example mismatch pairs that demonstrate more subtle differences between persons, by depicting two individuals who are very similar in appearance, may also be more useful for improving face matching.

In Experiment 7, observers completed an initial block of stimuli to establish an accuracy baseline. Observers were then divided into three groups, comprising low-difficulty examples, high-difficulty examples and no-examples conditions. For participants in the two example conditions, the target pairs in the second block were flanked by clearly-labelled match and mismatch pairs. The low-difficulty examples consisted of highly similar match pairs as well as mismatch pairs that depicted more different-looking individuals. By contrast, the high-difficulty examples comprised match pairs which incorporated more variability and mismatch pairs with more subtle differences. After each trial, the observers were then asked to indicate whether or not they had used the examples to reach a matching decision. Individuals in the no-examples condition saw the target pairs only. All participants then completed a final block of stimuli without examples to determine whether any benefits of viewing the example pairs could be retained when these were no longer in view.

For both example conditions, eye movements revealed that observers fixated the examples on considerably more trials than they had overtly reported using them. This finding is supported by previous research, which demonstrates that individuals find it difficult to accurately report their cognitive processes (see, e.g., Kok, Aizenman, Võ, & Wolfe, 2017; Nisbett & Wilson, 1977; Võ, Aizenman, & Wolfe, 2016). Furthermore, task improvement was not associated with self-reported example usage for the low- or high-difficulty examples. As observers appear to have limited insight into their own face processing ability (see, e.g.,

Bindemann et al., 2014; Bobak, Pampoulov, & Bate, 2016; Palermo et al., 2017, but see Ventura, Livingston, & Shah, 2018), it is possible that individuals have a similar lack of awareness for when examples are needed to help them reach a correct matching decision. If this is the case, then it may explain why self-report of example usage did not relate to task improvement.

However, fixation of the examples, as measured via eye movements, was also found *not* to be related to reported example usage. The fixation measure only assesses whether or not the examples were looked at during the task, but not the length of these fixations. Therefore, the duration of these fixations was also considered. A relationship between task improvement and fixation duration was found for high-difficulty examples, such that the longer these pairs were studied, the more likely an improvement was to occur. However, a similar association did not occur for the low-difficulty examples. It is possible for individuals to make accurate matching decisions within two seconds of viewing a face pair (see, e.g., Özbek & Bindemann, 2011; Bindemann, Fysh, Cross, & Watts, 2016; Fysh & Bindemann, 2017b). Thus, viewing the low-difficulty examples, which comprised of clear match and mismatch pairs, may not encourage sufficient learning from this example type. In contrast, the high-difficulty pairs likely needed to be processed for longer for learning to occur, as the match pairs were more dissimilar and the mismatch pairs more similar. Hence, the more time observers spent viewing the high-difficulty examples, the more likely they were to improve overall. These findings imply that the nature of the examples that are provided may influence the way the examples are processed and utilised to aid face matching.

In addition, Experiment 7 also examined whether the nature of the provided examples affected accuracy at a group-level, by comparing those given no-examples to observers given low-difficulty and high-difficulty examples. There was no benefit of the provision of low- or high- difficultly examples at a group-level. However, as in the previous chapter, this

experiment was primarily interested in whether examples produce an improvement at an individual level. Similarly, to Chapter 3, the provision of both example types improved overall accuracy for low-performing individuals both when the examples were displayed (Block 2) and after the examples were removed (Block 3). When accuracy was broken down by trial type, this also revealed a match improvement for the low- and high-difficulty examples conditions, but a mismatch improvement was limited to the low-difficulty examples only. However, a match improvement was also found for the no-examples group. A match improvement was found for this group in the previous chapter but is likely due to an increased propensity to make a match response over the course of a trial, which has been found in a number of recent studies (see, e.g., Alenezi & Bindemann, 2013; Bindemann et al., 2016; Fysh & Bindemann, 2017b; Papesh et al., 2018). However, contrary to the previous chapter, there was also a relationship between overall baseline accuracy and change in performance for Block 3 in the no-examples condition. This finding suggests that it is possible for poor-performing observers to increase their accuracy *without* the help of examples. Thus, Experiment 7 did not provide clear evidence of an examples-advantage at a group or an individual level. Further research is therefore needed to explore the examples method.

While Chapter 3 indicates that the provision of example pairs can improve matching accuracy, especially in low-performing individuals, this advantage was not clear in Chapter 4. The findings of Chapter 4 raise the question of how observers were able to improve face matching performance in the *absence* of the examples. At present it is unresolved why this is the case. One possible explanation is that there are fundamental task differences between the eye-tracking task reported in Experiment 7 and the tasks completed in Experiments 4, 5 and 6. For instance, the drift correction in the eye-tracking task requires the researcher to initiate every trial manually for the participant. This is likely to have slowed observers down during

the task, which may have increased their accuracy. Requiring observers to perform face-matching tasks under time-pressure reduces task accuracy (see, e.g., Bindemann et al., 2016; Fysh & Bindemann, 2017b). In turn, slowing observers down during the task may have had the opposite effect and improved performance.

Moreover, the experimenter had to be present in the room, sitting next to the participant in order to conduct the eye-tracking. Motivational incentives have been shown to increase accuracy on unfamiliar face matching tasks (see, e.g., Bobak, Dowsett, et al., 2016; Moore & Johnston, 2013). The presence of the experimenter may have acted as similar motivation and thus, may have increased task performance. Furthermore, as the experimenter was in the room with the participant and they had their eye-movements recorded throughout the task, participants may have felt that they were being closely monitored during the task. Awareness of being watched can change individuals' behaviour during a number of tasks (see, e.g., Bateson, Callow, Holmes, Redmond Roche, & Nettle, 2013; Bateson, Nettle, & Roberts, 2006; Fathi, Bateson, & Nettle, 2014). Furthermore, making observers aware that they are being monitored during a visual search task can reduce processing speed and improve performance (Miyazaki, 2013). Therefore, it is possible that the perception of being watched during the experiment could have improved task accuracy. It would be interesting to change the eye-tracking set-up so that close proximity of the experimenter is not required to determine if this could explain the different pattern of results obtained.

