Within-Person Variability in Facial Appearance

Heather J. Cursiter

School of Psychology

University of Glasgow

Submitted for the Degree of Ph.D. to the Higher Degree Committee of the College of Science and Engineering, University of Glasgow

May 2013

Abstract

Within-person variability has largely been neglected in face processing research, with research typically focusing instead on between-person variability. In experimental settings between-person variability often becomes between-image variability with research using one image to represent a face. This implies that an image is an adequate representation of a face, however one image cannot illustrate the variability that can occur in facial appearance. This thesis argues that overlooking within-person variability is a fundamental flaw in face processing research, as within-person variability is surprisingly large. The experiments in this thesis illustrate the effect of within-person variability in different face processing contexts: face identification, face perception and image memory. Experiments 1 – 7 demonstrate the difficulty of identifying familiar and unfamiliar faces across within-person variability using an image-sorting task. Experiment 8 illustrates the within-person variability that occurs in personality perception, and Experiments 9 and 10 illustrate the within-person variability that occurs in the perception of facial attractiveness. Lastly Experiments 11 and 12 introduce within-person variability to memory recognition and demonstrate the difficulties of remembering multiple images of the same face. From the results of Experiments 1 – 12 it is concluded that within-person variability is highly influential in all the discussed areas of face processing and therefore needs to be taken seriously in face processing research.

Acknowledgements

I would like to thank Dr Rob Jenkins, Professor A. Mike Burton and Professor Helmut Leder for their supervision, Dr David White and Robert Hazod for their help in recruiting participants for Experiments 8 and 9, and my family for their support.

Declaration

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

……………………………………………….

Heather Judith Cursiter

Table of Contents

List of Figures

Chapter One: Within-person variability and face processing: A General Introduction

Chapter Two: Within-person variability in face identification

Chapter Three: Between-person similarity in face identification

Chapter Four: Within-person variability in social dimensions

Chapter Five: Within-person variability in facial attractiveness

Chapter Six: Image Memory

Appendix A

List of Tables

Chapter One

Within-Person Variability and Face Processing: A General Introduction

1.1 General Introduction

Current understanding of face recognition is based almost exclusively on research of image recognition. Many research studies have represented faces by a single image that may be repeated throughout the experiment, or by pairs of images that differ only by one element, such as facial expression or pose. For example in recognition memory research and social perceptions research one image is often used to represent a face (Goldstein & Chance 1970; Bainbridge, Isola, Blank & Oliva 2012; Antonakis & Dalgas 2009; Little, Burriss, Jones & Roberts 2007) and in face identification and recognition research, two images are typically presented for participants to 'match' or remember. The two images are often varied by dimensions such as image quality (Burton, Wilson, Cowan & Bruce 1999; Bruce, Henderson, Greenwood, Hancock, Burton & Miller 1999; Bruce, Henderson, Newman & Burton 2001), expression (Bruce 1982) or pose (Bruce 1982; Clutterbuck & Johnston 2002, 2004).

In this thesis it is argued that comparing an image to a face is a serious misinterpretation of the face recognition problem. It is also argued that the approach of using images to represent faces is one that gives rise to misleading findings. This approach results in theories that miss fundamental issues, such as within-person variability. It is important to underline the definitions of the terms face and image used in this thesis. Face refers to a person, specifically their face, whereas image refers to a photograph of a face i.e. one set view of a face. The term identity is also frequently used; identity also refers to a person, and the term is used in identification tasks to clarify participants' instructions.

Face recognition is often characterised as a problem of telling different people apart. Given that all faces share the same basic template (two eyes above a nose above a mouth), it is perhaps surprising that we can distinguish between thousands of individuals. Previous research states that we are all face experts due to the number of different faces we can distinguish (Carey, De Schonen & Ellis 1992; Carey & Diamond 1977). Face expertise research has also shown that faces are typically recognised with identity-specific categorisation labels, such as the person's name, as opposed to basic level categorisation, such as 'man' or 'woman' (Tanaka 2001).

This suggests that face recognition is therefore a problem of within-category discrimination. This results in an emphasis on differences between individuals and so encourages research to focus on between-person variability in face recognition. In experimental settings between-person variability becomes between-image variability. The substitution of images for faces implies that an image can adequately capture a person's facial appearance, and suggests that exposure to an image is interchangeable with exposure to a face.

The idea that an image adequately represents a face is challenged in this thesis. It is proposed that an image is not a reliable indicator of facial appearance because it is blind to within-person variability. Face processing research typically overlooks within-person variability but it turns out that within-person variability is surprisingly large.

1.2 Within-person variability

When we see two photographs of the same person we can easily make simple observations as to how the images differ. These differences might be due to the face undergoing non-rigid changes in the short-term, such as muscular movement, and in the long-term, such as ageing. Face characteristics also cause variation, for example one image may show a person with facial hair while the other image shows the same person clean-shaven. Alternatively these differences might be due to image characteristics superimposed onto faces, for example the images may differ in quality of lighting. Finally images can also vary dependent on variations in the camera used to capture the image.

These differences are all examples of within-person variability. Considering the differences that can occur between just two images, it is important to consider differences that can occur between many different images of a face. To demonstrate within-person variability across many images, twenty images of one face are shown in Figure 1.1. From Figure 1.1 it is possible to see that within-person variability can occur across a number of different factors. These factors can be classified into three distinct categories: *1).* Variations in face characteristics, *2).* Environmental variations and *3).* Camera variations.

***Figure 1.1.* Within-person variability.**
An example of within-person variability is shown in twenty different images of one face.

1). Variations in face characteristics

Differences in facial appearance can have a big effect on a person's appearance. Differences can include superficial changes such as a change in hairstyle or hair colour, changes in facial hair or changes in make-up worn. Changes in facial appearance can also be influenced by facial structure. Long-term changes such as aging, health or weight can alter facial structure and so change facial appearance. For example research into facial symmetry has shown fluctuations in symmetry can occur with stress or health challenges during development (Perrett, et al., 1999). Structural facial changes can also occur with weight loss or weight gain. Health can also affect factors such as skin colour or texture (Coetzee, Perrett & Stephen 2009); a common example might be the effect tiredness can have on facial appearance.

Short-term changes in facial structure can also result in face variability, for example the muscular movements and changes that occur with different facial expressions. Research has explored variations in facial expressions through building databases of images that contain multiple images of different faces showing different expressions and used these images to analyse automatic face recognition performance but not human face recognition (Huang, Mattar, Berg & Learned-Miller 2008; Gross, Matthews, Cohn, Kanade & Baker 2010; Yang, Zhang, Frangi & Yang 2004). Research has also explored the effect different facial expressions have on the perception of attractiveness between-persons (Morrison, Morris & Bard 2013). However research typically focuses on deliberately controlled expressions as opposed to spontaneous expression, and explores variation in expressions between-persons not within-person.

Gaze direction and head pose are other aspects of within-person variability. For example, research has shown that faces with direct gaze are recognised better than faces with diverted gaze (Hood, Macrae, Cole-Davis & Dias 2003). Similarly research has shown that head position is an important factor in recognition of a face. For example, research has shown that faces are easier to recognise from a ¾ angle than from a frontal pose or profile pose (Logie, Baddeley & Woodhead 1987). Different head positions change the perspective of facial features and can emphasise or obscure different features. Together with changes in lighting, a difference in head pose can have a big effect on the perception of facial characteristics such as head or face shape.

2). Environmental variations

Changes in the environment such as lighting can have a big effect on the variability of an image. Variations in lighting can affect facial appearance in images and in person. Different sources of light can exaggerate or obscure facial features and lighting variations can influence the shape or appearance of facial features. Lighting is an element frequently controlled in experiments yet can have a substantial impact on the recognition of a face. Research using computer-generated representations of faces and 3-D face models has found lighting plays a significant role in face recognition (Hill & Bruce 1996; Chen, Chen & Tyler 2013).

Another environmental variation that may affect face recognition and identification is scene complexity. For example, object recognition research has explored the interaction between objects in the foreground and background of images; objects in context-consistent settings are identified more accurately than in inconsistent settings (Davenport & Potter 2004). This could occur with faces and cause problems in everyday life. For example we may meet a work colleague while walking down the street but because the face is in a new context, or environment, we cannot immediately place the face or identify the person (O'Toole, Abdi, Deffenbacher & Valentin 1993; Young, Hay & Ellis 1985). Face recognition research has highlighted scene complexity as an issue for automated recognition systems but not for human face recognition (Kanade, Cohn & Tian 2000).

3). Camera Variations

Variations in camera are image specific; there are several camera variations that can have an impact on variability in images. For example focal length i.e. the distance between the camera lens and the target can have a big impact on an image (Harper & Latto 2001). Shown in Figure 1.2 are three images of the same face, taken during the same session with only the focal length varied between 0.5 and 3 metres in length. The resulting images vary in face characteristics such as face shape and length. Different types of lenses can also have varying effects of images, for example research has shown that wide-angled lenses have a 'slimming' effect on appearance (Harper & Latto 2001).

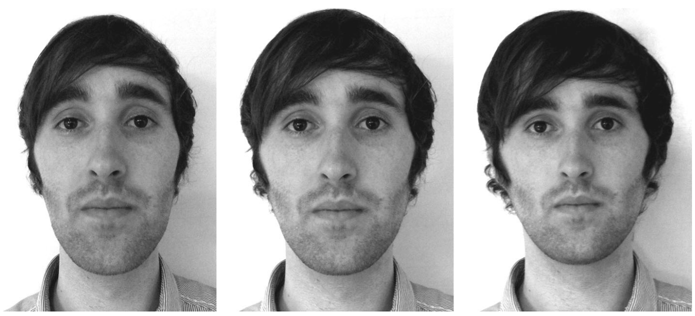


Figure 1.2. **Variability in a camera's focal length.**
An example of the effect of different focal lengths on three images of the same face (Burton, in press).

Different cameras can also have a surprising effect on the image produced. For example shown in Figure 1.3 are two sets of images taken on two different cameras, the poses are matched across camera with one small change in the person's presentation (hairstyle). The photos were taken within minutes of each other but despite this the images look quite different (Burton, in press). Research has shown that small changes in camera parameters can effect face recognition, for example, in the Glasgow Face Matching Test pairs of images of the same faces were captured on two different cameras within minutes of each other. Despite controlling for different variability such as expression and pose, when participants were asked to match pairs of same face images accuracy was as low as 62%, with a mean of 89%. Therefore enough variability occurred between the two cameras to affect participants' face matching performance (Burton, White & McNeil 2010).

Figure 1.3. **Variability between different cameras.**
An example of variations in images taken within minutes of each other on different cameras with just one change in presentation (hairstyle): the two images in the top row were taken with one camera, and the two images on the bottom row were taken on a different camera (Burton, in press).

Within-person variability is unavoidable in the real world, because no face casts the same image twice. Despite the extent of within-person variability, the natural range of facial appearance is rarely considered in experiments. Instead research strives to minimise image variability, using images that have controlled for any variability other than what is of interest to manipulate; variability is treated as 'noise' that can get in the way of the real problem. The images used in experimental settings therefore end up being far from the within-person variability that occurs with real faces. This response to image variability creates a fundamental disconnect between the process that we would like to understand and the process that is tested.

1.3 Previous studies of face recognition

A few studies have shown the recognition and identification difficulties that can occur across specific image manipulations. For example in an experiment by Bruce (1982) face recognition was explored across changed and unchanged images. A set of four different images of 48 faces was collected. The images were controlled for pose (front facing or at a ¾ angle to the camera) and expression (smiling or unsmiling). Participants were shown one image of each face for a period of eight seconds, and 15 minutes later were shown either the same image, an image that had one change (expression or pose) or an image that had two changes (expression and pose), and were asked to remember if they had seen the person before. Participants were made aware that they would be asked to remember the faces following first exposure.

Overall participants accurately recognised 89% of faces they had seen before if the image remained unchanged between exposure and test. Accuracy dropped to 76% if the image changed on one variable (expression or pose) between exposure and test. Accuracy further dropped to 60.5% if the image changed by two variables (expression and pose) between exposure and test. This experiment clearly illustrated the difficulties participants have in recognising unfamiliar faces across different images.

In a study from Bruce, Henderson, Greenwood, Hancock, Burton and Miller (1999) again two different images of faces were captured and this time used to test face identification in a mock identity line-up task. Images were taken from video stills and high quality photographs. The images of each face were captured on the same day to restrict variability in aspects such as hairstyle. Target images were presented above an array of ten high quality photographs (see Figure 1.4 for an example). The target image was either controlled to show the same forward facing, neutral expression as the array images or was changed to show a smiling expression, or a ¾-facing pose. The target was present in the array of ten images on 50% of the trials.

Participants were presented with the arrays and asked to indicate which image in the array matched the target image; participants were told that the target image would be present in approximately 50% of the arrays. Participants performance was poor, reaching 70% accuracy when the target image and array image matched for pose and expression, reducing to 66% when just the expression varied between images and 60% when just the pose varied between images. This finding illustrates that identification of unfamiliar faces across two images, with and without variations, is surprisingly poor.

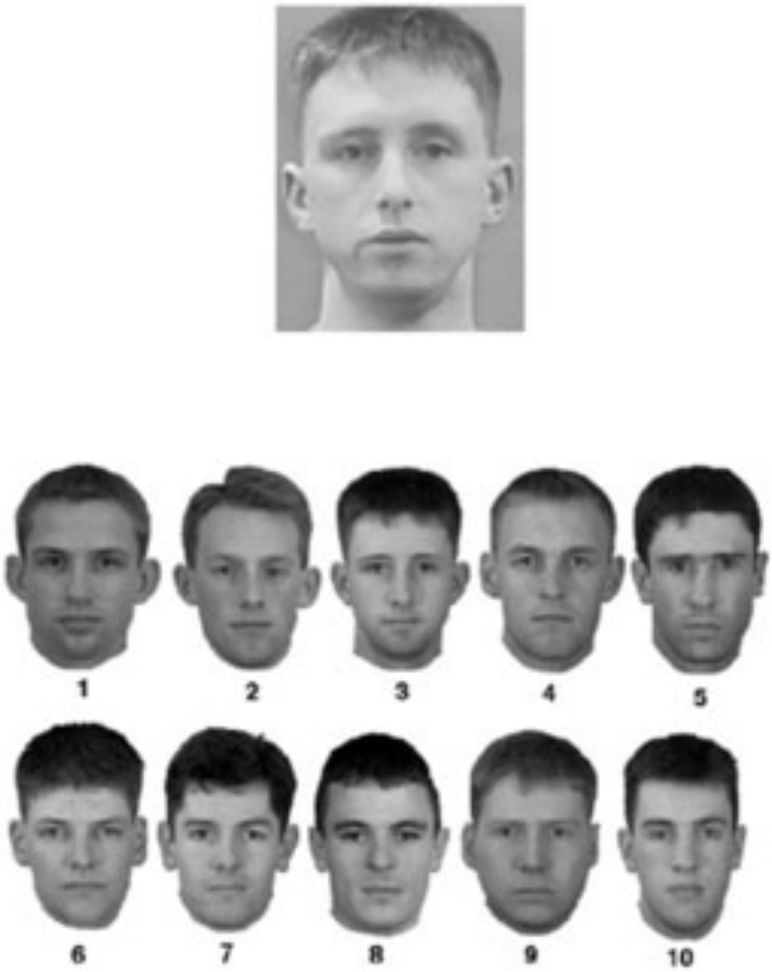


Figure 1.4. **Face matching in identification line-ups.**
An example of a target image and a line-up array of ten images presented to participants by Bruce et al., (1999).

In Bruce (1982) and Bruce et al., (1999) the idea of within-person variability through pose and expression was explored. However as described above, within-person variability can occur not just when face characteristics change but with variations in cameras used to capture images. The Glasgow Face Matching Test created by Burton, White and McNeil (2010) demonstrates the effect camera changes can have on face matching. Burton et al., (2010) created a database of 168 pairs of faces, 84 same-face pairs and 84 different-face pairs. The images of each face were of the same pose and expression, and were taken from high quality photographs and video stills (see Figure 1.5 for an example).

Participants were shown all pairs of face and asked to indicate if the pairs matched. Overall participants' accuracy ranged from 62 – 100% accuracy, with a mean of 89.9%. Considering the task given to participants, to match two images controlled for same pose and expression taken on two different cameras within minutes of each other, the results are surprisingly bad. This result highlights the difficulty of identifying the same face over different images and also demonstrates the effect that using different cameras can have on image variations.

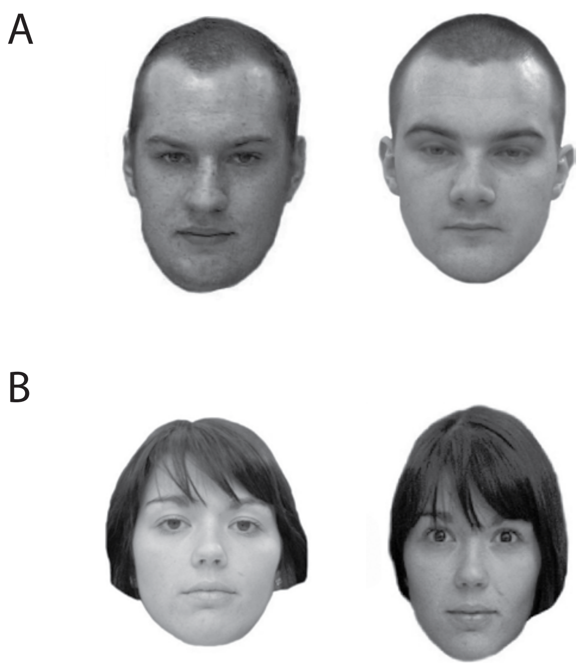


Figure 1.5. **The Glasgow Face Matching Test.**
An example of a different-face pair (A) and a same-face pair (B) taken from the Glasgow Face Matching Test (Burton, White & McNeil 2010).

Based on the findings from previous research outlined here it is clear that recognition memory and identification of unfamiliar faces across different, sometimes varied, images is a hard task for participants. However these studies demonstrate variability over a small number of images, often just two, and therefore do not illustrate the full extent of within-person variability. With this in mind it is surprising that these findings have not been developed to consider the full extent of within-person variability and the difficulties it causes for face recognition and identification.