The results in both Chapter 3 and Chapter 4 were limited by ceiling effects as a result of using the relatively easy GFMT stimuli in Experiments 4-7. Despite high normative performance for the GFMT of around 80-90% (Burton et al., 2010), there was a small group level improvement found in Chapter 3 when data was collapsed across Experiments 4-6. However, while the use of these stimuli was less of an issue for the studies reported in Chapter 3, using the GFMT faces in combination with the eye-tracking task in Experiment 7,

may have meant that there was very little room for improvement. In support of this reasoning, it is notable that baseline accuracy was higher in Experiment 7, and the range in individual performance was also reduced (*M* = 94.47, range: 77.5-100), compared with Experiments 4-6 (*M* = 92.10, range: 50-100). As such, accuracy was more likely to decline during the eye-tracking task than to increase (see, e.g., Alenezi et al., 2015; Fysh & Bindemann, 2017b), which could account for the pattern of correlations observed in both the examples and no-examples conditions in Experiment 7.

The GFMT stimuli were selected for this thesis as they have been widely tested (see e.g., Bindemann, Avetisyan & Rakow, 2012; White, Burton, Kemp & Jenkins, 2013) and utilised in other training studies (see, e.g., Towler, White & Kemp, 2014; Towler, White & Kemp, 2017). Furthermore, the GFMT has also been utilised to assess the impact of providing feedback on unfamiliar face matching (see, e.g., Alenezi & Bindemann, 2013; White, Kemp, Jenkins & Burton, 2014). As examples may act as a form of simultaneous feedback, using the GFMT to evaluate examples makes the work in this thesis more comparable to the existing literature. Furthermore, while normative performance is high, individual differences in performance on the GFMT can range from near chance to perfect accuracy (see, e.g., Bindemann et al., 2012; Burton et al., 2010). Therefore, as individual differences in performance were the primary focus of this thesis, the GFMT was deemed to be an appropriate for evaluating the impact of the examples at an individual level.

Despite this, the ceiling effects obtained in Chapters 3 and 4 suggest using a more challenging stimulus set may help to clarify the potential of providing examples as a method for improving unfamiliar face matching accuracy. Replication of the experiments reported in Chapters 3 and 4 with more difficult stimuli sets that have recently become available such as the Models Face Matching Test (MFMT, Dowsett & Burton, 2015) or Kent Face Matching Test (KFMT, Fysh & Bindemann, 2018a), may help clarify how beneficial examples can be

for this task. Other studies have started to include more challenging stimuli sets in addition to or instead of the GFMT (see, e.g., Bobak, Dowsett et al., 2016; Fysh, 2018) in order to combat such ceiling effects. Furthermore, as the KFMT has a normative performance of around 60% and individual performance can range from 40 - 88%, using this stimulus set may allow high performers to also benefit from the examples as well as the lower-performing individuals who were shown to improve with the provision of GFMT examples.

In conclusion, the experiments reported in this thesis suggest that observers can employ different strategies in order to accurately categorise face pairs as identity matches or mismatches (Experiments 1-3). Furthermore, the identity-related information for reaching a decision may be based on different features according to the specific face at hand. Thus, training individuals to use specific strategies or features may not be an effective method of improving accuracy. As an alternative approach, the examples method described in this thesis may provide observers with a platform to increase accuracy by developing their own matching criteria. Examples were found to aid individuals with lower baseline performance (Experiments 4-6), however, this effect was less clear when observers were eye-tracked during the task (Experiment 7). Future research employing a more difficult stimulus set may further reveal whether an improvement can be found for low-performing individuals under an eye-tracking set up and clarify the benefit of the examples training for unfamiliar face matching.

References

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision, 16*(3), 40. doi: 10.1167/16.3.40

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

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

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: A reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 997-1010. doi: 10.1037/0278-7393.25.4.997

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, *3*(1), 25. doi: 10.1186/s41235-018-0114-7

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, *12*(3), 219-228. doi: 10.3758/BF03197669

Bateson, M., Callow, L., Holmes, J. R., Redmond Roche, M. L., & Nettle, D. (2013). Do images of 'watching eyes' induce behaviour that is more pro-social or more normative? A field experiment on littering. *PLoS ONE, 8*(12), e82055. doi: 10.1371/journal.pone.0082055

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412-414. doi: 10.1098/rsbl.2006.0509

Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology, 1*(1), 1-35. 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*(6), 707-717. doi: 10.1002/acp.2970

Bindemann, M., Avetisyan, M., & Blackwell, K.-A. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied, 16*(4), 378-386. doi: 10.1037/a0021893

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

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition, 1*(2), 96-103. doi: 10.1016/j.jarmac.2012.02.001

Bindemann, M., Fysh, M. C., Cross, K., & Watts, R. (2016). Matching faces against the clock. *I-Perception, 7*(5), 1-18. 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*(5), 625-627. doi: 10.1068/p7008

Bindemann, M., Scheepers, C., & Burton, A. M. (2009). Viewpoint and center of gravity affect eye movements to human faces. *Journal of Vision, 9*(2), 7. doi: 10.1167/9.2.7

Bindemann, M., Scheepers, C., Ferguson, H. J., & Burton, A. M. (2010). Face, body, and center of gravity mediate person detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 36*(6), 1477-1485. doi: 10.1037/a0019057

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008a). Gaze selection in complex social scenes. *Visual Cognition, 16*(2-3), 341-355. doi: 10.1080/13506280701434532

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008b). Social attention and real-world scenes: The roles of action, competition and social content. *Quarterly Journal of Experimental Psychology, 61*(7), 986-998. doi: 10.1080/17470210701410375

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research, 49*(24), 2992-3000. doi: 10.1016/j.visres.2009.09.014

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PLoS ONE, 3*(8), e3022. doi: 10.1371/journal.pone.0003022

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

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

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2018). Facing the facts: Naïve participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology*. Advance online publication. doi: 10.1177/1747021818776145

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology, 7*(1378), 1-11. doi: 10.3389/fpsyg.2016.01378