1.4 Variability in photographs

One study that has explored within-person variability as outlined in this chapter, is a collection of experiments from Jenkins, White, Van Montfort and Burton (2011). Jenkins et al., explored within-person variability in image identification, image likeness and in facial attractiveness perceptions. In the identification tasks an image-sorting design task was used: participants were asked to sort by identity a set of forty images comprising of twenty different images of two different faces. Participants were not made aware of the number of identities present in the images. The results found that participants familiar with the faces had near perfect performance, while participants unfamiliar with the faces were surprisingly bad, creating 7.5 different identities. These findings show that identification across within-person variability is possible but is especially difficult if you are unfamiliar with a face.

From this Jenkins et al., (2011) suggested that perhaps some images are better at capturing a person's facial appearance than others, and so are therefore better representations of the person. To test this participants were shown 12 different images of 40 different familiar famous faces at random and were asked to give a likeness rating on a 7-point scale. A low rating indicated the image was of poor likeness to the famous person and a high rating indicated the image was of good likeness to the famous person. Twelve images of Bill Clinton shown in Figure 1.6 demonstrate the range of within-person variability encountered by participants. The results showed that a substantial amount of within-person variability occurred in likeness ratings for images of the same face. This suggests that part of the problem of within-person variability in identification is that some images are not a good likeness, or a good representation of a person.

Figure 1.6. **Within-person variability of Bill Clinton.**
Twelve ambient images of Bill Clinton representative of the type of within-person variability participants experienced in the likeness task in Jenkins et al., (2011).

Lastly Jenkins et al., (2011) examined within-person variability in the perception of facial attractiveness. Participants were given 20 images of 20 different unfamiliar faces and were asked to give a yes/no judgment of attractiveness for each image. The results showed that instead of finding a clear distinction between-person of faces that were rated attractive or unattractive, perceptions of attractiveness varied widely within-person. This occurred to the extent that for any pair of faces it was possible to select an image that could reverse attractiveness perceptions i.e. any face could be perceived as more attractive than another by simply changing the images of the faces compared. This further highlights the extent of within-person variability by showing that social judgments are not stable across different images.

The findings from Jenkins et al., (2011) illustrate the range of within-person variability that has not been fully explored in face processing research. Another important conclusion to be made is that within-person variability does not solely influence face identification but can also influence face perception. Building on these findings, the experiments set out in this thesis explore within-person variability in face identification, personality and attractiveness perceptions from facial appearance and image memory.

1.5 Familiar and unfamiliar face processing

The above findings all illustrate the difficulties in identifying across different images of unfamiliar faces. Previous research has also shown familiar face recognition to be significantly better across different image variations and even with poor quality images (Bruce 1982; Bruce, Henderson, Newman & Burton 2001; Burton, Wilson, Cowan & Bruce 1999; Jenkins, White, van Monfort & Burton 2011). This suggests that when we are familiar with a face we are very good at seeing past within-person variability to identify the person successfully.

For example, in a second experiment from Bruce (1982) face recognition across changed and unchanged images was examined for unfamiliar and familiar faces. Participants were shown images of 48 faces and following a delay of 15 minutes, were shown either the same image of each face or an image that had two variations (expression and pose) and were asked to indicate if they remembered seeing the person in the image previously. Participants were familiar with 24 of the 48 faces. Participants' accuracy was high for familiar faces, recognising 95.8% of faces if the images shown remained unchanged between exposure and test. Accuracy was similarly high even with changed images of familiar faces, with participants recognising 94.5% of faces if the images were changed across expression and pose. Participants' accuracy was lower for unfamiliar faces, recognising 88.8% of faces if the images shown remained unchanged between exposure and test, and recognising 54.8% if the images were changed across expression and pose. This finding clearly illustrates the ease with which participants can recognise familiar faces across different, varied images while struggling to recognise unfamiliar faces in same images and across different, varied images.

In a study from Bruce et al., (2001) familiarity was also shown to benefit performance in an identification task. Participants were shown a 3 second video clip, 3 images taken from video stills or a single image of 12 familiar target faces. A different high quality image of either the same or a different face was then presented to participants and they were asked to identify if the high quality image was of the target face. Participants' identification accuracy was high, with an overall score of 92% correct for target-present and target-absent trials. This finding illustrates that participants are very good at recognising and identifying familiar faces across different images of varying image quality.

From previous research it is clear that familiar face recognition is very good while unfamiliar face recognition is poor. These findings have led to research demonstrating that we process familiar and unfamiliar faces in different ways and have different representations of familiar and unfamiliar faces (see Johnston & Edmonds 2009 for a review). For example Bruce and Young (1986) propose that different codes are accessed when we see a face; pictorial codes and structural codes. Pictorial codes are precise descriptions of a face's appearance. It has been theorised that unfamiliar face recognition relies heavily on pictorial codes. Therefore if an image of an unfamiliar face is presented we process the image through pictorial codes. If we are then shown a different image of the same face, we struggle to recognise it as the same face because the pictorial codes will have changed. This is shown in the previously outlined identification and recognition memory experiments (Bruce 1982; Bruce et al., 2001; Burton et al., 1999). Structural codes are more abstract descriptions of a face's features and configuration; familiar face recognition has been proposed to rely mainly on structural encoding (Bruce & Young 1986). This difference is again reflected in previous research that has shown participants to be able to identify and recognise familiar faces across variability.

As discussed familiarity has been shown to benefit identification across within-person variability (Jenkins et al., 2011). But how do we recognise familiar faces across within-person variability? Bruce and Young's functional model of face processing (see Figure 1.7) suggests that we have facial recognition units (FRUs) for faces familiar to us and we use these FRUs to successfully identify a familiar face. When we see a face we build an abstract mental representation for that face, which is refined every time we see the face. Each encounter improves the quality of the representation by underlining the consistencies and inconsistences of the face. When a familiar face is viewed the specific FRU for that person will respond and an identification can be made i.e. who is this person, what is their name. Therefore the FRU acts as a generalised visual representation of a face that is not dependent on a specific view of the face or pictorial codes (Bruce & Young 1986).

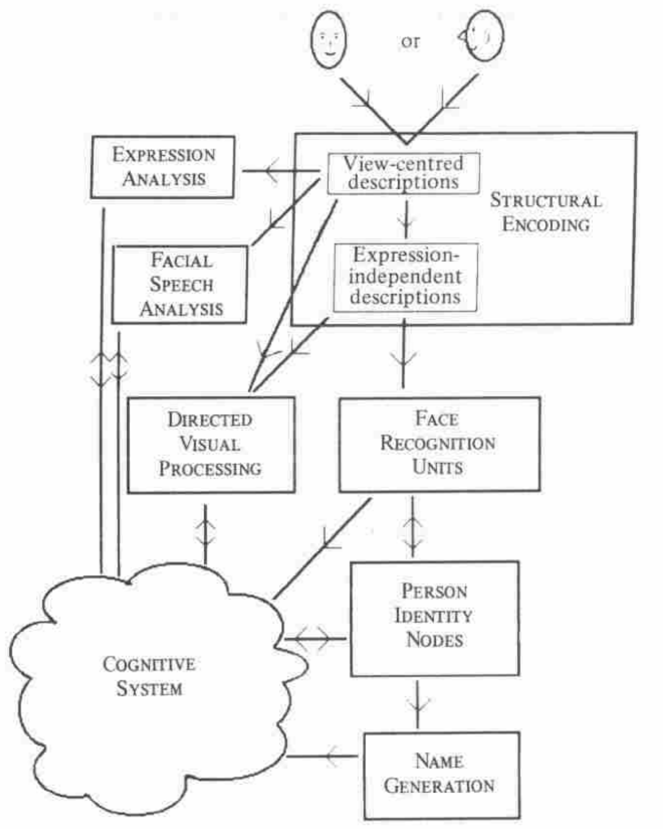


***Figure 1.7*. Bruce and Young's functional model of face processing (1986).**

Considering the extent of within-person variability, how long does it take to create FRUs that incorporate changes across within-person variability? Research has shown that gradual familiarisation of a face can occur through repeat exposure to different images of the face (Clutterbuck & Johnston 2004, 2005). For example, in an experiment from Clutterbuck and Johnston (2004) participants were shown images of novel faces over a period of two days. Participants were then required to complete a matching task with the same faces across images that had been varied across pose (frontal pose and ¾ angle pose). It was found that identification performance was better for the learned faces than for unfamiliar faces, but not better than performance with familiar faces. It was concluded that even a limited learning phase of previously unfamiliar faces was enough to observe a change in the way the faces were processed. This finding suggests that learning to recognise faces across variability is a gradual process.

Clutterbuck and Johnston's findings suggest that awareness of a face's variability might simply come from repeat exposure to the face (2004, 2005). In reference to Bruce and Young's model (1986), Burton, Jenkins and

Schweinberger (2011) proposed that at some stage during face learning there is a shift of focus from relying on pictorial codes to structural codes about a face. In the context of within-person variability this suggests that when we repeatedly see a face at some stage we start to look past variations in visual pictorial codes so that we can build a more structural representation of the face, therefore being able to recognise it across within-person variability.

1.6 Previous studies of face perception

So far within-person variability has been discussed in the research areas of face recognition and identification. In addition to face recognition and identification, face-processing research has explored the social perceptions that we read from faces. Previous research has shown that we can quickly perceive various different social signals from facial appearance, such as attractiveness (e.g. Olson & Marsheutz 2005), competence (e.g. Todorov, Mandisodza, Goren & Hall 2005), and trustworthiness and dominance (e.g. Oosterhof & Todorov 2008). Research has also demonstrated how facial perceptions can influence our judgments and decision-making (e.g. Mueller & Mazer 1996; Antonakis & Dalgas 2009). It is therefore important to know whether facial perceptions vary as much within-person as facial appearance can vary within-person.

Previous research in face perception typically focuses on variations between-persons and often uses artificial computer-generated faces that can be manipulated across different dimensions to explore different effects (Perrett et al., 1999; Little, Burt, Penton-Voak & Perrett 2001; Oosterhof & Todorov 2008). For example, Oosterhof and Todorov (2008) created a 2-dimensional ‘face space’ from which unlimited faces could be generated based on the dimensions trustworthiness and dominance (see Figure 1.8). Oosterhof and Todorov also demonstrated that important social judgments such as threat could be reproduced in faces as a result of this 2-dimensional face space.

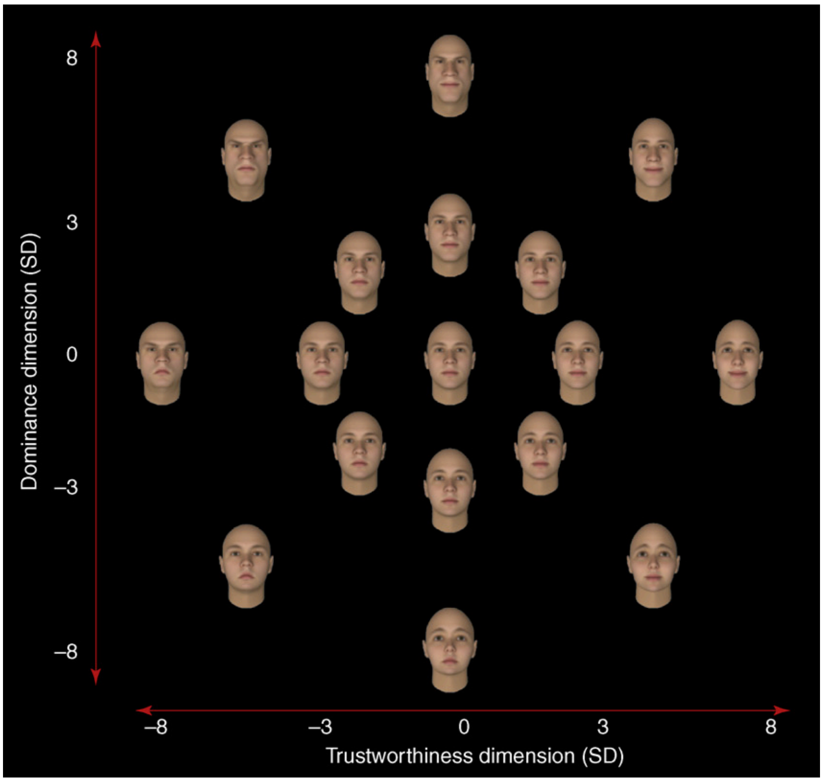


Figure 1.8. **Trustworthy and Dominance 2-dimensional face space.**
Example of computer-generated faces within the dimensions trustworthiness and dominance (Oosterhof & Todorov 2008).

The use of artificial faces is often a result of attempting to control for all other facial variability than what is being manipulated in the research question. Therefore a lot of information about real variability is overlooked. If variability occurs in facial appearance it seems logical to think that variability may also occur in face perceptions. This then raises the question of whether perceptions are stable and whether we can manipulate perceptions simply by manipulating images within-person.

1.7 Thesis aims and methodology

Following the review of previous research in face processing, it is the aim of this thesis to extend the findings of Jenkins et al., (2011) to examine whether the effect of within-person variability on face identification is robust. Developing on from within-person variability in face identification, a second aim of this thesis is to examine the effect of within-person variability in different areas of face processing, focusing on within-person variability in social perceptions and image memory. The final aim of this thesis is to test the effect of within-person variability on both familiar and unfamiliar face processing.

In Chapter 2 within-person variability in identification is examined through an image-sorting task as employed by Jenkins et al., (2011). An image-sorting task simply involves giving participants a ‘deck’ of images to identify. Any additional information given to participants about the images can then be manipulated to measure the effect on identification. This image-sorting task was developed and used in Chapter 3 with different variations made on the design. In Chapters 4, 5 and 6 participants’ perceptions of faces and image memory are measured through randomised computer presentations of different images of faces, with participants making keyboard-responses to different task questions. Throughout the experiments, there is an emphasis on collecting face stimuli that show realistic within-person variability. This is achieved through the use of faces which are easy to access variable images for, for example famous faces (from different countries), images of which can be accessed through Internet image searches. Throughout the experiments, there is also an emphasis on recruiting participants new to the concept of within-person variability.

1.8 Overview of thesis

Chapters 2: Within-person variability in face identification, and Chapter 3: Between-person similarity in face identification

In Experiment 1 within-person variability in identification is explored by expanding the image-sorting task used in Jenkins et al., (2011) to establish whether within-person variability occurs with images of different faces. In Experiments 2 and 3 the effect of within-person variability on identification is examined when additional information is given about the identities present in the images. In Experiments 4 and 5 similarities of the faces in the images is examined along with the effect of familiarity on identification. The issue of similarity is then controlled by conducting an image-sorting task with 20 images of just one face in Experiment 6. Finally in Experiment 7 the image sorting design is applied to images selected from a database designed to capture within-person variability. This allows for the ambient images used in Experiments 1-6 (see Figure 1.1 for an example of one face) to be compared to images controlled for different dimensions such as pose and expression.

Chapter 4: Within-Person Variability in Social Dimensions

In Experiment 8 the effect of within-person variability on the perception of social dimensions is examined. Participants rate images of 80 different faces that encompass within-person variability for two personality traits: trustworthiness and dominance. The effect of familiarity on within-person variability is tested with participants rating both familiar and unfamiliar faces. As in Experiments 1-6, the images of each face are ambient images as oppose to images controlled for different dimensions.

Chapter 5: Within-Person Variability in Facial Attractiveness

In Experiments 9 and 10 the effect of within-person variability on the perception of facial attractiveness is examined. In Experiment 9 participants rate multiple ambient images of 20 different familiar and unfamiliar faces for attractiveness. In Experiment 10 a selection of the images rated in Experiment 9 are presented in ascending order of attractiveness or descending order of attractiveness to test whether order effects would influence the overall attractiveness perception of familiar and unfamiliar faces.

Chapter 6: Within-Person Variability in Image Memory

In Experiments 11 and 12, within-person variability is applied to memory by exploring image memory for multiple images of the same face. The same images of familiar and unfamiliar faces as used in Experiment 8 are used in Experiments 11 and 12. This enabled participants in Experiment 8 to be recruited for Experiment 11, exploring their image memory over a long period of delay. The average length of delay in Experiment 11 was sixteen weeks. In Experiment 12 memory for the same images is tested but the length of delay between participants' initial exposure to the images and their memory test is reduced to just 24 hours. Using images of both familiar and unfamiliar faces allows for the effect of familiarity on image memory to be examined.

Chapter Two

Within-Person Variability in Face Identification

2.1 Introduction

There has been a substantial amount of face recognition and identification research that has concluded unfamiliar face recognition is highly error prone (Henderson, Bruce & Burton 2001; Davies and Valentine 2009; Bruce, Henderson, Greenwood, Hancock, Burton & Miller 1999; Kemp, Towell & Pike 1997) while familiar faces recognition is surprisingly good (Bruce, Henderson, Newman & Burton 2001; Burton, Wilson, Cowan & Bruce 1999). For example in a study from Burton et al., (1999) familiar face recognition was shown to be better than unfamiliar face recognition in a face 'matching' task. Participants were shown poor quality video clips of familiar and unfamiliar faces and were asked to identify if the same faces were present in high quality images. Results found that participants familiar with the target faces performed significantly better than participants unfamiliar with the target faces. This finding clearly illustrates the high level of familiar recognition ability even in poor-quality images, and the poor ability of unfamiliar face recognition.

Similarly Bruce (1982) found familiar face recognition to be better than unfamiliar face recognition. Participants were shown images of familiar and unfamiliar faces, and following a short delay were shown either the same images or images of the same faces varied by expression and pose. Participants were asked to indicate whether they had seen each person before. Performance accuracy for correctly identifying a familiar face was consistently high at over 90% while performance accuracy for unfamiliar faces dropped to as low as 54% when the image had been changed by either viewpoint or expression. This finding clearly illustrates that familiar face recognition across variations such as expression and pose is a lot better than unfamiliar face recognition across similar variations.

In a study examining unfamiliar identification, Kemp, Towell and Pike (1997) found that participants struggled to identify whether a photographic identification card matched the 'live' target presenting the card. In a real life setting of a supermarket, trained employees acted as participants and were required to assess whether a photographic identification card (a credit card including a face image) matched the cardholder. The image on the card showed either the real target person looking similar to their current appearance, the real target with a change to their appearance (such as a change in hairstyle), a different person matched for similarity, or a different person dissimilar to the target matched only for race and gender. The results showed that making accurate identifications was surprisingly difficult, with participants accepting just over 50% of the fraudulent cards i.e. the cards that displayed an image of a different person than the cardholder. This finding clearly illustrates the difficulties of unfamiliar face identification across multiple image presentations and also demonstrates the applied implications of unsuccessful face identification.

Previous research has shown participants struggle to accurately identify unfamiliar faces, which raises the question, is it possible to improve unfamiliar face recognition? Research has begun to explore how performance with unfamiliar faces can be improved, for example exploring the process of face familiarisation. Bruce et al., (2001) introduced a familiarisation phase to a series of face matching experiments. The first familiarisation phase consisted of viewing a 30 second or 60 second animated video of the target faces prior to the matching task. This was found not to have a significant effect on identification performance. In a subsequent experiment, the familiarisation phase was developed to involve viewing a 28 second long animated video of the target with another participant while discussing the faces shown. Results showed that this level of familiarisation had a significantly positive effect on identification performance. Therefore it was concluded that the introduction of familiarisation had little improvement on identification performance unless participants' in-depth and/or social processing was encouraged. This could suggest that unfamiliar face processing is difficult because we don't process unfamiliar faces on a deeper level, attributing social information to the faces in the same way that we do with familiar faces.

In a series of experiments Clutterbuck and Johnston (2004; 2005) also found that familiarising participants with previously unfamiliar faces resulted in improved identity matching performance. For example in their 2004 study, participants completed two familiarisation phases prior to an image-matching task. The familiarisation phases took place over two consecutive days with the matching task following the second familiarisation phase. During both familiarisation phases faces were shown at random for a total of 2 seconds repeated 15 times.

Participants were asked to give each face an honesty rating to ensure that participants were attending to each face. Following the second familiarisation phase, participants were presented with 72 pairs of images made up of 36 same pairs and 36 different pairs. Performance for the learned faces was significantly better than for the unfamiliar faces. This finding suggests that repeated exposure to faces and being required to attribute personality perceptions to the same faces, benefits identification performance. This supports findings from Bruce et al., (2001) and suggests that unfamiliar faces are harder to process because we have not been repeatedly exposed to the face in the same way as we have been with familiar faces, and we have not processed unfamiliar faces in the same way as familiar faces i.e. attributing more in depth information about the person to the face.

Burton, Jenkins, Hancock and White (2005) proposed that classifying a more robust representation of a face could improve face processing. Burton et al., found that participants were faster to correctly identify an image of a famous person that had been averaged from multiple images, than a non-average image of the same person. The process of averaging a face involves incorporating the variability that can occur in a face into one image. Therefore to improve face identification the variability that occurs within one person's facial appearance had to be overcome by averaging out the variability. Burton et al., (2005) illustrated that by averaging out across multiple images of one face, face identification of famous people was improved. To determine whether within-person variability is causing poor recognition of unfamiliar faces, face processing across multiple images of unfamiliar faces needs to be tested.

In a study from Jenkins, White, van Monfort and Burton (2011) within-person variability in face identification of both familiar and unfamiliar faces was examined. Jenkins et al., (2011) suggest that previous research has a constricted idea of how much one person's facial appearance can vary. To examine realistic within-person variability ambient images must be sampled, i.e. images taken from the surrounding environment rather than from an experimental pool. Participants were given twenty ambient images of two different faces and were asked to sort the images by identity, i.e. group together the images that represented one person. Participants who were *unfamiliar* with the two faces made a mean number of 7.5 identities from the 40 images. In comparison participants who were *familiar* with the two faces made a mean number of 2 identities with almost all participants reaching the correct outcome. These results illustrate that identification across within-person variability is surprisingly difficult, especially for faces that are unfamiliar to us. The near perfect results from the participants familiar with the faces suggests that part of being familiar with a face is being able to acknowledge the variability that can occur in a person's facial appearance and being able to recognise the face despite that variability. Comparing these findings to the familiarisation results of Bruce et al., (2001), and Clutterbuck and Johnston (2004), it could be suggested that this awareness of within-person variability in familiar faces is attained through repeated exposure to the person.

In this chapter the findings of Jenkins et al., (2011) are expanded to examine whether the effect of within-person variability occurs with images of different faces. With ambient images (as defined by Jenkins et al., 2011) of two new faces, participants' identification performance was observed in Experiment 1 using the same 'free sorting' image design outlined by Jenkins et al., (2011) in which participants were asked to sort the images by identity without being made aware of how many identities to expect. It was then possible to manipulate the experimental design in Experiment 2 and 3 to establish in what conditions accurate identification can be achieved.

2.2 Experiment 1: Free sorting

In Experiment 1 participants' identification behaviour was explored with a stimuli set comprising of multiple images of two different faces. A 'free sorting' image task was implemented with participants sorting the images by identity with no further information about the images or identities.

Method

Participants

Forty participants recruited from the School of Psychology's subject pool (22 females: 18 males; age range: 18 – 55 years) took part in this experiment in return for a small payment.

Stimuli and Design

To test participants' identification ability with multiple images of different faces, first images of different faces had to be collected. Twenty images of 2 different male faces, that captured within-person variability as discussed in Chapter 1, were gathered. The images were not taken specifically for the experiment therefore are not controlled for variations such as lighting or camera specifications, or limited to capturing appearance over a short time period. Rather the images were ambient images that captured any variability in the appearance of each face, providing an accurate representation of that person. The images were predominantly personal photographs that were supplied by the two men. The faces will be referred to as Face 1 and Face 2.

The images were cropped to include only the face and rescaled to replicate the size of an official passport photograph (45mm x 35mm). The images were then printed in black and white and laminated to create a set of 40 images (see Figure 2.1). It is important to underline that although the images were kept the constant size and grey scale, the images were not controlled any further.

Participants were presented with the images in a 'free sorting' design task. All participants were asked to sort the same stimuli. Familiarity with either face was established following the task by asking participants to indicate with a "Yes/No" response whether they recognised or were familiar with either face. Some participants were familiar with Face 2 (a current lecturer at the University of Glasgow) therefore familiarity is later examined.

Procedure

Participants were given the set of 40 images and asked to: "sort the image by identity so that photos of the same face, or person, are grouped together". Participants were then informed that there was no time restriction although the trial time was recorded, and that they were free to create as many or as few groups as they saw fit.

Participants' familiarity with Face 2 was recorded, along with the number of identities participants reached and the grouping of the identities i.e. the specific images that made up the identities. The main measure of interest from participants' performance is the number of identities participants reached; two identities is the actual number of identities present in the stimuli therefore it is of interest to see how participants perform in comparison.

Figure 2.1. **Images used in Experiment 1.**
Twenty images of two different faces made up the stimuli for Experiment 1: Twenty images of Face 1 are shown in the top two rows (labelled 1 – 20) and twenty images of Face 2 are shown in the bottom two rows (labelled 21 – 40).

Results

Surprisingly none of the 40 participants arrived at the correct number of 2 identities present amongst the 40 images. The mean number of identities participants' reached was 9.7 identities (median 8; mode 5; range 3-24). A one-sample t-test confirmed that the mean number of identities produced was significantly higher than the two identities actually present in the image set, [t (39) = 8.93, $p < 0.01$]. Therefore participants' mean outcome is significantly higher than the correct outcome.

Identity-split errors and identity-merge errors

None of the participants reached the correct number of identities present in the images therefore all participants made identification errors. These errors can be classified as identity-*split* errors and identity-*merge* errors. An identity-split error occurs within-person; it is an error consisting of images of the same person being divided into more than one identity. An identity-merge error occurs between-person; it is an error consisting of one or more images of one face being grouped with images of another face resulting in an identity that merges two different identities.

All 40 participants made identity-split errors while 13 out of 40 participants made identity-merge errors (see Figure 2.2). Of the 13 participants who made identity-merge errors, a mean number of 1.2 errors were made. This suggests that the difficulty participants experienced was not telling images of the two identities apart but telling images of each identity *together*.

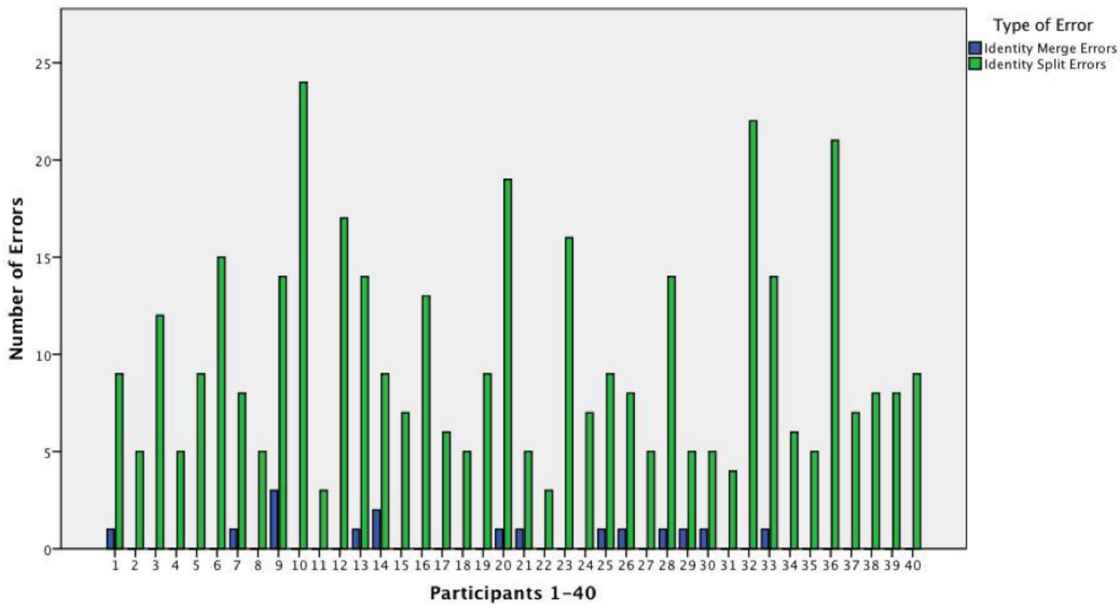


Figure 2.2. **Experiment 1: Histogram of participants' identification errors.**
Histogram illustrating the number of identity-split errors and the number of identity-merge errors each participant made in Experiment 1.

Face 1 and 2 considered independently

The majority of participants only made identity-split errors. To explore whether more split errors were made for one face over the other, participants' identification performance was analysed separately for Face 1 and Face 2. If one face was more difficult to identify than the other, this may have caused the overall poor performance.

The mean number of identities produced for *Face 1* was 5.9 identities (median 4.5; mode 2; range 1-16). A one-sample t-test of the identities made for Face 1 shows that the mean number of identities produced is significantly higher than the correct outcome of one identity [t (39) = 7.75, $p < 0.01$]. The mean number of identities produced for *Face 2* was 4.25 identities (median 3; mode 3; range 1-9). A one-sample t-test for identities made for Face 2 show that the mean number of identities produced is significantly higher than the correct outcome of one identity [t (39) = 9.63, $p < 0.01$]. Therefore looking at the number of identities made from images of Face 1 and Face 2 it appears that the faces are similarly difficult to identify (see Figure 2.3).

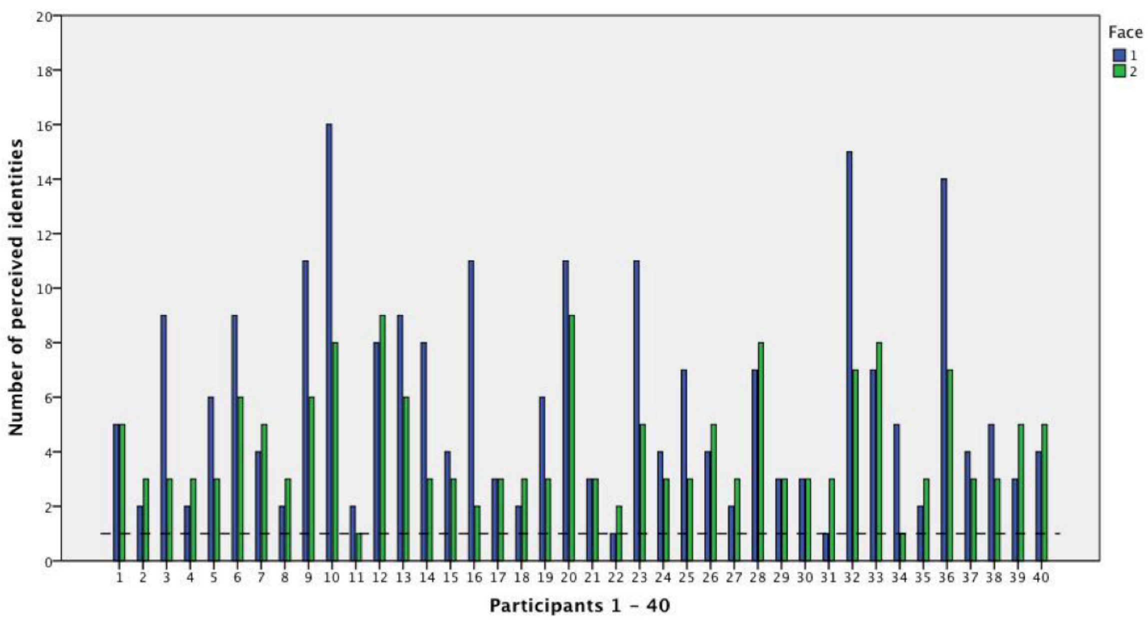


Figure 2.3. **Experiment 1: Number of perceived identities.**
Histogram illustrating participants' perceived identities for Face 1 and Face 2, with the correct number of identities for each face (i.e. one) indicated with a dashed line.

Effects of Familiarity

As it turned out, 9 of the 40 participants were familiar with Face 2 prior to Experiment 1 because Face 2 was a current lecturer at the University of Glasgow. This allowed for any effects of familiarity to be explored.

The results were considered separately for participants familiar and unfamiliar with Face 2. It is important to note that the number of identities reached for Face 1 and Face 2 were both considered independently for familiarity. This was to investigate whether being familiar with one of the faces affected participants' identification performance of the remaining face. From this point participants described as 'familiar participants' are participants familiar with Face 2 and participants described as 'unfamiliar participants' are participants unfamiliar with Face 2.

Firstly the overall performance (i.e. the number of identities reached for Face 1 and Face 2) from participants familiar and unfamiliar with Face 2 was compared using a Mann Whitney-U test. The mean number of identities made by familiar participants was 7 identities (median 6; mode 3 and 5; range 3-13). The mean number of identities made by unfamiliar participants was 10.6 identities (median 9; mode 5; range 4-24) (see Figure 2.4). It was found that this difference was not quite significant (U = 85, $p > 0.05$), suggesting that familiarity with Face 2 does not result in significantly better overall performance with both Face 1 and Face 2.

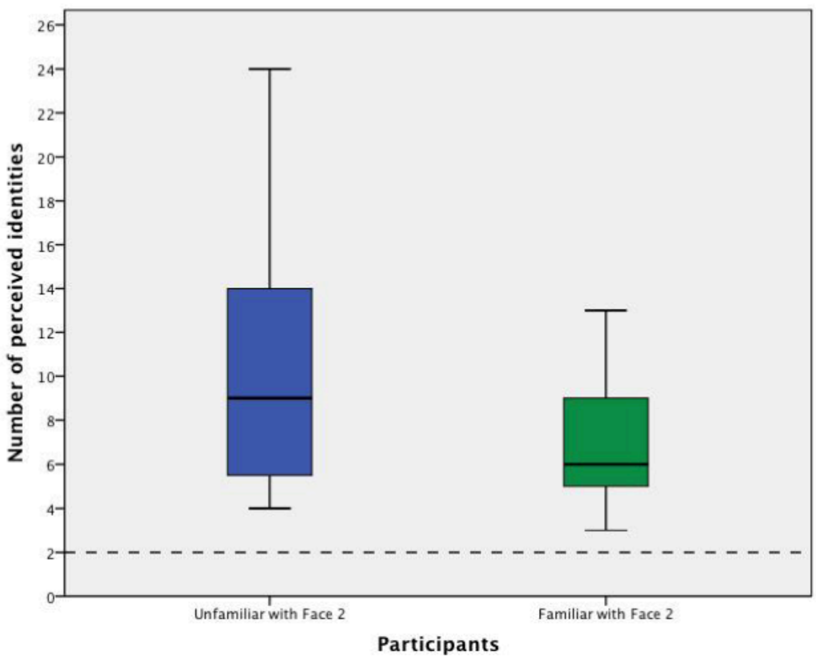


Figure 2.4. **Experiment 1: Performance by familiarity.**
A boxplot illustrating the overall range of identities made (for both faces) by participants familiar and unfamiliar with Face 2, with actual number of identities represented with a dotted line.

Face 1

All participants were unfamiliar with Face 1. Participants familiar with Face 2 made a mean number of 4.7 identities for images of Face 1. Participants unfamiliar with Face 2, i.e. the remaining participants, made a mean number of 6.1 identities for images of Face 1. These results were analysed using a Mann-

Whitney U test and no significant difference was found between the numbers of identities made for Face 1 ($U = 108.50$, $p > 0.05$). This suggests that familiarity with Face 2 does not benefit identification performance for Face 1.

Face 2

Nine of the 40 participants were familiar with Face 2. The 31 remaining participants were unfamiliar with Face 2. Participants familiar with Face 2 made a mean number of 2.3 identities for images of Face 2. Participants unfamiliar with Face 2 made a mean number of 4.8 identities for images of Face 2. These results were analysed using a Mann-Whitney U and the mean number of identities formed by familiar participants was significantly different from the mean number of identities formed by unfamiliar participants, ($U = 37.50$, $p < 0.01$). This suggests that familiarity with Face 2 has a positive effect on identification performance, but only for Face 2.

Effects of time

Participants' performance in Experiment 1 was surprisingly poor. A possible reason for poor performance could be participants rushing through the task, resulting in a negative effect on their accuracy. To investigate this, trial time i.e. the amount of time participants spent to complete the identification task, and the number of identities produced by participants were analysed. It was expected that the more time spent on the task would lead to a more accurate result.