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

Border Force. (2018, June 29). Guide to faster travel through the UK border. Retrieved October 16, 2018, from https://www.gov.uk/government/publications/coming-to-the-uk/faster-travel-through-the-uk-border

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*(1), 105-116. doi: 10.1111/j.2044-8295.1982.tb01795.x

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339-360. doi: 10.1037/1076-898X.5.4.339

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught of CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207-218. doi: 10.1037/1076-898X.7.3.207

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

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

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

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

Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: A comparison across image formats. *Vision Research, 41*(24), 3185-3195. doi: 10.1016/S0042-6989(01)00186-9

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

Campbell, R. (1999). When does the inner-face advantage in familiar face recognition arise and why?. *Visual Cognition, 6*(2), 197-215. doi: 10.1080/713756807

Clarke, A. D. F., Mahon, A., Irvine, A., & Hunt, A. R. (2017). People are unable to recognize or report on their own eye movements. *The Quarterly Journal of Experimental Psychology, 70*(11), 2251-2270. doi: 10.1080/17470218.2016.1231208

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

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology, 17*(1), 97-116. doi: 10.1080/09541440340000439

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision, 10*(4), 16. doi: 10.1167/10.4.16

Davidoff, J., & Donnelly, N. (1990). Object superiority: A comparison of complete and part probes. *Acta Psychologica, 73*(3), 225-243. doi: 10.1016/0001-6918(90)90024-A

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology, 23*(4), 482-505. doi: 10.1002/acp.1490

Deffenbacher, K. A., Johanson, J., Vetter, T., & O'Toole, A. J. (2000). The face typicality-recognizability relationship: Encoding or retrieval locus?. *Memory & Cognition, 28*(7), 1173-1182. doi: 10.3758/BF03211818

Dhir, V., Singh, A., Kumar, R., & Singh, G. (2010). Biometric recognition: A modern era for security. *International Journal of Engineering Science and Technology*, *2*(8), 3364-3380. Retrieved from https://www.researchgate.net/publication/50315614_BIOMETRIC_RECOGNITION_A_MODERN_ERA_FOR_SECURITY

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*(2), 107-177. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.555.3596&rep=rep1&type=pdf

Donnelly, N., & Davidoff, J. (1999). The mental representations of faces and houses: Issues concerning parts and wholes. *Visual Cognition*, *6*(3-4), 319-343. doi: 10.1080/135062899395000

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

Ellis, H. D. (1986). Face recall: A psychological perspective. *Human Learning: Journal of Practical Research & Applications, 5*(4), 189-196.

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

Estudillo, A. J., & Bindemann, M. (2017). Can gaze-contingent mirror-feedback from unfamiliar faces alter self-recognition?. *The Quarterly Journal of Experimental Psychology*, *70*(5), 944-958. doi: 10.1080/17470218.2016.1166253

Fathi, M., Bateson, M., & Nettle, D. (2014). Effects of watching eyes and norm cues on charitable giving in a surreptitious behavioural experiment. *Evolutionary Psychology*, *12*(5), 878-887. doi: 10.1177/147470491401200502

Favelle, S., Hill, H., & Claes, P. (2017). About face: Matching unfamiliar faces across rotations of view and lighting. *i-Perception*, *8*(6). doi: 10.117/2041669517744221

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

FRONTEX. (2015). Best practice technical guidelines for automated border control (ABC) systems. Retrieved from https://frontex.europa.eu/assets/Publications/Research/Best_Practice_Technical_Guidelines_ABC.pdf

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

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

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

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

Fysh, M. C., & Bindemann, M. (2018b). Human-computer interaction in face matching. *Cognitive Science*. Advance online publication. doi: 10.1111/cogs.12633

Gilad, S., Meng, M., & Sinha, P. (2009). Role of ordinal contrast relationships in face encoding. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(13), 5353-5358. doi: 10.1073/pnas.0812396106

Goffaux, V., & Rossion, B. (2006). Faces are “spatial” -- holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology*: *Human Perception and Performance, 32*(4), 1023-1039. doi: 10.1037/0096-1523.32.4.1023

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research, 41*(17), 2261-2271. doi: 10.1016/S0042-6989(01)00097-9

Heathrow Airport Limited. (2018). *Company information: Facts and figures* [Fact sheet]. Retrieved from https://www.heathrow.com/company/company-news-and-information/company-information/facts-and-figures

Heisz, J. J., & Shore, D. I. (2008). More efficient scanning for familiar faces. *Journal of Vision, 8*(1), 9. doi: 10.1167/8.1.9

Henderson, J. M. (2003). Human faze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11), 498-504. doi: 10.1016/j.tics.2003.09.006

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16*(4), 219-22. doi: 10.1111/j.1467-8721.2007.00507.x

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

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

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

Hills, P. J., Cooper, R. E., & Pake, J. M. (2013). Removing the own-race bias in face recognition by attentional shift using fixation crosses to diagnostic features: An eye-

tracking study. *Visual Cognition, 21*(7), 876-898. doi: 10.1080/13506285.2013.834016

Hills, P. J., & Pake, J. M. (2013). Eye-tracking the own-race bias in face recognition: Revealing the perceptual and socio-cognitive mechanisms. *Cognition, 129*(3), 586-597. doi: 10.1016/j.cognition.2013.08.012

Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception, 23*(1), 65-74. doi: 10.1068/p230065

Hsiao, J. H. W., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological Science, 19*(10), 998-1006. doi: 10.1111/j.1467-9280.2008.02191.x

Hussain, Z., Sekuler, A. B., & Bennett, P. J. (2009). Perceptual learning modifies inversion effects for faces and textures. *Vision Research, 49*(18), 2273-2284. doi: 10.1016/j.visres.2009.06.014

Jenkins, R., & Burton, A. M. (2008a). 100% accuracy in automatic face recognition. *Science, 319*(5862), 435. doi: 10.1177/03063127067078012

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

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

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

Keil, M. S. (2009). “I look in your eyes, honey”: Internal face features induce spatial frequency preference for human face processing. *PLoS ONE Computational Biology*, *5*(3), e1000329. doi: 10.1371/journal.pcbi.1000329