As it turned out there was a significant positive correlation between trial time and the number of identities produced ($rs = .36$, $p < 0.05$) (see Figure 2.5). That is, the longer participants spent on the task the more identities they perceived. This rules out any possible speed-accuracy trade-off as an explanation of the observed performance. Not only was participants' performance not due to rushing the task but participants' performance actually got worse the longer they spent on the task. This suggests participants' first impressions when identifying an image are more accurate than if participants over-think identification.

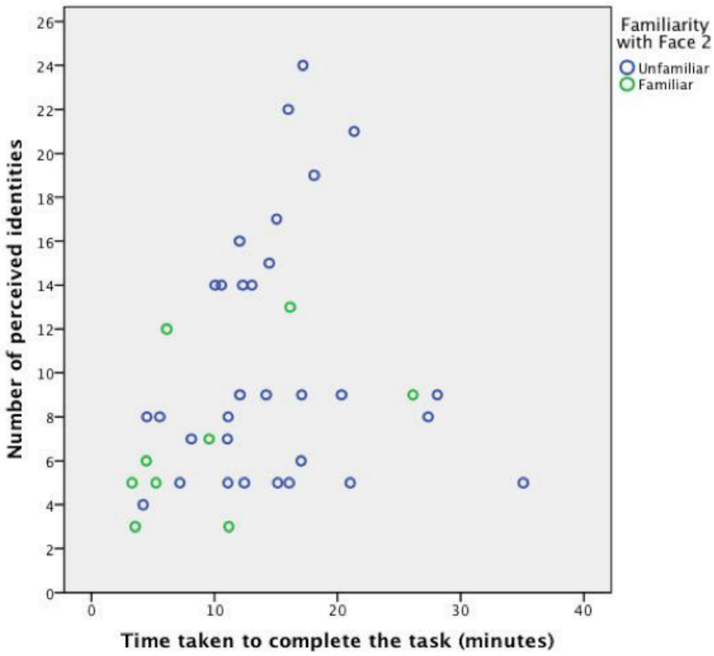


Figure 2.5. **Experiment 1: Effects of Time.**
A scatterplot illustrating the number of perceived identities (y-axis) and the time taken by participants to complete Experiment 1 (x-axis). Blue data points represent participants unfamiliar with Face 2 and green data points represent participants familiar with Face 2.

Confusion Matrix

To examine the identification errors made by participants more closely a confusion matrix was created (see Figure 2.6). Images of Face 1 (listed 1-20) and images of Face 2 (listed 21-40) were listed down either side of the matrix. Every occurrence of one image being categorised with another image was independently recorded to build up a visual representative of the identities created by participants. If participants had performed perfectly the matrix would show all red cells and the number 40 (representative of 40 participants) in the top left and bottom right quadrants, and all green cells in the top right and bottom left quadrants. As participants did not reach perfect performance instead their identity-split errors and identity-merge errors can be visualised. Identity-split errors are represented in the top left and bottom right quadrants; green cells indicate when images were infrequently grouped together therefore resulting in split errors, while red cells indicate when images were frequently, correctly grouped together. Identity-merge errors are represented in the top right and bottom left quadrants; all green cells indicate that images were rarely, or never grouped together therefore resulting in few, or no merge errors.

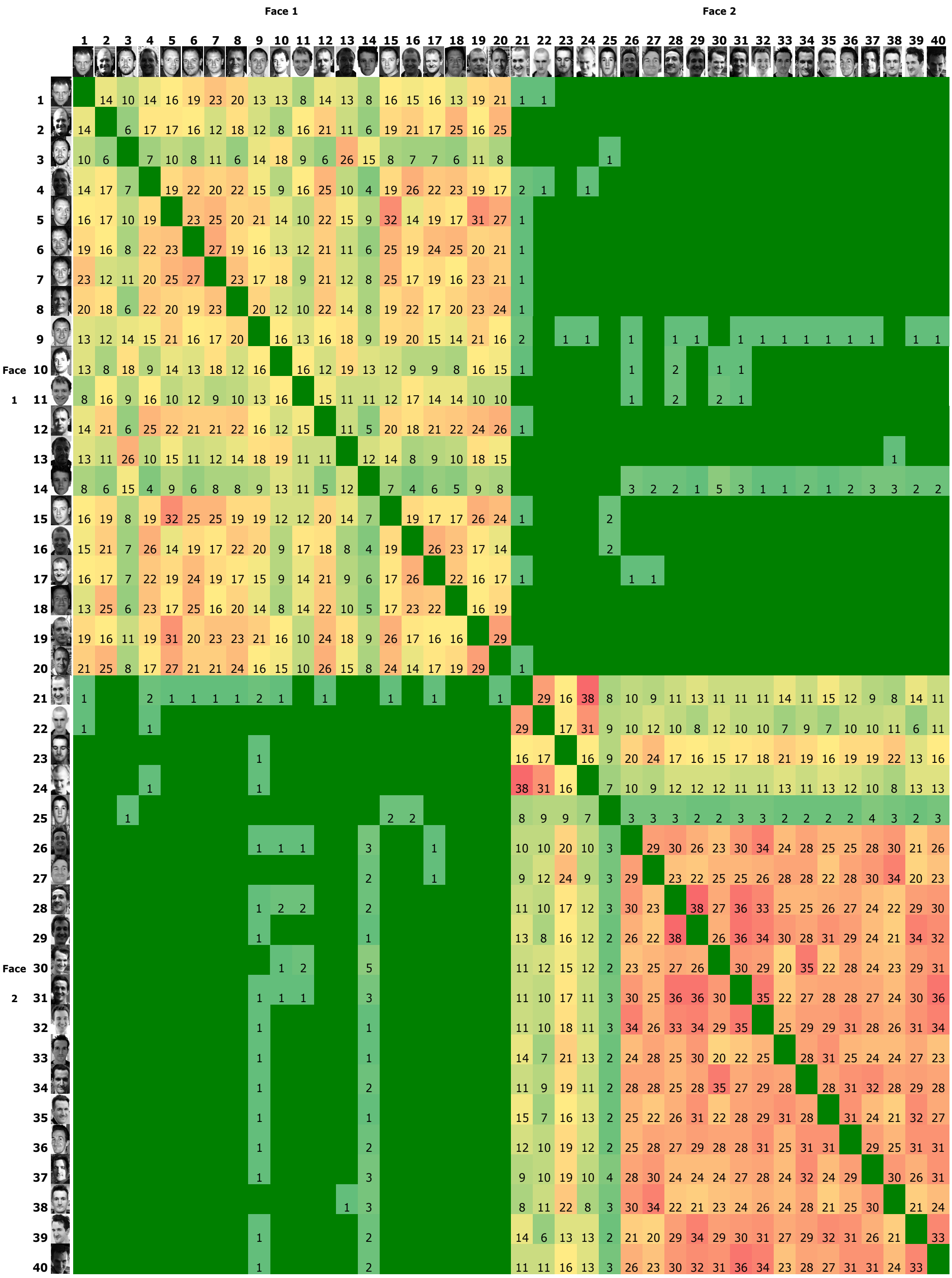


Figure 2.6. **Experiment 1: Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.

From the confusion matrix it is possible to visualise participants' identification behaviour: shown in Figure 2.7 is an example of images frequently identified correctly, a frequent identity-split error and a frequent identity-merge error for each face. Most notably it is clear from the matrix that in comparison to identity-split errors, very few identity-merge errors were made, represented by the green cells. Within the identity-split errors it is possible to observe patterns that indicate images that frequently caused difficulties for participants. For example it is clear to see from the range of green cells that images 3 and 14 of Face 1 and image 25 of Face 2 were infrequently categorised with other images within each identity. Finally it is worth noting that very few red cells occur in the matrix, illustrating the difficulty of identifying images of within-person variability together.

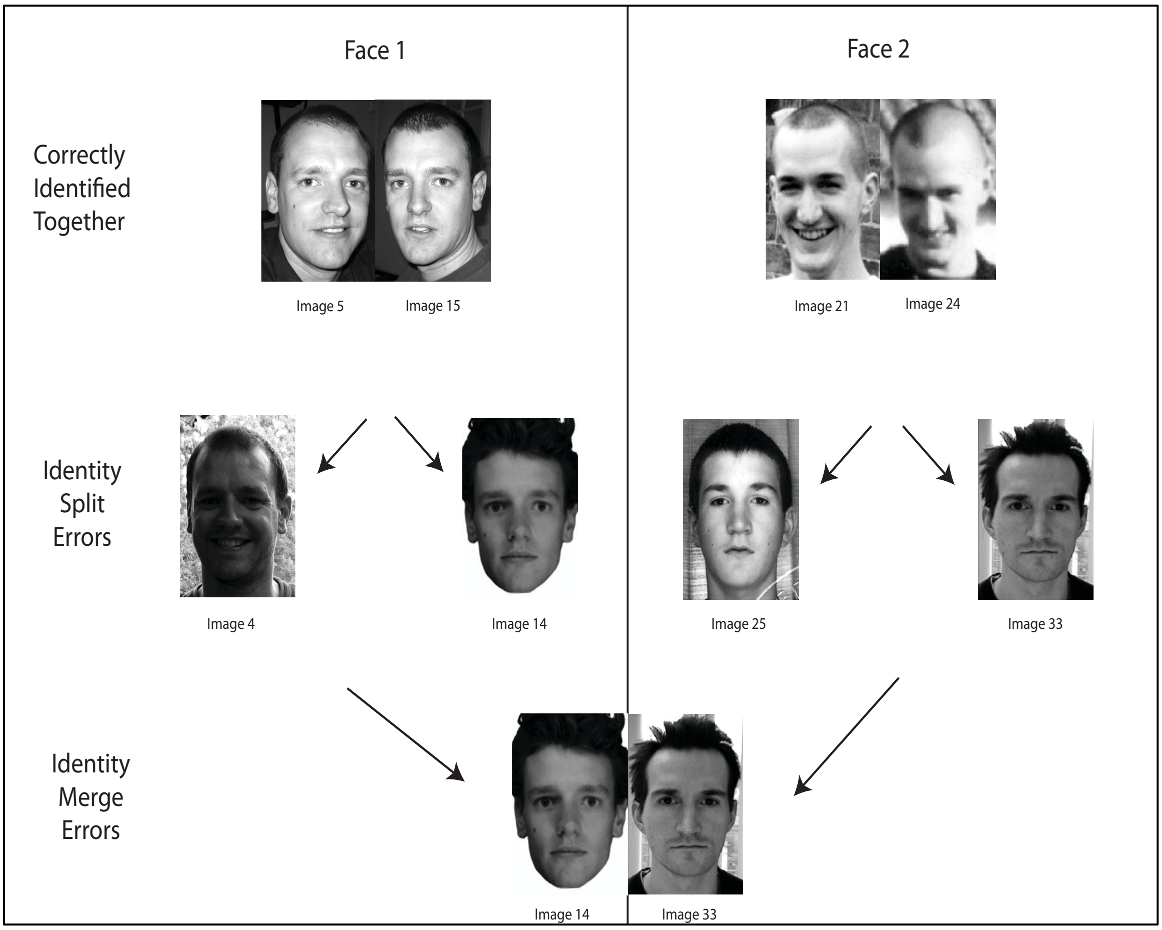


Figure 2.7. **Experiment 1: Examples of identification errors.**
Examples are shown of frequently correctly identified images in the top row, frequent identity-split errors in the second row and an example of an identity-merge error in the bottom row.

Discussion

Participants performed surprisingly poorly in Experiment 1 when asked to identify multiple images of two different faces. Instead of reaching the correct outcome of 2 identities amongst the stimuli, participants made a mean number of 9.7 different identities. This clearly demonstrates the difficulties in identifying across within-person variability. The findings replicate and extend the findings of Jenkins et al., (2011), illustrating that within-person variability occurs with images of different faces.

Analysis of the identification errors shows that a high number of identity-split errors were made. Identity-split errors occur when images of the same face are identified as two different identities. Participants made a small number of identity-merge errors, errors resulting from identifying images of different faces as the same identity. This suggests that the main difficulty of identification was grouping different images of the same face together as opposed to grouping images of different faces apart.

The time spent on identification was analysed to explore whether participants' poor identification performance was based on rushing through the task. However trial time and participants' performance were significantly positively correlated, suggesting not just that rushing the task did not cause poor performance but that the longer participants spent on the task, the worse their performance was.

It was discovered after the task that 9 of the 40 participants were familiar with Face 2 in the stimuli. This was due to Face 2 being a current lecturer at the University of Glasgow. Familiarity could therefore be explored to a certain extent. Analysis showed that participants familiar with Face 2 were significantly better at identifying images of Face 2. This finding is in line with previous research that has shown face processing to be very good when a face is familiar (e.g. Bruce 1982; Burton et al., 1999). Being familiar with one of the faces did not improve identification for Face 1. This could have been expected as once participants are exposed to the variability that can occur within one face they may have been more aware of variability within the other faces. However the findings instead suggest that awareness of within-person variability is learnt on a face-by-face basis.

It could be suggested that the high number of identities made by participants is a reflection of the task given to them, asking participants to sort “by identity” may have implied that multiple identities exist within the images. This is an important consideration, however examining the identification behaviour of participants shows that participants typically grouped images together based on similar cues such as changes in expression, and similar aged images. This could suggest that participants were being led by visual information in the images, as opposed to verbal information given in the task instructions. Furthermore, participants were informed that the number of identities made could be as few or as many as they saw fit. To improve the instructions in future research and minimise task effect, it may be helpful to ask participants following the task whether they had felt compelled to reach their outcome based on the instructions or whether they were led by the images alone. The feedback could then be applied to adjust the instructions where appropriate.

In Experiment 1 participants were not aware of how many identities were present amongst the stimuli. Successfully identifying the two faces turned out to be a difficult task for participants. If participants are made aware of the number of identities present, would this information help identification performance? Furthermore if participants are given a visual aid in the form of a reference image for each identity, would this information help identification? Establishing the conditions in which participants can accurately identify two faces across within-person variability is examined in Experiment 2.

2.3 Experiment 2: Image sorting with two constraints

In Experiment 1 participants struggled to identify the two different faces present in the stimuli. To establish if and when participants would be able to correctly identify the two faces, constraints were introduced to Experiment 2. Constraints are pieces of information that will prevent participants from behaving in a certain way. Here, two constraints are introduced: a descriptive constraint and a visual constraint. The descriptive constraint consists of the information that two identities exist amongst the stimuli. The visual constraint consists of a reference image of each identity present in the stimuli. The current passport photographs for each face were used here as the reference image. If constraints can help performance then it can be established that in certain conditions, more accurate identification across within-person variability is achievable.

Method

Participants

Forty participants recruited from the School of Psychology's subject pool (31 females: 9 males; age range: 18 – 38 years) took part in Experiment 2 in return for a small payment.

Design and Stimuli

The same stimuli was used as in Experiment 1, 40 different images of Face 1 and Face 2 (see Figure 2.1). In addition stimuli was prepared to act as the visual constraint for participants. These consisted of the current passport photograph for Face 1 and Face 2 (see Figure 2.8). The images were again printed and laminated but were sized slightly larger (65mm x 55mm) so that they could be distinguished from the rest of the images.

All participants were given the same stimuli, the visual constraint of the reference images and the descriptive constraint that two identities are present in the images. Participants were asked to complete the "constrained sort" design task, sorting the images by identity.

***Figure 2.8.* Reference images used in Experiment 2.**
The visual constraint in Experiment 2 consisted of two reference images: the current passport photographs of Face 1 (left) and Face 2 (right).

Procedure

Participants were given the set of 40 images and told that two identities were present amongst the images. They were then presented with two additional reference images and told that these images represented each of the two identities present in the forty images. Participants were asked to: “sort the 40 images into the two identities using the information given in the constraints” i.e. with the knowledge that only two identities were present and using the reference images as a guide to match to.

Participants were then informed that there was no time restriction although the trial time was recorded, and that they were free to create the identities as they saw fit. Participants’ performance was recorded, along with the specific images that were grouped with each reference image and the time it took for them to reach their final outcome.

Familiarity with either face was established following the task by asking participants to indicate with a “Yes/No” response whether they recognised or were familiar with either face. Some participants were familiar with Face 2 (a current lecturer at the University of Glasgow) therefore familiarity was later examined.

Results

Unlike Experiment 1, participants in Experiment 2 knew that only two identities exist among the 40 images. Therefore no identity-split errors can be made: if an identity is ‘split’ then it becomes part of the other identity and therefore is an identity-merge error.

Despite constraints being introduced to the task, the research question remains the same: How well will participants perform at identifying two identities from 40 images? However with the knowledge of the two constraints - the descriptive constraint that two identities exist amongst the stimuli set and the visual constraint of a reference image for each identity - the result to focus on is not the number of identities reached, which will remain constant at two identities, but the number of identity-merge errors made.

Identity-merge errors

The mean number of identity-merge errors made by participants was 3.27 errors (median 1; mode 0; range 0-23). Fourteen out of 40 participants did not make identity-merge errors therefore 14 out of 40 participants reached the correct outcome (see Figure 2.9).

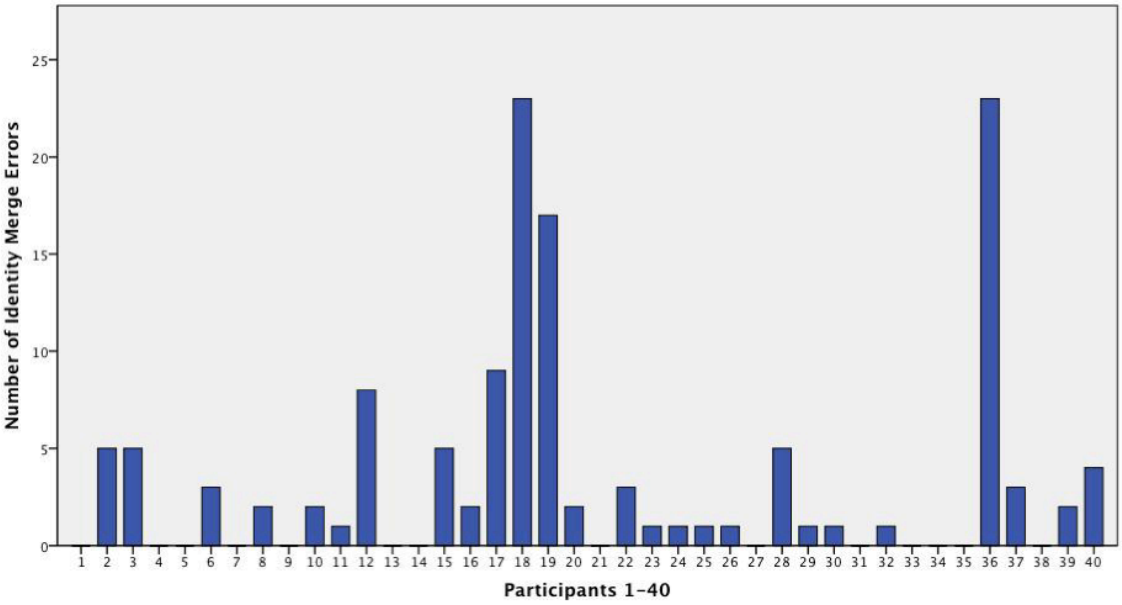


Figure 2.9. **Experiment 2: Histogram of participants' identification errors.**
A histogram illustrating the number of identity merge errors each participant made in Experiment 2.

Face 1 and Face 2 considered independently

To examine whether one face was harder to identify than the other, participants' errors were considered separately for each face. To consider errors independently for Face 1 and Face 2, what is meant by a Face 1 error and a Face 2 error must first be defined. A Face 1 identity-merge error occurs when an image of Face 1 is identified with the reference image of Face 2. A Face 2 identity-merge error occurs when an image of Face 2 is identified with the reference image of Face 1.

Identity-merge errors are more frequent for Face 2 than Face 1 (see Figure 2.10). Participants made a mean number of 1.25 identity-merge errors for Face 1, and a mean number of 2 identity-merge errors for Face 2. This difference is statistically significant [t (39) = -2.67, p < 0.05]. This suggests that the images of Face 2 were too varied to justify grouping the images together and that some images varied more widely from the reference image of Face 2, in comparison to Face 1.

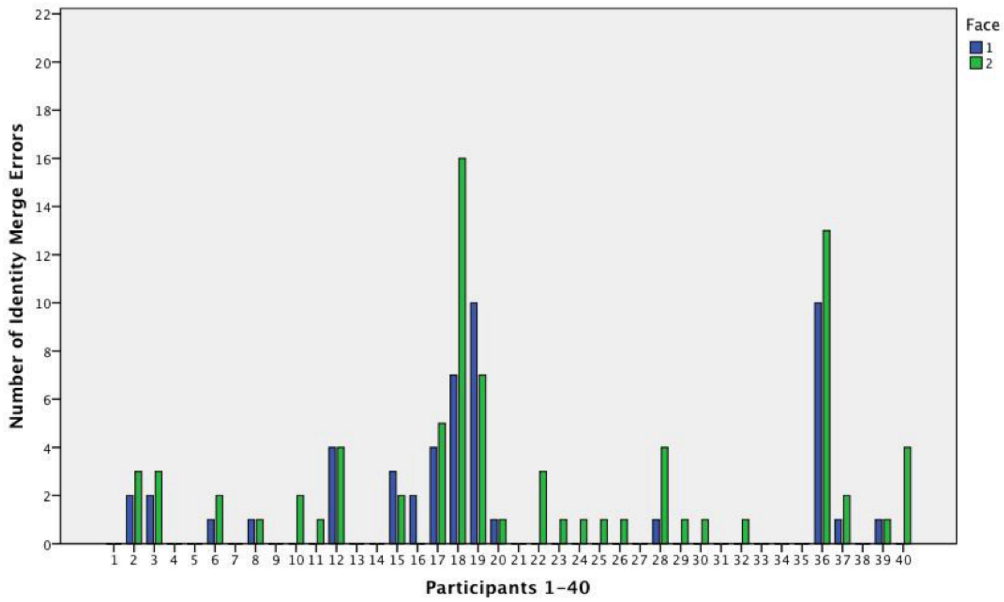


Figure 2.10. **Experiment 2: identity-merge errors for each face.**
A histogram illustrating the number of identity-merge errors, shown independently for Face 1 and Face 2, made by participants in Experiment 2.

Familiarity

Following the experiment it was established that 11 of the 40 participants were familiar with Face 2 (a current lecturer at the University of Glasgow). Therefore the results were considered separately for participants familiar and unfamiliar with Face 2 to investigate the possible effect of familiarity. It is important to note that the number of identity-merge errors for Face 1 and Face 2 were both considered independently, to investigate whether familiarity with Face 2 also affected the performance to identify Face 1, as opposed to performance solely for Face 2 being investigated for a familiarity effect. From this point on, participants familiar with Face 2 will be referred to as familiar participants, and participants unfamiliar with Face 2 will be referred to as unfamiliar participants.

Firstly the overall mean numbers of identity-merge errors made by familiar and unfamiliar participants were considered (see Figure 2.11). Unfamiliar participants made a mean number of 4.31 errors (median 2; mode 0 and 1, range 0 - 23) and familiar participants made a mean number of 0.55 errors (median 0; mode 0; range 0-4). Analysis shows that the mean number of identity-merge errors made by participants unfamiliar with Face 2 is significantly higher than the mean number of identity-merge errors made by participants familiar with Face 2 [U = 62, $p < 0.01$]. Therefore there is a familiarity effect: participants familiar with Face 2 make significantly fewer identity-merge errors.

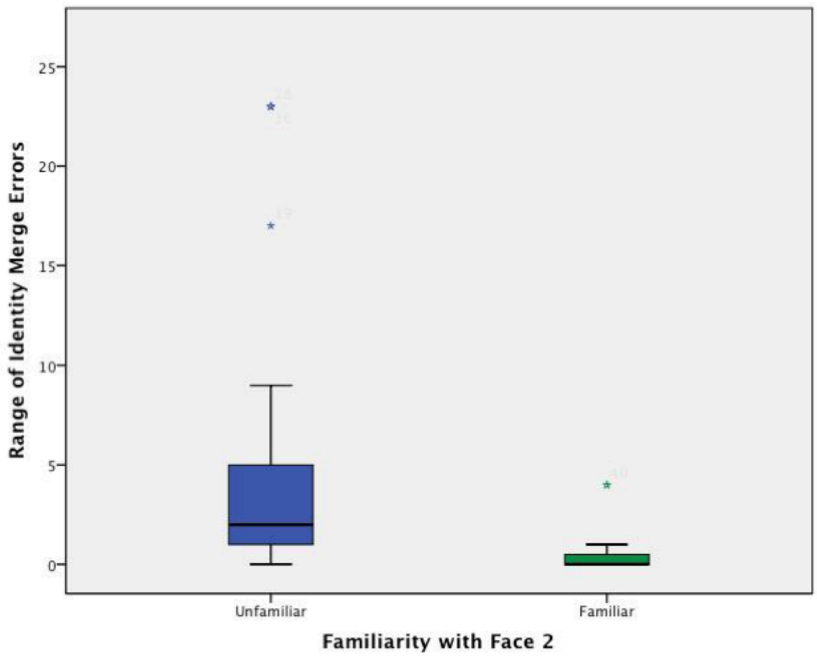


Figure 2.11. **Experiment 2: Performance by familiarity.**
A boxplot illustrating the range of identity-merge errors made by participants familiar and unfamiliar with Face 2.

Face 1

Familiar participants made no identity-merge errors for Face 1, while unfamiliar participants made a mean number of 1.72 identity-merge errors for Face 1 (median 1; mode 0; range 0-10). Performance for Face 1 was found to be significantly different: familiar participants were significantly better at identifying images of Face 1 ($U = 77$, $p < 0.01$).

Face 2

Familiar participants made a mean number of 0.55 identity-merge errors for Face 2 (median 0; mode 0; range 0-4) while unfamiliar participants made a mean number of 2.59 identity-merge errors for Face 2 (median 1; mode 9; range 0-16). This difference was found to be significant: familiar participants were significantly better at identifying images of Face 2 ($U = 75$, $p < 0.01$).

Trial time

To discover whether participants' errors in performance were the result of rushing through the task, the time it took for participants to complete the task and the number of identity-merge errors made were analysed (see Figure 2.12). It was found that trial time and performance were not significantly correlated ($rs = 0.192$, $p > 0.05$). Therefore participants' performance was not related to the amount of time they spent on the image-sorting task.

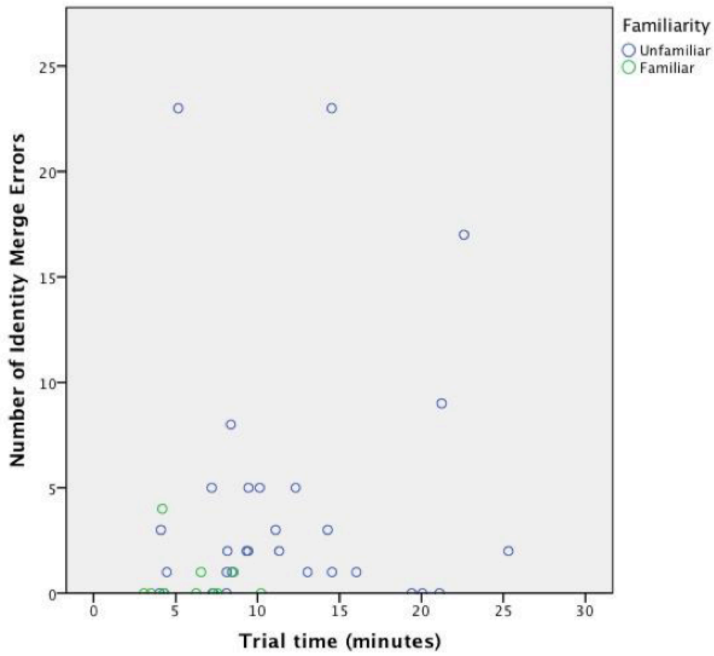


Figure 2.12. **Experiment 2: Effects of Time.**
A scatterplot illustrating the number of identity-merge errors made by participants (y-axis) and the time taken by participants to complete Experiment 2 (x-axis). Blue data points represent participants unfamiliar with Face 2 and green data points represent participants familiar with Face 2.

Confusion Matrix

To more closely examine the identity-merge errors made, a confusion matrix was created with images of Face 1 (listed 1-20) and images of Face 2 (listed 21-40), (see Figure 2.13). Every occurrence of one image being categorised with another image was independently recorded to create a visual representative of participants' performance. If participants had performed perfectly the matrix would show all red cells and the number 40 (representative of 40 participants) in the top left and bottom right quadrants, and all green cells in the top right and bottom left quadrants. As participants did not reach perfect performance the matrix clearly shows where images were grouped together and where images were categorised with the wrong reference image, making identity-merge errors.

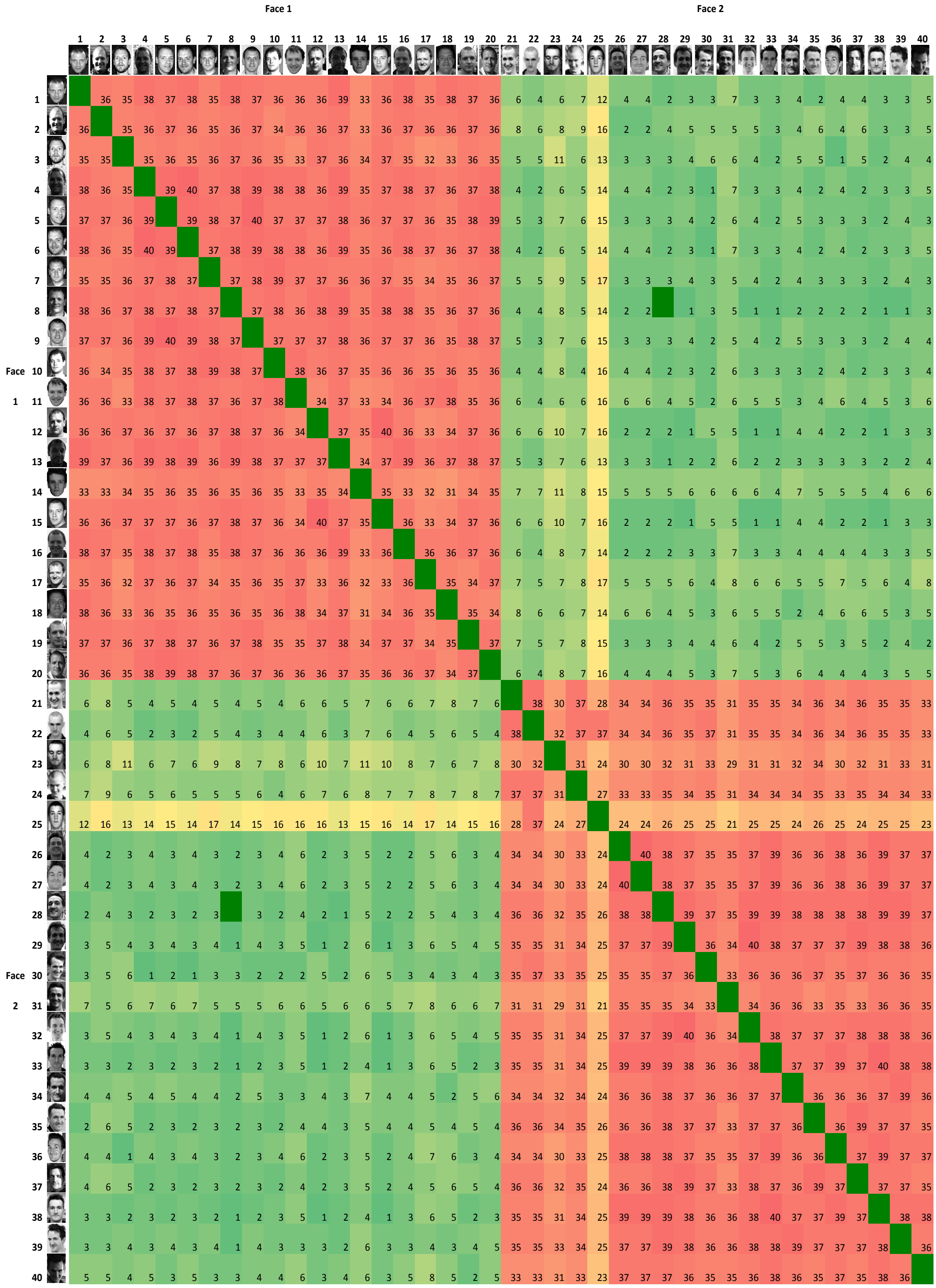


***Figure 2.13.* Experiment 2: Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.

From the confusion matrix participants' identification behaviour can be observed (see Figure 2.14 for examples). For example, it is possible to see that *all* images were at some point wrongly categorised with images from the other face. Only two images were never identified together: image 8 of Face 1 and image 28 of Face 2. All other combinations of images were at some point mixed resulting in identity-merge errors.

What the matrix doesn't show is how participants grouped the images with the two reference images. Analysis of participants' performance shows that *all* images of Face 2 were at some point 'matched' to the reference image of Face 1, while *all* images of Face 1 *except* image 4 were at some point 'matched' to the reference image of Face 2. This is a very surprising result. It illustrates the extent to which images were identified to the wrong reference image: *all* images were wrongly identified apart from image 4 of Face 1.

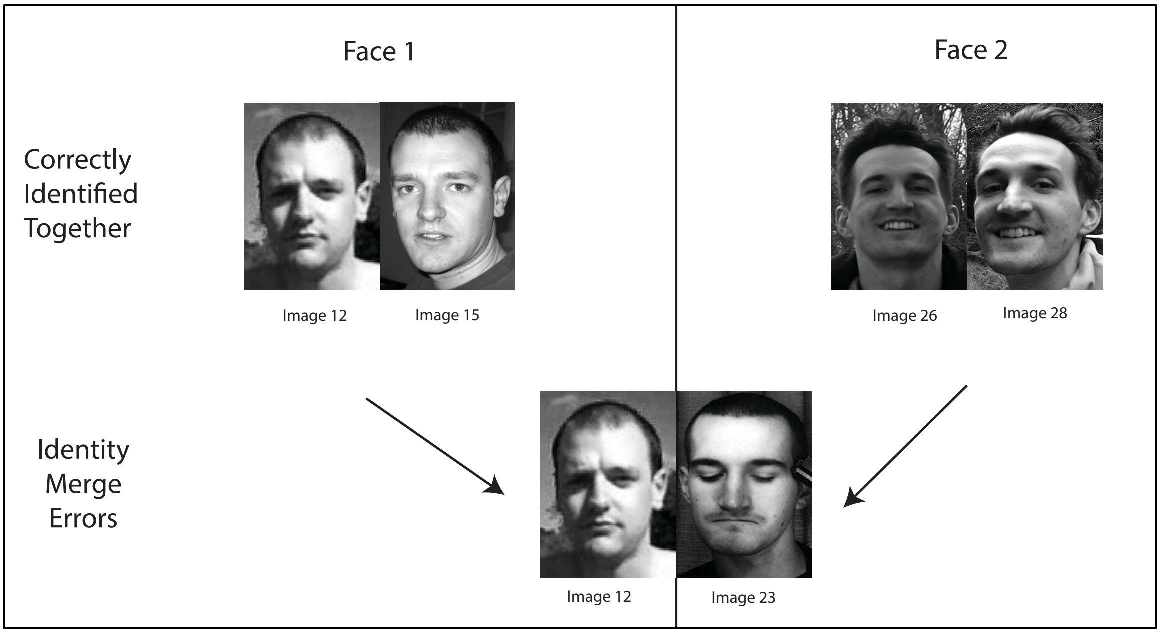


Figure 2.14. **Experiment 2: Examples of identification errors.**
Examples are shown of frequently correctly identified images in the top row and a frequent identity-merge error in the bottom row (participants were unable to make identity-split errors).

Discussion

In this experiment participants were given additional information about the identities present in the images and were able to reach the correct number of two identities. However only 14 of the 40 participants correctly identified all 40 images with the right reference image. The remaining participants made a range of 1 – 23 errors. Participants were unable to make identity-split errors as whenever they 'split' images of the same identity they created an identity-merge error. Therefore all errors made in Experiment 2 were identity-merge errors.