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

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

Kok, E. M., Aizenman, A. M., Võ, M. L.-H., & Wolfe, J. M. (2017). Even if I showed you where you looked, remembering where you just looked is hard. *Journal of Vision*, *17*(12), 1-11. doi: 10.1167/17.12.2

Kokje, E., Bindemann, M., & Megreya, A. M. (2018). Cross-race correlations in the abilities to match unfamiliar faces. *Acta Psychologica*, *185*, 13-21. doi: 10.1016/j.actpsy.2018.01.006

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

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychological Science*, *15*(11), 762-768. doi: 10.1111/j.0956-7976.2004.00753.x

Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science*, *11*(5), 379-385. doi: 10.1111/1467-9280.00274

Livingston, L. A., & Shah, P. (2017). People with and without prosopagnosia have insight into their face recognition ability. *Quarterly Journal of Experimental Psychology*, *71*(5), 1260-1262. doi: 10.1080/17470218.2017.1310911

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 77-100. doi: 10.1037/0096-1523.34.1.77

Luria, S. M., & Strauss, M. S. (1978). Comparison of eye movements over faces in photographic positives and negatives. *Perception*, *7*(3), 349-358. doi: 10.1068/p070349

Mahon, A., Clarke, A. D. F., & Hunt, A. R. (2018). The role of attention in eye-movement awareness. *Attention, Perception, & Psychophysics*, *80*(7), 1691-1704. doi: 10.3758/s13414-0181553-4

Mansour, J. K., & Flowe, H. D. (2010). Eye-tracking and eyewitness memory. *Forensic Update*, *101*, 11–15. Retrieved from https://dspace.lboro.ac.uk/2134/20322

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

McIntyre, A. H., Hancock, P. J., Kittler, J., & Langton, S. R. (2013). Improving discrimination and face matching with caricature. *Applied Cogntive Psychology*, *27*(6), 725-734. doi: 10.1002/acp.2966

McKone, E. (2004). Isolating the special component of face recognition: Peripheral identification and a Mooney face. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, *30*(1), 181-197. doi: 10.1037/0278-7393.30.1.181

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

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

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS ONE*, *13*(3), e0193455. doi:10.1371/journal.pone.0193455

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

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

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

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

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364-372. doi: 10.1037/a0013464

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

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

Meinhardt-Injac, B., Persike, M., & Meinhardt, G. (2010). The time course of face matching by internal and external features: Effects of context and inversion. *Vision Research*, *50*(16), 1598-1611. doi: 10.1016/j.visres.2010.05.018

Meinhardt-Injac, B., Persike, M., & Meinhardt, G. (2011). The context effect in face matching: Effects of feedback. *Vision Research*, *51*(19), 2121-2131. doi: 10.1016/j.visres.2011.08.004

Mileva, M., & Burton, A. M. (2018). Smiles in face matching: Idiosyncratic information revealed through a smile improves unfamiliar face matching performance. *British Journal of Psychology*, *109*(4), 799-811. doi: 10.1111/bjop.12318

Millen, A. E., Hope, L., Hillstrom, A. P., & Vrij, A. (2017). Tracking the truth: The effect of face familiarity on eye fixations during deception. *The Quarterly Journal of Experimental Psychology*, *70*(5), 930-943. doi: 10.1080/17470218.2016.1172093

Miyazaki, Y. (2013). Increasing visual search accuracy by being watched. *PLoS ONE*, *8*(1), e53500. doi: 10.1371/journal.pone.0053500

Moore, R. M., & Johnston, R. A. (2013). Motivational incentives improve unfamiliar face matching accuracy. *Applied Cognitive Psychology*, *27*(6), 754-760. doi: 10.1002/acp.2964

Nettle, D., Nott, K., & Bateson, M. (2012). 'Cycle thieves, we are watching you': Impact of a simple signage intervention against bicycle theft. *PLoS ONE*, *7*(12), e51738. doi: 10.1371/journal.pone.0051738

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review, 84*(3), 231-259. doi: 10.1037/0033-295X.84.3.231

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

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*(1), 44-64. doi: 10.1016/0010-0285(75)90004-3

O'Donnell, C., & Bruce, V. (2001). Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception, 30*(6), 755-764. doi: 10.1068/p3027

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

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

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

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

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

Peissig, J., Goode, T., & Smith, P. (2009). The role of eyebrows in face recognition: With, without, and different. *Journal of Vision, 9*(8), 554-554. doi: 10.1167/9.8.554

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422. Retrieved from http://www.infinitychallenge.com/clamlist/Rayner_1998.pdf

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

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

Robertson, D. J., Middleton, R., & Burton, A. M. (2015). From policing to passport control: The limitations of photo ID. *Keesing Journal of Documents and Identity*, *February*, 3-8. Retrieved from https://www.researchgate.net/profile/David_Robertson31/publication/305429407_From_policing_to_passport_control_The_limitations_of_photo_ID/links/578e690c08aecbca4caad01a.pdf

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

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

Sadr, J., Jarudi, I., & Sinha, P. (2003). The role of eyebrows in face recognition. *Perception*, *32*(3), 285-293. doi: 10.1068/p5027

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

Schulz, C., Kaufmann, J. M., Walther, L., & Schweinberger, S. R. (2012). Effects of anticaricaturing vs. caricaturing and their neural correlates elucidate a role of shape for face learning. *Neuropsychologia*, *50*(10), 2426-2434. doi: 10.1016/j.neuropsychologia.2012.06.013

Smilek, D., Birmingham, E., Cameron, D., Bischof, W., & Kingstone, A. (2006). Cognitive ethology and exploring attention in real-world scenes. *Brain Research*, *1080*(1), 101-119. doi: 10.1016/j.brainres.2005.12.090

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology*, *46*(2), 225-245. doi: 10.1080/14640749308401045

Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & Cognition*, *25*(5), 583-592. Retrieved from: https://www.researchgate.net/profile/J_Tanaka3/publication/13889704_Features_and_their_configuration_in_face_recognition/links/55c0c2df08ae092e9666e0d2/Features-and-their-configuration-in-face-recognition.pdf

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

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

Van Belle, G., De Graef, P., Verfaillie, K., Busigny, T., & Rossion, B. (2010). Whole not hole: Expert face recognition requires holistic perception. *Neuropsychologia*, *48*(9), 2620-2629. doi: 10.1016/j.neuropsychologia.2010.04.034

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

Vinette, C., Gosselin, F., & Schyns, P. G. (2004). Spatio-temporal dynamics of face recognition in a flash: It's in the eyes. *Cognitive Science*, *28*(2), 289-301. doi: 10.1016/j.cogsci.2004.01.002

Võ, M. L.-H., Aizenman, A. M., & Wolfe, J. M. (2016). You think you know where you looked? You better look again. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(10), doi: 10.1037/xhp0000264

Walker-Smith, G. J., Gale, A. G., & Findlay, J. M. (2013). Eye movement strategies involved in face perception. *Perception*, *42*(11), 1120-1133. doi: 10.1068/p060313n

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

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

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

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

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

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

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 1814-1822. doi: 10.1098/rspb.2015.1292

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, *55*, 493-518. doi: 10.1146/annurev.psych.55.090902.141954

Winograd, E. (1981). Elaboration and distinctiveness in memory for faces. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(3), 181-190. doi: 10.1037/0278-7393.7.3.181

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

Young, A. W., McWeeny, K. H., Hay, D. C., & Ellis, A. W. (1986). Matching familiar and unfamiliar faces on identity and expression. *Psychological research*, *48*(2), 63-68. doi: 10.1007/BF00309318

Appendix A

Experiment 1: *d'* and *criterion*

The accuracy data for the features was also translated into signal detection measures of sensitivity (*d*') and bias (*criterion*), which are illustrated in Figure A1. For *d'* a one-way repeated measures ANOVA revealed a main effect of region, $F(4,88) = 56.59$, $p < .001$, $\eta_p^2 = .72$. A series of paired samples *t*-tests (with alpha corrected to .05/10 = .005 for ten comparisons) revealed higher *d'* for the whole face compared to the hair, eyes, nose and mouth, all $ts \geq 6.58$, $ps \leq .001$. In addition, *d'* was higher for the eyes than the hair, nose and mouth, all $ts \geq 4.12$, $ps < .001$. No other comparisons were significant, all $ts \leq 2.64$, $ps \geq .02$. Thus, sensitivity for the whole faces exceeds that of any of the individual features. Furthermore, sensitivity for the eyes is higher than the other facial features.

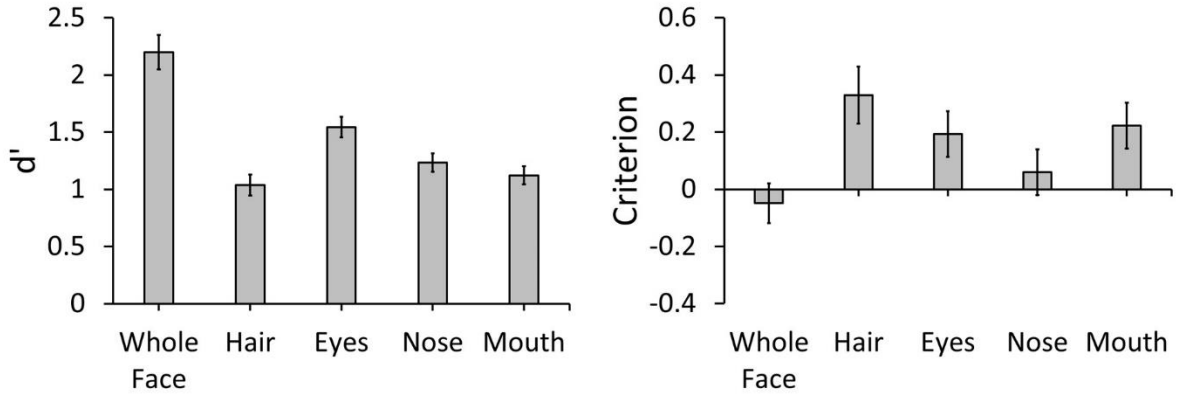

FIGURE A1. *Sensitivity (d') and bias (criterion) for the examples and no-examples conditions in Experiment 1.*

For *criterion*, an analogous ANOVA found a main effect of region, $F(4,88) = 5.98$, $p < .001$, $\eta_p^2 = .21$. A series of paired samples *t*-tests (with alpha corrected to .05/10 = .005 for ten comparisons) revealed a change in criterion from whole face to hair, $t(22) = 3.37$, $p < .005$. Whereby, bias towards match decisions changed to bias towards mismatch decisions. In

addition, bias towards mismatch decisions was reduced for the nose compared to the hair, $t(22) = 3.44$, $p < .005$. No other comparisons were significant, all ts ≤ 3.09, ps $\geq .01$.

The mean *criterion* values were also compared to zero to determine whether there was a bias towards a particular response (match or mismatch) for any of the features. One-sample *t*-tests revealed a mismatch bias for the hair, $t(22) = 3.15$, $p < .01$, the eyes, $t(22) = 2.49$, $p < .05$, and mouth, $t(22) = 2.76$, $p < .05$. Thus, observers were more likely to classify these feature pairs as belonging to two different individuals. No such bias was found for the whole face or nose, both ts ≤ 0.70, ps $\geq .49$.

Experiment 2: *d'* and *criterion*

As in Experiment 1, the accuracy data for the features was translated into signal detection measures of sensitivity (*d*') and bias (*criterion*), which are illustrated in Figure A2. For *d'* a one-way repeated measures ANOVA revealed a main effect of region, $F(4,76) = 72.17$, $p < .001$, $\eta_p^2 = .79$. A series of paired samples *t*-tests (with alpha corrected to .05/10 = .005 for ten comparisons) revealed higher *d'* for the whole face compared to the hair, eyes, nose and mouth, all ts ≥ 7.45, ps $\leq .001$. In addition, *d'* was higher for the eyes than the hair, nose and mouth, all ts ≥ 5.07, ps $\leq .001$. No other comparisons were significant, all ts ≤ 2.85, ps $\geq .01$. Hence, sensitivity for the whole face was higher than for all of the isolated features. Sensitivity for the eyes also exceeded that of the other features.