The frequency of identity-merge errors increased from Experiment 1: *all* images were at some point wrongly identified with an image of another face. This clearly illustrates the difficulty of identification across within-person variability, and suggests that one of the constraints introduced in Experiment 2 created confusion for participants and resulted in an increase of identity-merge errors.

Interestingly 39 out of the 40 images were *all* at some point identified with the wrong reference image. This is a surprising result and clearly illustrates the difficulties that can occur when matching to a target image. As the reference images were current passport photographs for each identity, this finding has a clear applied implication.

The results from Experiment 2 clearly show an increase of identity-merge errors in comparison to Experiment 1. In Experiment 2 participants were given reference images of each identity, therefore the task developed to involve identifying the images of the two different identities together and identifying the images to the right reference image. Without the reference images the participants would have to decide for themselves what the two identities looked like and group the images into the two possible identities. To explore whether participants are able to do this correctly, in Experiment 3 a new set of participants are presented with the same images and this time given just one descriptive constraint that two identities exist amongst the images.

2.4 Experiment 3: Image sorting with one constraint

In Experiment 2 participants complete an image-sorting task with the same images as used in Experiment 1, and were given additional information about the two identities present. This additional information was given in the form of a descriptive constraint: two identities exist amongst the images, and a visual constraint: a reference image for each identity. Participants were able to reach the correct number of identities but the number of identity-merge errors increased from Experiment 1. To establish whether participants are able to reach the correct outcome with less information, Experiment 3 reduces the constraints to just one descriptive constraint that two identities are present amongst the images. This also allows for the opportunity to explore whether too much information can actually cause confusion for participants, i.e. more identity merge errors.

Method

Participants

Forty participants recruited from the School of Psychology's subject pool (25 females: 15 males; age range: 18 – 54 years) took part in Experiment 3 in return for a small payment.

Stimuli and Design

The same stimuli was used as in Experiment 1, 40 different images of Face 1 and Face 2 (see Figure 2.1). All participants were given the same 40 images and in addition were given the descriptive constraint that only two identities were present in the images, resulting in a one constraint sorting task design.

Procedure

Participants were given the set of 40 images and told that two identities were present amongst the images. Participants were asked to: "sort the 40 images into the two identities using the information given" i.e. with the knowledge that only two identities were present amongst the images.

Participants were then informed that there was no time restriction although the trial time was recorded, and that they were free to create the identities as they saw fit. Participants' performance was recorded, along with the specific images that were grouped as each of the two identities and the time it took for them to reach their final outcome.

Familiarity with either face was established following the task by asking participants to indicate with a "Yes/No" response whether they recognised or were familiar with either face. A small number of participants were familiar with Face 2 (a current lecturer at the University of Glasgow) therefore familiarity was examined.

Results

As in Experiment 2, participants in Experiment 3 were aware that only two identities existed among the 40 images. Therefore no identity-split errors can be made; if an identity is ‘split’ then it becomes part of the other identity and is an identity-merge error. Despite a descriptive constraint being provided in this task, the research question remains the same: How well will participants perform at identifying two identities from 40 images? Again as in Experiment 2, the result to focus on is not the number of identities reached, which will remain constant at 2 identities, but the number of identity-merge errors participants made.

Unlike Experiment 2 no reference images were given, therefore an identity-merge error occurs when an image of one face is wrongly identified with the *majority* of images of the other face. For example, a Face 2 identity-merge error occurs when an image of Face 2 is wrongly identified with the *majority* of images of Face 1. It was clear to see (from all participants performance) which identity was the ‘majority’ identity for Face 1 and Face 2.

Identity-merge errors

The mean number of identity-merge errors made by participants was 2.15 errors (median 2; mode 0; range 0-12). Out of 40 participants only 9 participants reached the correct outcome (see Figure 2.15). This illustrates that although participants were able to reach the correct outcome of two identities, with the additional descriptive constraint, the majority of participants struggled to categorise *all* twenty images of each identity together.

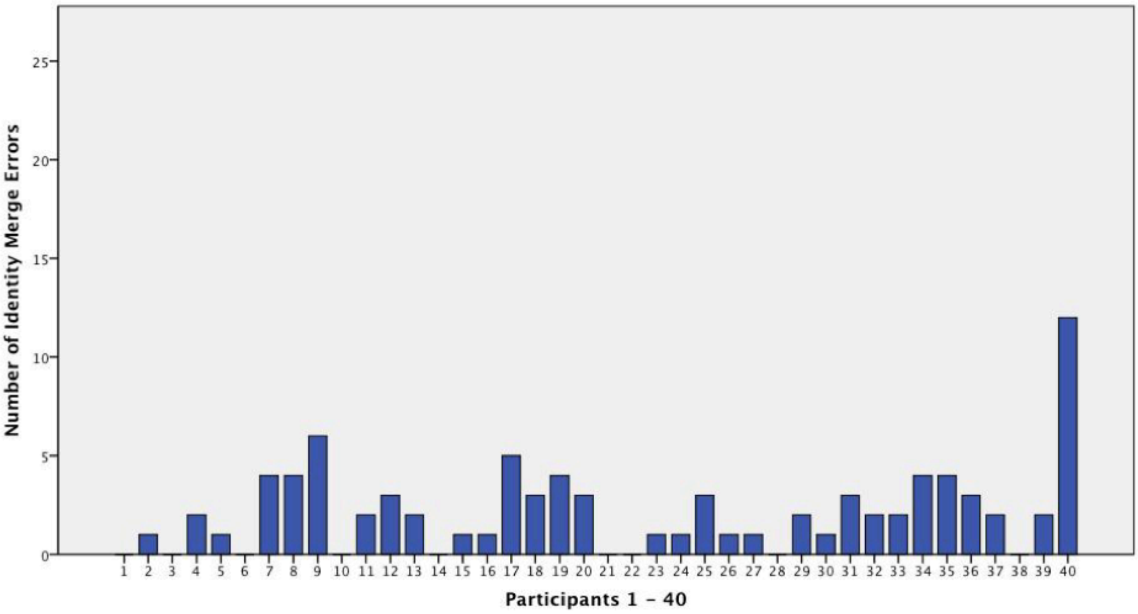


Figure 2.15. **Experiment 3: Histogram of participants' identity-merge errors.**
A histogram illustrating the number of identity-merge errors each participant made in Experiment 3.

Face 1 and Face 2 considered independently

Participants made a mean number of 0.97 identity-merge errors for images of Face 1 (median 1; mode 0; range 0-8) and a mean number of 1.17 identity-merge errors for images of Face 2 (median 1; mode 0; range 0-4) (see Figure 2.16). It was found that there is no significant difference between identity-merge errors made for Face 1 and Face 2 [t (39) = -0.941, $p > 0.05$]. This suggests that images of Face 1 and images of Face 2 were equally difficult to identify in this task.

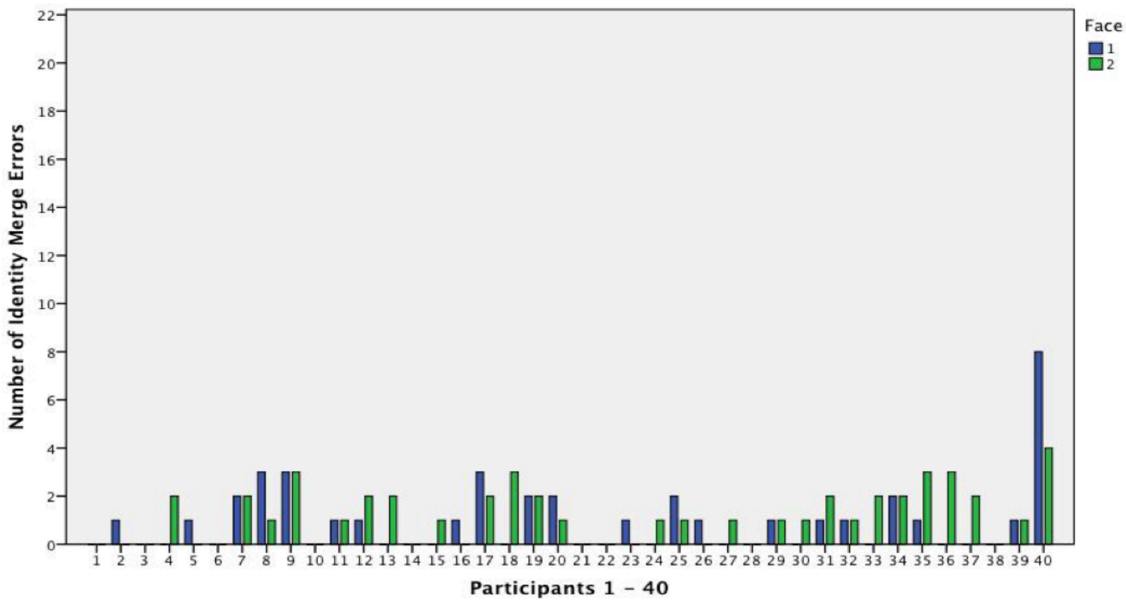


Figure 2.16. **Experiment 3: Histogram of identity-merge errors for each face.**
A histogram illustrating the number of identity-merge errors, shown independently for Face 1 and Face 2, made by participants in Experiment 3.

Effects of Familiarity

It was established that 5 of the 40 participants were familiar with Face 2 (a current lecturer at the University of Glasgow). Therefore the results were considered separately for participants familiar and unfamiliar with Face 2 to investigate the possible effect of familiarity. It is important to note that the number of identity-merge errors for Face 1 and Face 2 were both considered independently, to investigate whether familiarity with Face 2 also affected the performance to identify Face 1, as opposed to performance solely for Face 2 being investigated for a familiarity effect. From this point on, participants familiar with Face 2 will be referred to as familiar participants, and participants unfamiliar with Face 2 will be referred to as unfamiliar participants.

Firstly the mean numbers of identity-merge errors made by familiar participants and unfamiliar participants were considered (see Figure 2.17). Unfamiliar participants made a mean number of 2.34 identity-merge errors, (median 2; mode 0 and 1, range 0-12) and familiar participants made a mean number of 0.8 identity-merge errors (median 0; mode 0; range 0-3). Analysis shows that the difference between the mean number of identity-merge errors made by familiar participants familiar and unfamiliar participants is not quite at significance level [$U = 43.5$, $p > 0.05$]. Therefore unlike the findings in Experiment 1 and 2, in Experiment 3 there does not appear to be an effect of familiarity. Given the small number of familiar participants (5 out of 40), perhaps significance could be reached if the task was repeated with more familiar participants.

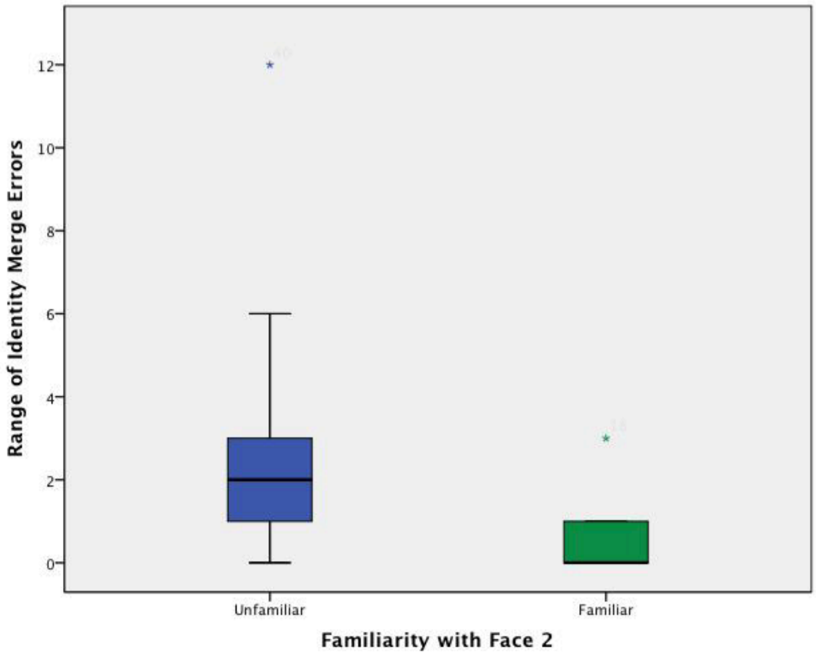


Figure 2.17. **Experiment 3: Performance by familiarity.**
A boxplot illustrating the range of identity-merge errors made by participants familiar and unfamiliar with Face 2.

Face 1

Familiar participants made no identity-merge errors for Face 1 while unfamiliar participants made a mean number of 1.11 identity-merge errors for Face 1 (median 1; mode 0; range 0-4). Performance for Face 1 was found to be significantly different: familiar participants made significantly less identity-merge errors for Face 1 ($U = 35$, $p < 0.05$).

Face 2

Familiar participants made a mean number of 0.8 identity-merge errors for Face 2 (median 0; mode 0; range 0-3) while unfamiliar participants with Face 2 made a mean number of 1.23 identity-merge errors for Face 2 (median 1; mode 0; range 0-8). No significant difference was found between performance of familiar and unfamiliar participants ($U = 65$, $p > 0.05$).

Error rates considered independently for Face 1 and Face 2 are very low. It can be concluded that familiar participants perform better with images of Face 1 than unfamiliar participants. However familiar participants perform no differently with images of Face 2 than unfamiliar participants. This suggests that being familiar with Face 2 allowed participants to rule out categorising any images of Face 1 as Face 2 but did not exclude them from confusing images of Face 2 as Face 1. It is important to note however that any identity-merge errors with Face 2 were very low suggesting that participants familiar with Face 2 may have struggled to correctly identify only a small number of images of Face 2.

Effects of Time

To discover whether participants' errors in performance were the result of rushing through the task, the time it took for participants to complete the task and the number of identity-merge errors made were analysed (see Figure 2.18). It was found that the time participants took to complete the task and their performance was not significantly correlated (*rs* = 0.36, $p > 0.05$). Therefore participants' performance was not related to the amount of time they spent on the image-sorting task.

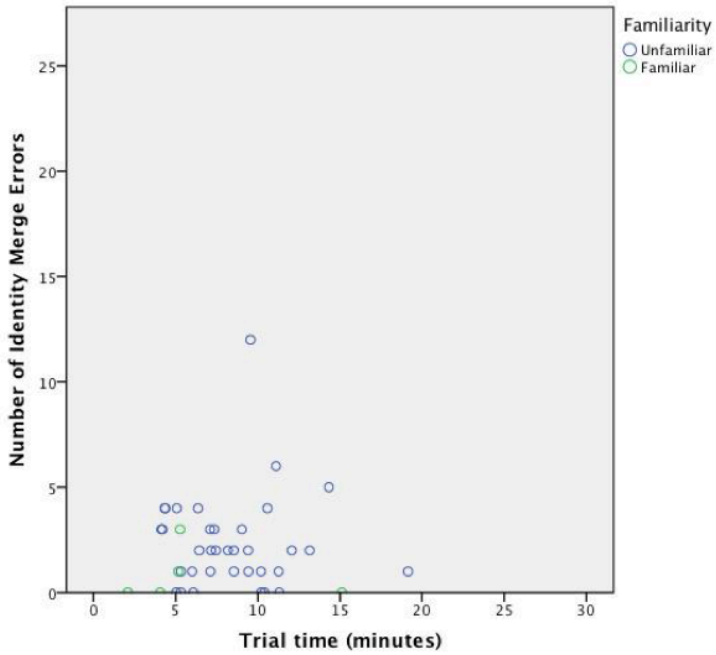


Figure 2.18. **Experiment 3: Effects of Time.**
A scatterplot illustrating the number of identity-merge errors made by participants (y-axis) and the time taken by participants to complete Experiment 3 (x-axis). Blue data points represent participants unfamiliar with Face 2 and green data points represent participants familiar with Face 2.

Confusion Matrix

To examine participants' identification behaviour in Experiment 3, a confusion matrix was created with images of Face 1 (listed 1-20) and images of Face 2 (listed 21-40), (see Figure 2.19). Every occurrence of one image being categorised with another image was independently recorded to create a visual representative of participants' performance. If participants had performed perfectly the matrix would show all red cells and the number 40 (i.e. 40 participants) in the top left and bottom right quadrants, and all green cells in the top right and bottom left quadrants. As participants did not reach perfect performance the matrix illustrates where errors were made.