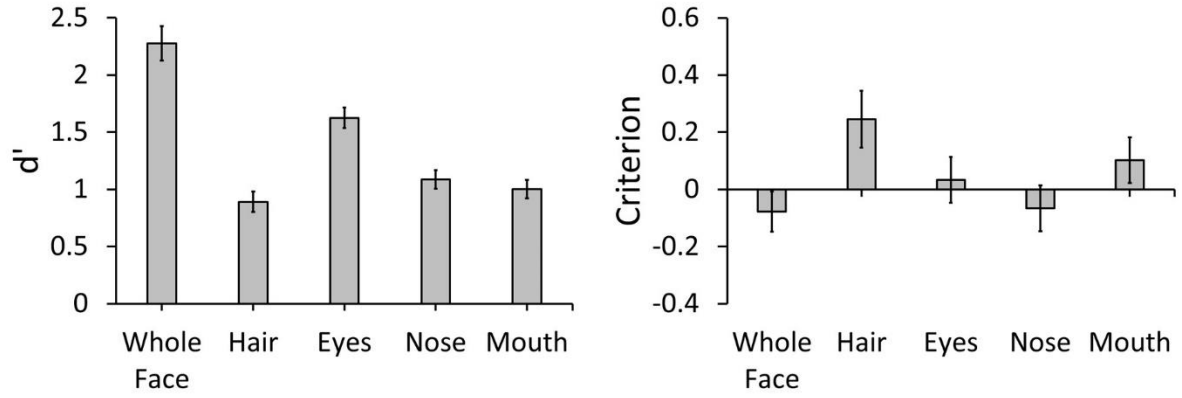

FIGURE A2. *Sensitivity (d') and bias (criterion) for the examples and no-examples conditions in Experiment 2.*

For *criterion*, an analogous ANOVA found a main effect of feature, $F(4,76) = 4.34$, $p < .01$, $\eta_p^2 = .19$. A series of paired samples *t*-tests (with alpha corrected to .05/10 = .005 for ten comparisons) did not reveal a difference in criterion for any of the features, all $ts \leq 3.19$, $ps \geq .01$.

The mean *criterion* values were also compared to zero to determine whether there was a bias towards a particular response (match or mismatch) for any of the features. One-sample *t*-tests did not reveal a bias for the whole face, eyes, nose or mouth, all $ts \leq 1.02$, $ps \geq .32$. For the hair a bias towards a mismatch response was approaching significance, $t(22) = 2.05$, $p = .05$. Thus, the addition of the 'don't know' option is likely to have reduced the mismatch bias found for most of the features in Experiment 1.

Experiment 3: *d'* and *criterion*

The accuracy data for the different face types was converted into signal detection measures of sensitivity (*d*') and (*criterion*), which are illustrated in Figure A3. For *d*' a one-way repeated measures ANOVA revealed a main effect of face type, $F(2,46) = 9.23$, $p < .001$, $\eta_p^2 = .29$. A series of paired samples *t*-tests (with alpha corrected to .05/3 = .017 for three comparisons) revealed higher *d'* for whole face pairs compared to split face pairs, *t*(23)

= 4.40, $p < .001$, and part face pairs, $t(23) = 3.23$, $p < .01$. For split and part face pairs, *d'* was similar, $t(23) = 0.23$, $p = .82$.

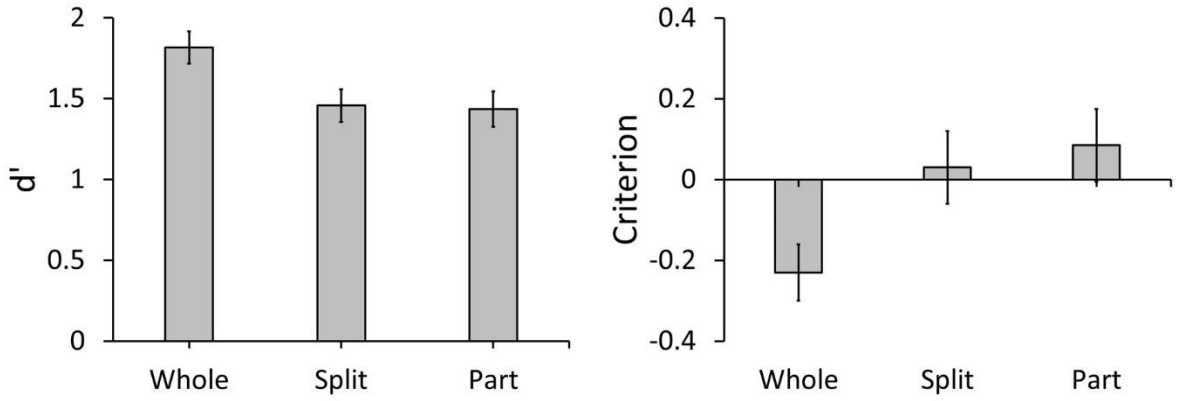

FIGURE A3. *Sensitivity (d') and bias (criterion) for the examples and no-examples conditions in Experiment 3.*

For *criterion*, an analogous ANOVA found a main effect of face type, $F(2,46) = 6.47$, $p < .01$, $\eta_p^2 = .22$. A series of paired samples *t*-tests (with alpha corrected to .05/3 = .017 for three comparisons) revealed a change in *criterion* from whole face to part face, $t(23) = 3.35$, $p < .017$. Whereby, bias towards match decisions was reduced and shifted to a mismatch bias. A similar pattern was observed for the whole face and split face pairs, and was approaching significance, $t(23) = 2.56$, $p = .018$. *Criterion* was comparable for split face and part face pairs, $t(23) = 0.65$, $p = .52$.

The *criterion* values for each face type were also compared to zero to determine whether there was a significant bias towards either trial type (match or mismatch). One-sample *t*-tests revealed a match bias for whole face pairs, $t(23) = 3.17$, $p < .01$, but no bias was found for split face or part face pairs, both $ts \leq 0.96$, $ps \geq .35$.