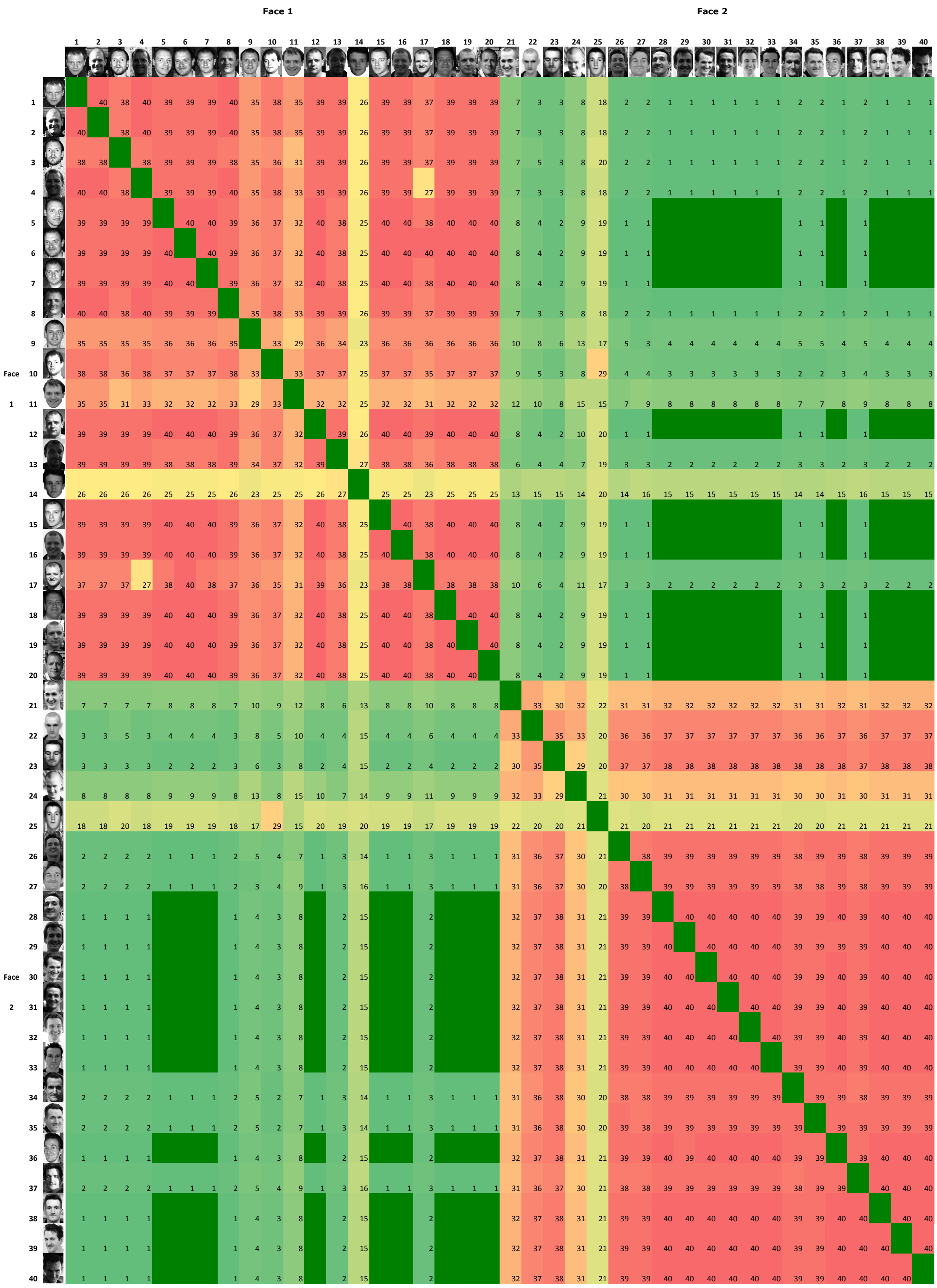


***Figure 2.19.* Experiment 3: Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.

From the confusion matrix participants' identification behaviour can be observed (see Figure 2.20 for examples). In the matrix a red – yellow – green colour scale has been implemented: red cells indicate where images have been frequently grouped together while green cells indicate where images have been less frequently grouped together. A clear observation to make from the matrix is that the range and frequency of the identity-merge errors can be seen in the top left/bottom right quadrants. Interestingly there are blocks of green indicating that certain groups of images were never merged with each other. Reflecting on the images this occurs with, it could be suggested that these blocks represent the images of each face that are very similar to each other and very dissimilar to the other face, therefore are never grouped together. Another key observation is that there appear to be two or three images that participants particularly struggled with and which resulted in frequent errors, specifically images 11 and 14 of Face 1 and image 25 of Face 2. This could suggest that these images are especially varied and so difficult to categorise.

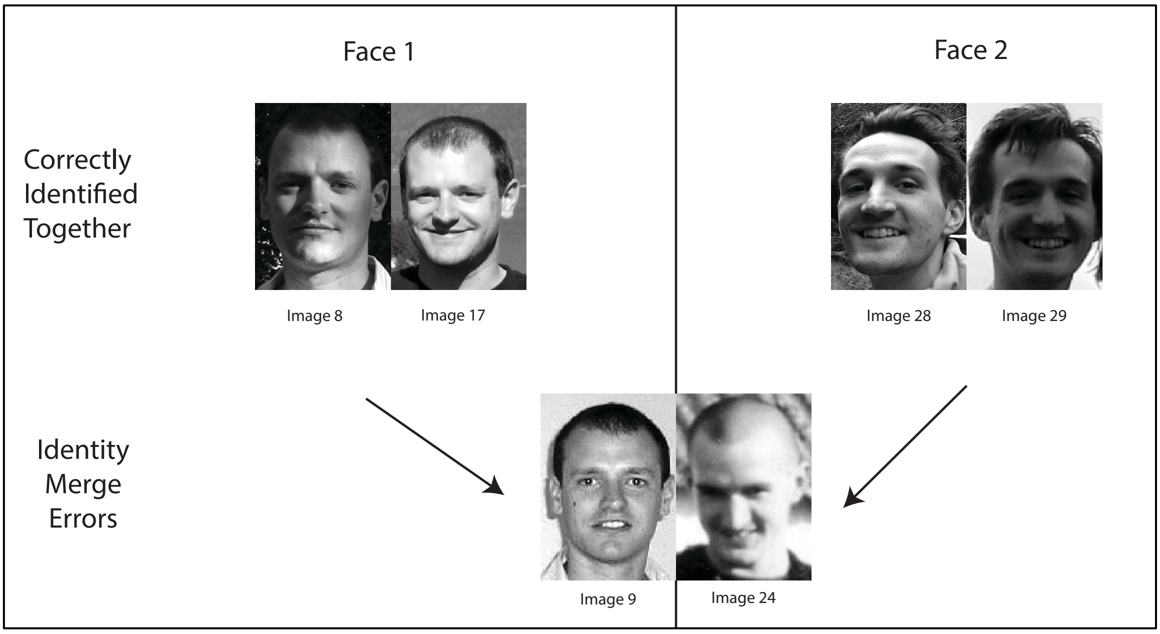


Figure 2.20. **Experiment 3: Examples of identification errors.**
Examples are shown of frequently correctly identified images in the top row and a frequent identity-merge error in the bottom row (participants were unable to make identity-split errors).

Discussion

In this experiment participants were provided with the additional information that two identities are present amongst the 40 images and so were able to reach the correct number of two identities. However only 9 participants accurately identified all 40 images. The remaining participants made a range of 1 – 12 errors. As in Experiment 2, participants were unable to make identity-split errors as whenever they 'split' images of the same identity they created an identity-merge error; therefore all errors made in Experiment 3 were identity-merge errors.

The frequency of identity-merge errors decreased from Experiment 2. This reduction in identity-merge errors could suggest that the visual constraint in Experiment 2 was causing some of the difficulties experienced by participants. However all images were still at some point wrongly identified with an image of the other face. Interestingly from the confusion matrix it is possible to see where certain images of Face 1 and Face 2 were *never* wrongly identified. From this it could be suggested that participants were grouping together very similar images and were able to see that some images were too dissimilar to group together, for example images 5, 6 and 7 of Face 1 were never categorised with images 28, 29 and 30 of Face 2. Very similar images were often images that showed the faces with similar hairstyles, facial expressions, or images that looked as though they were taken within a short space of time. Therefore, it appears as though participants used superficial cues such as hairstyle, as well as more in depth facial cues such as to identify these clusters of images. This illustrates that for the more similar, less varied images, participants were better at within-person identification.

The results from Experiment 3 show an increase of identity-merge errors in comparison to Experiment 1 and a slight decrease in comparison to Experiment 2. In Experiment 3 participants were made aware that two identities were present in the stimuli but did not have a reference image for each identity. Therefore the task developed to become a task of grouping images together and create identities based on what participants thought each identity looked like. As a result participants made frequent identity-merge errors, clearly illustrating the difficulties that can be faced in identification across varied images.

2.5 General Results: Experiments 1, 2 and 3

In Experiment 1 participants were asked to sort 40 images by identity, unaware of the number of identities present in the images. Performance was surprisingly poor with participants struggling to accurately identify the two faces, instead creating a mean of 9.7 distinct identities. Identity-split errors were very frequent with participants splitting up images of the same face to create different identities. Identity-merge errors were less frequent but did occur, with 13 of the 40 participants grouping images of the different faces together to create a merge identity.

In Experiment 2 participants were asked to sort the same 40 images by identity and were told that two identities were present, and given a reference image for each identity. Participants all made two identities but their error performance was poor: all except one image were at some point matched to the wrong reference image, and all images were at some point wrongly identified with an image of the other identity.

In Experiment 3 participants were again asked to sort the same 40 images by identity and this time were only made aware of the number of identities present in the images. Participants all made two identities but much like Experiment 2 their error performance was poor: all images were at some point wrongly identified with an image of the other identity.

It is important to consider the low statistical power throughout Experiments 1, 2 and 3 caused by comparing low numbers of observations, such as the low number of familiar participants, or the relatively low error values, such as the comparison of the number of identity-split and identity-merge errors. As the numbers of errors compared between participants, and between Face 1 and Face 2, are relatively low, it is important to be conservative in the interpretation of results. To improve how these comparisons are reported, a lower significance value could be used i.e. p </> 0.01 as opposed to p </> 0.05. To improve the reliability of the familiarity comparisons, data from more participants familiar with Face 2 would need to be collected, ideally so that the two groups of participants are of equal size. This point is developed in the recruitment of participants for Experiment 4 in Chapter 3.

By using the same stimuli and same basic design of an image-sorting task, Experiments 1, 2 and 3 can be closely compared. Any differences and patterns throughout the experiments may allow for a better understanding of what is happening in each experiment and also how identification behaviour changes across different levels of information.

Identity-split errors

The number of identities created by participants in Experiment 1, 2 and 3 are shown in Figure 2.21. This effectively illustrates the extent of the identity-split errors made by participants in Experiment 1: participants were unable to group images of the same face together. Asking participants to identify across within-person variability was a simple way of demonstrating how difficult it is.

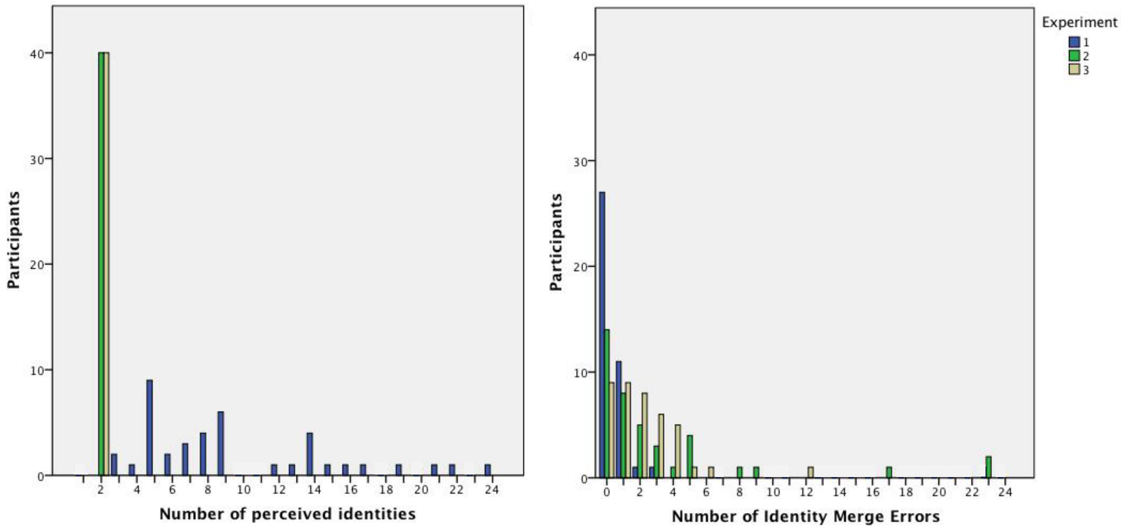


Figure 2.21. **Experiments 1, 2 and 3: Perceived identities and frequency of errors.** Histograms illustrating the number and range of identities, and the number and range of identity-merge errors created by participants in Experiments 1, 2 and 3.

Identity-merge errors

Due to the introduction of constraints in Experiments 2 and 3 the only errors participants could make were identity-merge errors. In Experiment 1 participants made a mean number of 0.4 identity-merge errors, in Experiment 2 participants made a mean number of 3.2 identity-merge errors and in Experiment 3 participants made a mean number of 2.1 identity-merge errors (see Figure 2.21).

A one-way ANOVA was conducted to analyse whether the frequencies of identity-merge errors were significantly different across all three experiments. A significant difference was found (F (2,117) = 6.79, p < 0.05, η_p^2 = 0.10) and posthoc tests showed that significant differences occurred between Experiments 1 and 2 ($p < 0.01$) and between Experiment 1 and 3 ($p < 0.01$) therefore significantly more identity-merge errors were made in Experiments 2 and 3 than in Experiment 1. No significant difference was found between errors made in Experiments 2 and 3 ($p > 0.05$). These results suggests that the additional information provided in Experiments 2 and 3 caused significant problems for participants trying to reach the two identities specified.

Matrices

The identification behaviour of all participants in Experiments 1, 2 and 3 can be visualised and compared through the confusion matrices (see Figure 2.22). The different patterns of identity-merge errors are clear across all three matrices. A way to explain the differences in identification behaviour across all three experiments is simply that participants' strategy of identification must change with each task.

In Experiment 1 participants were asked to complete a "free sorting" task with the images and no further information. Here the difficulties of grouping different images of the same face together are demonstrated. In Experiment 2 participants are again asked to identify the images but with the information that two identities exist and with a reference image. Participants therefore need to match each image to the right reference image, resulting in multiple pairwise comparisons being made between each image to each of the reference images. Finally in Experiment 3 participants only know that two identities exist in the images. They therefore begin to identify clusters of similar looking images, slowly building their own representation of each identity until they can group all of the images into two. This concept of different identification strategies could be applied to face identification tasks within and outwith experimental settings.

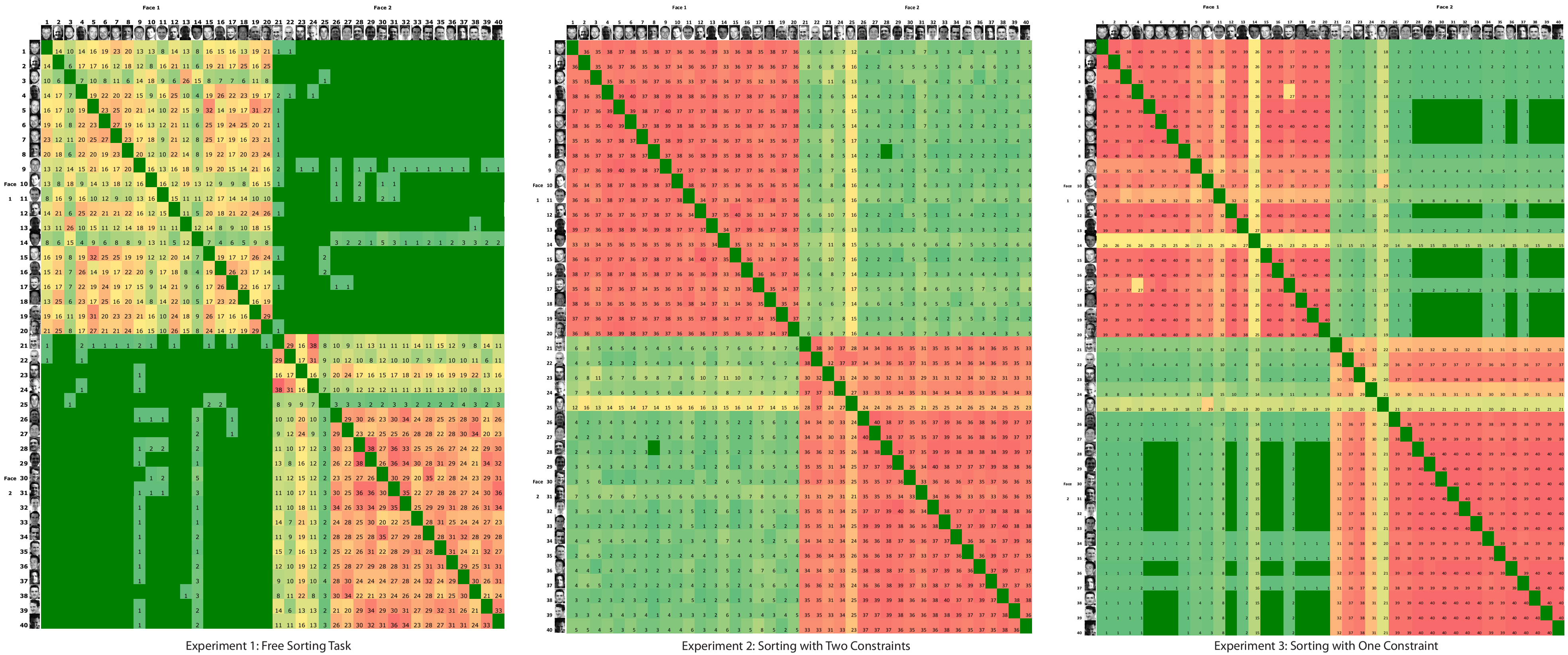


***Figure 2.22*. Confusion Matrices from Experiments 1, 2 and 3.**

2.6 General Discussion: Experiments 1, 2 and 3

Within-person variability

In Experiment 1 participants were unable to correctly identify the two different faces present in the 40 images. Instead participants made a mean number of 9.7 distinct identities, significantly more than the two identities. Interestingly only 13 of the 40 participants made identity-merge errors i.e. grouped images of Face 1 with images of Face 2. This suggests that the problem of identification is not grouping two identities apart but grouping the same identity together. This result replicates findings from Jenkins et al., (2011) that images of two identities can be confused as multiple identities, and also demonstrates that this effect of within-person variability can occur with different faces.

In Experiments 1 – 3, and in the study from Jenkins et al., (2011), identification across within-person variability is examined with multiple images of two different faces. It is important to point out that this is a very small number and so generalising conclusions about identification based on these findings should be done so carefully. Two faces are not an adequate representation of the wider population; it could be argued that all faces may not vary in the same way that Face 1 and Face 2 vary. To overcome this limitation in future identification research, within-person variability should be considered across a larger number of faces. A possible future experiment could simply be based on the image sorting tasks laid out in this chapter, with the additional introduction of new faces.

This study also explored the effect of familiarity with the faces used as stimuli. In Experiment 1, 9 of the 40 participants were familiar with Face 2 (a lecturer at the University of Glasgow). Of the familiar participants only two correctly identified all 20 images of Face 2. Overall the familiar participants made a mean number of 2.3 identities for images of Face 2. Jenkins et al., (2011) also explored the effect of familiarity in an image-sorting task and found better performance with low error rates from participants familiar with the faces used. Previous research has also widely shown that identification of familiar faces is surprisingly good even in poor quality images (Burton et al., 1999; Bruce 1982).

Therefore better performance could have been expected from participants familiar with Face 2.