Appendix B

Experiment 4: *d'* and *criterion*

The accuracy data was also converted into signal detection measures of sensitivity (*d'*) and bias (*criterion*), which are illustrated in Figure B1. For *d'*, a 2 (block) x 2 (condition) mixed-factor ANOVA did not show a main effect of condition, $F(1,58) = 0.07$, $p = .79$, $\eta_p^2 = .00$, but revealed a main effect of block, $F(1,58) = 6.07$, $p < .05$, $\eta_p^2 = .10$, and an interaction between factors, $F(1,58) = 5.04$, $p < .05$, $\eta_p^2 = .08$. Analysis of simple main effects showed that *d'* was comparable across conditions in Block 1, $F(1,58) = 1.12$, $p = .29$, $\eta_p^2 = .02$, and in Block 2, $F(1,58) = 0.36$, $p = .55$, $\eta_p^2 = .01$. In addition, *d'* was comparable across Block 1 and Block 2 in the no-examples condition, $F(1,58) = 0.02$, $p = .88$, $\eta_p^2 = .00$. In contrast, *d'* increased from Block 1 to Block 2 in the examples condition, $F(1,58) = 11.08$, $p < .01$, $\eta_p^2 = .16$.

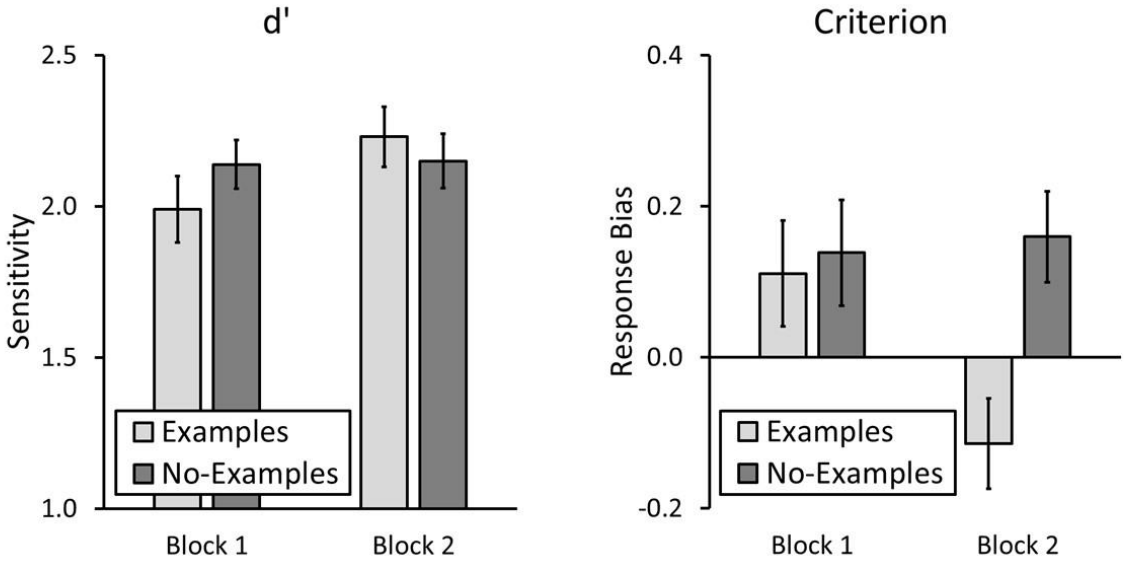

FIGURE B1. *Sensitivity (d') and bias (criterion) for the examples and no-examples conditions in Experiment 4.*

For *criterion*, an analogous ANOVA found no main effect of condition, $F(1,58) = 3.42$, $p = .07$, $\eta_p^2 = .06$, but a main effect of block, $F(1,58) = 6.55$, $p < .05$, $\eta_p^2 = .10$, and an

interaction between factors, $F(1,58) = 9.60$, $p < .01$, $\eta_p^2 = .14$. Analysis of simple main effects showed that *criterion* was comparable across conditions in Block 1, $F(1,58) = 0.08$, $p = .77$, $\eta_p^2 = .00$, but not in Block 2, $F(1,58) = 9.50$, $p < .01$, $\eta_p^2 = .14$. In addition, *criterion* was comparable across Block 1 and 2 in the no-examples condition, $F(1,58) = 0.15$, $p = .70$, $\eta_p^2 = .00$. By contrast, a change in *criterion* was observed from Block 1 to Block 2 in the examples condition, $F(1,58) = 16.00$, $p < .001$, $\eta_p^2 = .22$, whereby observers' initial bias towards making mismatch decisions was reduced and they became more likely to make identity-match responses.

Experiment 5: *d'* and *criterion*

The accuracy data was also converted into *d'* and *criterion* (see Figure B2). For *d'*, a 2 (condition) x 4 (block) mixed-factor ANOVA, did not show a main effect of block, $F(3,174) = 0.73$, $p = .54$, $\eta_p^2 = .01$, or a main effect of condition, $F(1,58) = 0.48$, $p = .49$, $\eta_p^2 = .01$, or an interaction between these factors, $F(3,174) = 0.25$, $p = .86$, $\eta_p^2 = .00$.

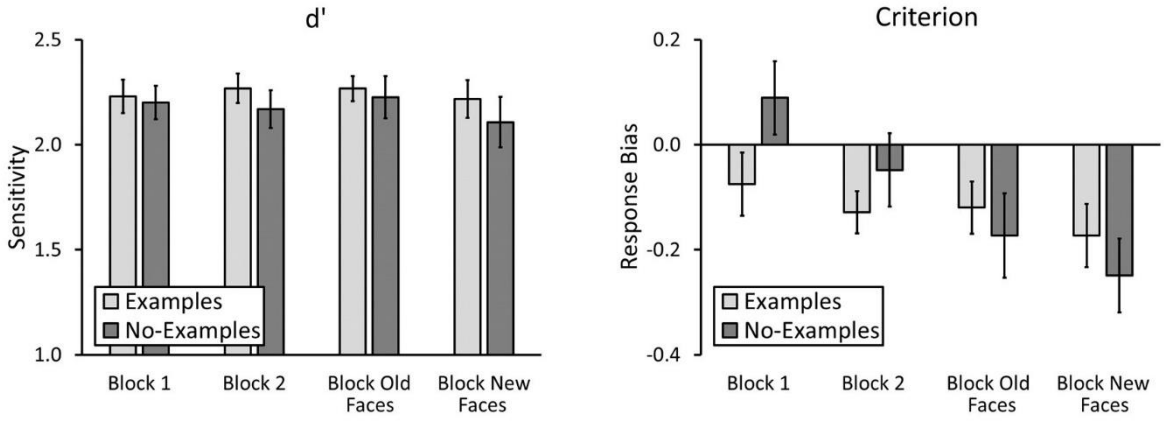

FIGURE B2. *Sensitivity (d') and bias (criterion) for the examples and no-examples conditions in Experiment 5.*

For *criterion*, an analogous ANOVA found no main effect of condition, $F(1,58) = 0.14$, $p = .71$, $\eta_p^2 = .00$, but a main effect of block, $F(3,174) = 11.16$, $p < .001$, $\eta_p^2 = .16$, and

an interaction between factors, $F(3,174) = 4.21$, $p < .01$, $\eta_p^2 = .07$. Further analysis revealed no simple main effects of condition for any of the blocks, all Fs ≤ 3.33, ps $\geq .07$, and no simple main effect of block for the examples condition, $F(3,56) = 0.85$, $p = .48$, $\eta_p^2 = .04$. However, a simple main effect of block was observed for the no-examples condition, $F(3, 56) = 12.15$, $p < .001$, $\eta_p^2 = .39$. A series of paired-samples *t*-tests (with alpha corrected to .05/6 = .008 for six comparisons) revealed an increased bias towards an identity-match response from Block 1 to Block Old Faces, $t(29) = 4.50$, $p < .001$ and Block New Faces, $t(29) = 5.95$, $p < .001$. There was also an increase in match bias from Block 2 to Block Old Faces, $t(29) = 3.75$, $p < .008$, and Block New Faces, $t(29) = 4.17$, $p < .001$. No other comparisons were significant, all ts ≤ 2.21, ps $\geq .04$.

Experiment 6: *d'* and *criterion*

The accuracy data was also translated into signal detection measures of sensitivity (*d'*) and bias (*criterion*), and is illustrated in Figure B3. For *d'*, a 2 (condition) x 4 (block) mixed-factor ANOVA did not reveal a main effect of condition, $F(1,58) = 0.02$, $p = .90$, $\eta_p^2 = .00$, or an interaction between condition and block, $F(3,174) = 0.88$, $p = .45$, $\eta_p^2 = .02$. However, ANOVA revealed a main effect of block, $F(3,174) = 482.66$, $p < .001$, $\eta_p^2 = .89$. A series of paired-samples t-tests (with alpha corrected to .05/6 = .008 for six comparisons) indicated this was due to a decrease in *d'* in Block KFMT compared to Block 1, $t(59) = 26.45$, $p < .001$, Block 2, $t(59) = 29.57$, $p < .001$ and Block GFMT, $t(59) = 24.72$, $p < .001$. Furthermore, *d'* was lower in Block 1 compared to Block 2, $t(59) = 3.53$, $p < .008$, and Block GFMT, $t(59) = 3.21$, $p < .008$. There was no change in *d'* from Block 2 to Block GFMT, $t(59) = 0.21$, $p = .83$.

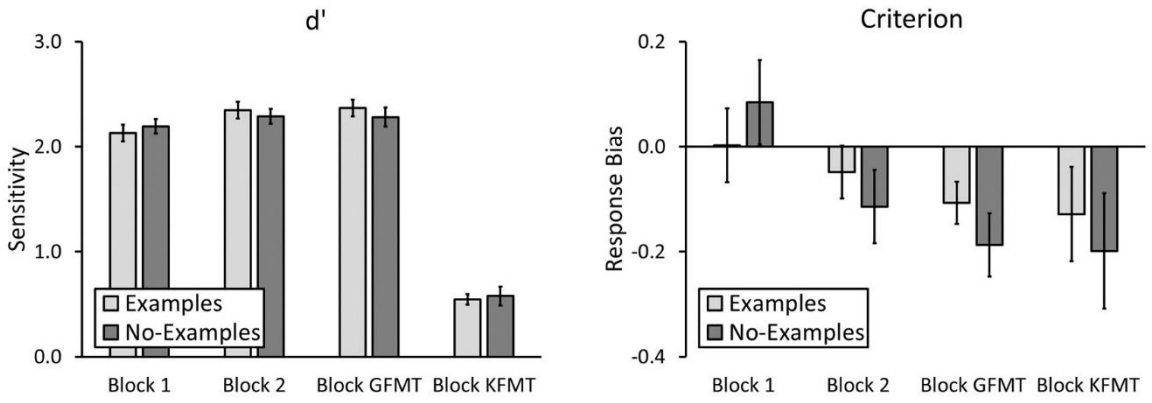

FIGURE B3. *Sensitivity (d') and bias (criterion) for the examples and no-examples conditions in Experiment 6.*

For *criterion*, an analogous ANOVA did not reveal a main effect of condition, $F(1,58) = 0.16$, $p = .69$, $\eta_p^2 = .00$, or an interaction between condition and block, $F(3,174) = 1.17$, $p = .32$, $\eta_p^2 = .02$. However, ANOVA found a main effect of block, $F(3,174) = 6.94$, $p < .001$, $\eta_p^2 = .11$. A series of paired-sample *t*-tests (with alpha corrected to .05/6 = .008 for six comparisons) indicated this was due to shift in bias from mismatch in Block 1 to match in Block 2, $t(59) = 2.93$, $p < .008$, Block GFMT, $t(59) = 3.83$, $p < .001$, and Block KFMT, $t(59) = 3.25$, $p < .008$. No other comparisons were significant, all $ts \leq 2.15$, $ps \geq .04$.