However the errors made by participants here could simply be explained by the extent of within-person variability shown in the images of Face 2. Within-person variability includes variations that come with aging; participants familiar with Face 2 had known the person over a short period of time as a university lecturer. Therefore their awareness of the within-person variability of Face 2 was limited to within a certain age range. Any images that represented Face 2 as a younger person may have been particularly difficult to identify for the participants familiar with him as an older person. This raises the interesting question of whether our awareness of a familiar face's variability can reach past the restriction of the period of time in which we have known them.

Alternatively this result could be a reflection on the level of familiarity of Face 2 to participants. If familiarity is thought of as a graded concept perhaps identification of a lecturer cannot be compared to identification of a highly familiar face such as a famous face or a personally familiar face. However previous research has found familiarity effects between students and their lecturers before (Burton et al., 1999). Therefore this result could simply be a reflection on the difficulty of identification across such varied images, as described above.

Interestingly being familiar with Face 2 did not help participants with the task of identification of Face 1. It could be suggested that if participants were familiar with Face 2 (and aware of the variability in the images of Face 2) they would be able to apply that within-person knowledge to images of Face 1. This did not happen, as performance with Face 1 was no different between familiar and unfamiliar participants. Therefore being aware of the within-person variability that can occur with one face does not help to see the within-person variability that might occur with another face.

This finding is important to consider as it has a key implication for future research. It could be suggested that if we have a better awareness of the extent of within-person variability of a face then our identification ability across multiple images of the same person can be improved. However the results of this study suggest that awareness of one person's variability does not apply to other people. Instead understanding within-person variability involves learning one face at a time. In future research if improvement through learning is investigated, there needs to be an element of encouraging participants to apply their knowledge about one face to images of another face, as opposed to solely focusing on improving identification through awareness of within-person variability for faces independently.

The findings of Experiment 1 illustrate the importance of considering within-person variability in face processing research. Given the poor identification performance, it is surprising that previous research has not explored the variability that can occur in one person's facial appearance more thoroughly. Perhaps within-person variability has not been explored because we are unaware of the extent of variability. Face recognition and identification tasks typically use a limited number of different images of the same person (Bruce 1982; Burton et al., 1999; Bruce et al., 2001). This suggests that previous research is aware that one face can vary but the full scale of within-person variability has not been completely acknowledged.

Free Sorting and Constrained Sorting

Following the poor identification performance in Experiment 1 constraints were introduced to Experiments 2 and 3 to establish if there were conditions in which accurate identification could be achieved. In this study the term constraints was given to additional information that would limit participants' identification behaviour. In Experiment 2 the number of identities present was provided as a descriptive constraint and reference images were provided as a visual constraint. With the constraints participants were able to come to the right outcome of two identities, however the number of identity-merge errors increased from Experiment 1; participants were more likely to wrongly identify an image by combining images of Face 1 and Face 2. This suggests that the constraints given to participants were causing confusion and resulted in an increase of identity-merge errors. It could be said that the constraints were actually negatively influencing participants' accuracy.

In Experiment 3 the number of identities present was provided as a descriptive constraint. Again participants were able to reach the correct outcome of two identities and again the number of identity-merge errors was increased from Experiment 1, however the number of identity-merge errors decreased in comparison to Experiment 2. This again suggests that giving participants a constraint actually confused identification across the two faces. The slight decrease in identity-merge errors in comparison to Experiment 2 could suggest that the visual constraint was particularly confusing for participants.

This difference could also be a reflection of the change in task strategy in Experiments 2 and 3. In Experiment 2 participants have a reference image for each identity requiring each image to be matched to one or other of the references. In Experiment 3 participants have to decide for themselves what each identity looks like, grouping similar images together until all images have been identified into two face representations. For each face there were set images that were never grouped together. This suggests that participants determined what each identity looked like by grouping together very similar images and so grouping apart very dissimilar images. From there they could then build up each identity by discarding images that could not be categorised with these principle groups of images.

This difference in approach to identification may tell us a lot about how images are processed. Research has described face processing as split into pictorial codes and structural codes (Bruce & Young 1986). In Experiment 3 it appears as though participants were first using pictorial codes to group together similar images, and then using structural codes to process images in more depth and group together images in which the face appears similar. In Experiment 3 participants were forced to make two identities therefore their processing would have been highly skewed by the knowledge that they could not divide the images into more than two groups. In Experiment 1 this was not the case and participants divided the images into multiple groups. This suggests that relying on pictorial and structural codes will not always result in accurate identification and awareness of how these codes may vary is needed.

Overall the results of Experiments 1, 2 and 3 can tell us a lot about how within-person variability is processed in faces. The main finding is that participants struggled to identify across within-person variability for unfamiliar faces. This suggests that when we process images of unfamiliar faces any within-person variability perceived is often concluded to be between-person variability. This could be due to having little awareness of within-person variability. However we must have some understanding of within-person variability as participants' performance with images of a familiar face was better. This suggests that familiarity with a face includes the awareness of how the face can vary.

However it is important to understand that having awareness of within-person variability for familiar faces does not apply to unfamiliar faces. Being familiar with a face might equate to having awareness of within-person variability but it is for that face alone; awareness of within-person variability does not seem to be a transferrable skill. Instead it would appear that within-person variability must be learnt for each face independently. This could suggest that faces vary differently and as a result general rules of within-person variability cannot be applied. From this it can be concluded that a greater understanding of within-person variability is needed in research.

The results from Experiments 1, 2 and 3 strongly outline the extent of within-person variability and the wide effect it can have on identification behaviour. A key finding especially from Experiments 2 and 3 is that images of different faces are often confused as one face. To understand this finding a closer look at the two faces used in this study is needed. The faces used in this study are actually quite similar looking for example both faces are of a similar age with dark facial features. The faces used by Jenkins et al., (2011) in their image-sorting task were also similar in appearance. The low level of between-person variability could be causing the confusion for participants.

This raises the question of how between-person variability relates to within-person variability in image-sorting identification tasks. If between-person variability is high does within-person variability seem less varied by comparison? In Experiments 1, 2 and 3 it could be said that there were no salient differences between the two faces; if there were salient differences between the faces would identification improve? In terms of the errors that can be made in image-sorting tasks, if the number of identity-merge errors reduces due to an increase in between-person variability what would happen to the frequency of identity-split errors? To explore the interaction of between-person variability and within-person variability, similarity between the faces used in an image-sorting task could be manipulated to show highly similar or highly dissimilar faces.

The results in Experiment 1 clearly illustrate the difficulties in identifying multiple images of one face. Experiments 2 and 3 illustrate that even with additional information about the images it is still possible to wrongly identify images of one face as another person. These findings strongly outline the extent of within-person variability and the wide effect it can have on identification behaviour. Developing on from these findings it would be of interest to explore whether identification difficulties occur with different faces and whether between-person variability in faces influences the perceived within-person variability. Finally it can be concluded that face-processing research needs to seriously consider the extent of within-person variability and it's effect on our ability to recognise and identify a person.

Chapter Three

Between-Person Similarity in Face Identification

3.1 Introduction

As discussed in Chapter 2, previous face identification and recognition research typically does not focus on the real extent of within-person variability. Instead studies have used one or two different controlled images to represent a face resulting in a limited view of within-person variability (Bruce 1982; Burton et al., 1999; Bruce et al., 2001; Clutterbuck and Johnston 2004; 2005). Research that has examined within-person variability in depth is a study from Jenkins, White, van Montfort and Burton (2011). Using an image-sorting task, Jenkins et al., (2011) found that participants were surprisingly bad at identifying two faces from multiple images of the faces. This finding was replicated in Experiment 1 and extended to show that difficulties in identifying across within-person variability also occur with image of other faces.

To develop the research on within-person variability it is important to establish whether the effect is limited to the images of Face 1 and Face 2 used in Chapter 2. Considering the errors made in Experiments 1, 2 and 3 when grouping images of Face 1 with images of Face 2 it is also important to establish whether identification difficulties occurred due to the similarity of the two faces.

With this is mind this chapter explores the identification performance for multiple images of two highly similar faces and two highly dissimilar faces (Experiments 4 and 5). To achieve highly similar face stimuli images of two identical twins were used. Images of identical twins have previously been used in research exploring the learning process of a face (Robbins & McKone 2003) but have not been used in identification tasks. To achieve highly dissimilar face stimuli images of faces used in previous experiments (Face 1 and Face 2) and the highly similar faces (Face 3 and Face 4) were combined i.e. Face 1 with Face 3, Face 2 with Face 4.

To overcome any issues of similarity a free sorting task is implemented with multiple images of just one face in Experiment 6. Finally to put the findings into the context of previous research, a free sorting with multiple images of just one face is conducted using images taken from a database aimed at encompassing variations in appearance across pose, illumination and expression (Gross, Matthews, Cohn, Kanade & Baker 2010). If the database successful captures within-person variability the results from this task will be directly comparable with the results from Experiment 6 using ambient images.

3.2 Experiment 4: Free sorting with highly similar faces

To explore identification with highly similar faces 20 images of two identical twins were collected, 40 images in total. The twins were both previously politicians in Poland therefore images of them were easily accessible online. In Experiments 1, 2 and 3 familiarity with Face 1 and 2 was established after participants competed the task. In Experiment 4 familiarity was deliberately tested by recruiting Polish and non-Polish participants. The two faces in this experiment will be referred to as Face 3 and Face 4.

Method

Participants

Twenty-four undergraduate students (5 male: 19 female, age range 19 – 56) recruited from the School of Psychology subject pool, participated in Experiment 4 in return for a small payment. Of the 24 participants, 12 participants were familiar with the faces in the stimuli set and 12 participants were unfamiliar with the faces in the stimuli set.

Familiarity was established at the end of the experiment; participants were asked if they had recognised or were familiar with any of the identities and indicated their familiarity with a “Yes/No” response. All of the Polish (familiar) participants confirmed that they had recognised the two Polish politicians and were aware that the men were twin brothers. All of the non-Polish (unfamiliar) participants confirmed that they were not familiar with the men.

Stimuli and Design

Twenty images of each face were gathered online, edited so that the images were cropped to show just the face and re-sized to the regulated passport photograph size (45mm x 35mm). The criteria for images selected online was for high-quality images, with no other figures or faces included, showing the full face and with neutral backgrounds. The images were printed in black and white, and laminated to form a set of images participants could handle and work with (see Figure 3.1). All participants were given the same set of images to identify and the "free sorting" design of Experiment 1 was replicated.

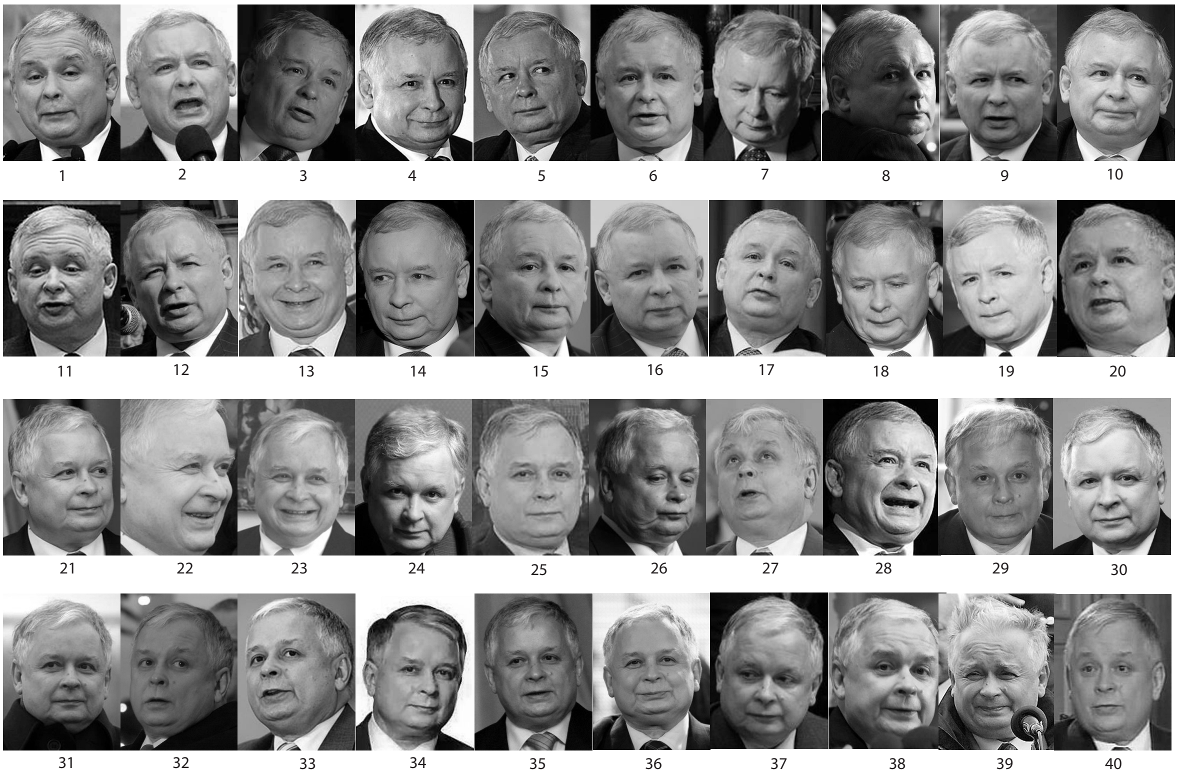


Figure 3.1. **Images used in Experiment 4.**
Twenty images of two different faces made up the stimuli for Experiment 4: Twenty images of Face 3 are shown in the top two rows (numbered 1 – 20) and twenty images of Face 4 are shown in the bottom two rows (numbered 21 – 40).

Procedure

Participants were given the set of 40 images and asked to: "sort the images by identity so that images of the same face, or person, are grouped together". Participants were also informed that there was no time restriction, although the trial time was recorded, and that they were free to create as many or as few identities as they saw fit. Participants' outcome was recorded in addition to how the images were categorised.

Results

Participants familiar with the Face 3 and 4 will be referred to as familiar participants and participants unfamiliar with the Face 3 and 4 will be referred to as unfamiliar participants. All familiar participants made 2 identities from the

40 images, while unfamiliar participants made a mean number of 6.17 identities (median 4.5; mode 3; range: 3 – 19) reflecting the number of distinct identities perceived in the set (see Figure 3.2). Comparing the results with a paired sample t-test shows that unfamiliar participants made significantly more identities than familiar participants [t (11) = 3.01, p < 0.05]. One of the participants made 19 identities, removing this outlier reduced the overall mean number of identities made by unfamiliar participants to 5, but still remained significantly higher than the number of identities made by familiar participants [t (11) = 3.88, p < 0.01]. Therefore unfamiliar participants made significantly more identities than familiar participants.

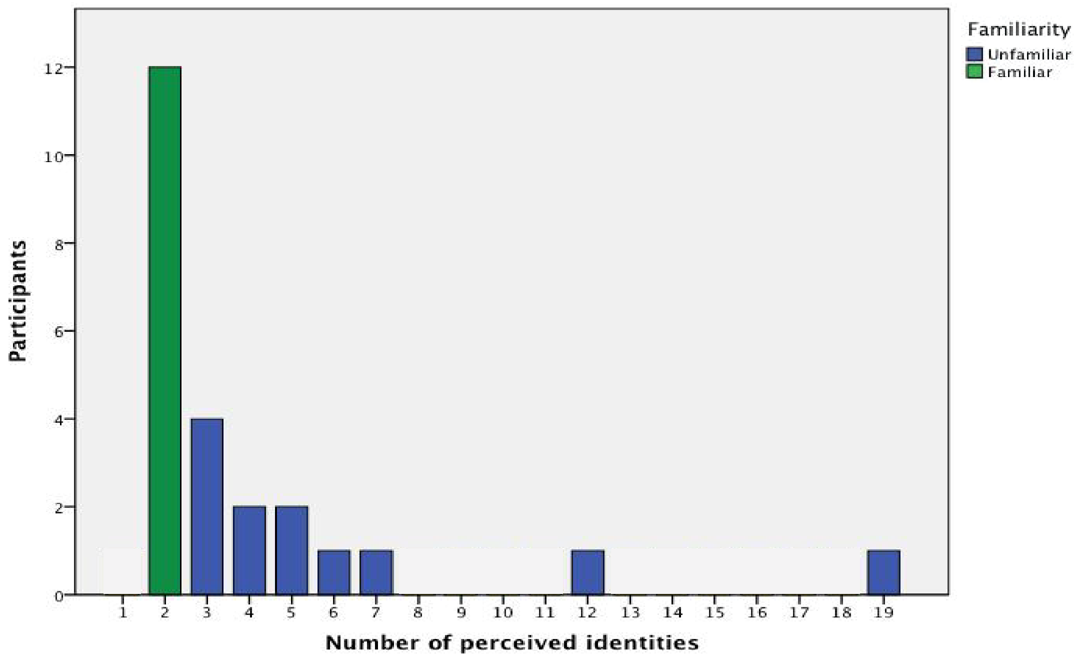


Figure 3.2. **Experiment 4: Number of perceived identities.**
A histogram illustrating the number and frequency of identities perceived by familiar (green) and unfamiliar (blue) participants in Experiment 4.

Familiar participants all made two identities from the images of Face 3 and 4. However both familiar and unfamiliar participants made identity-merge errors i.e. they identified images of one face with another creating a merge identity.

Unfamiliar Participants

Unfamiliar participants made both identity-split errors and identity-merge errors. Unfamiliar participants made a mean number of 5.33 identity-merge errors (median 2; mode 1; range 1-16). This indicates that unfamiliar participants had difficultly telling the two identities in addition to grouping each identity together.

Familiar Participants

Familiar participants knew there was two identities amongst the images and so divided the images into two. Therefore any errors they made could only be identity-merge errors as any splitting of the two faces resulted in the images being merged with the other face. Familiar participants made a mean number of 5.08 identity-merge errors (median 2.5; mode 2; range 1-10). This indicates that despite knowing there are two identities present, participants still found it difficult to accurately identify the two faces.

No significant difference was found between the mean numbers of identity-merge errors made by familiar participants compared with unfamiliar participants [t (11) = 0.09, p > 0.05]. Therefore although familiar participants were expected to be able to identify Face 3 and 4 accurately, they made similar identity-merge errors as unfamiliar participants. This suggests that although familiar participants had the semantic information (that the faces were identical twins), they did not have the perceptual knowledge allowing them to distinguish each twin.

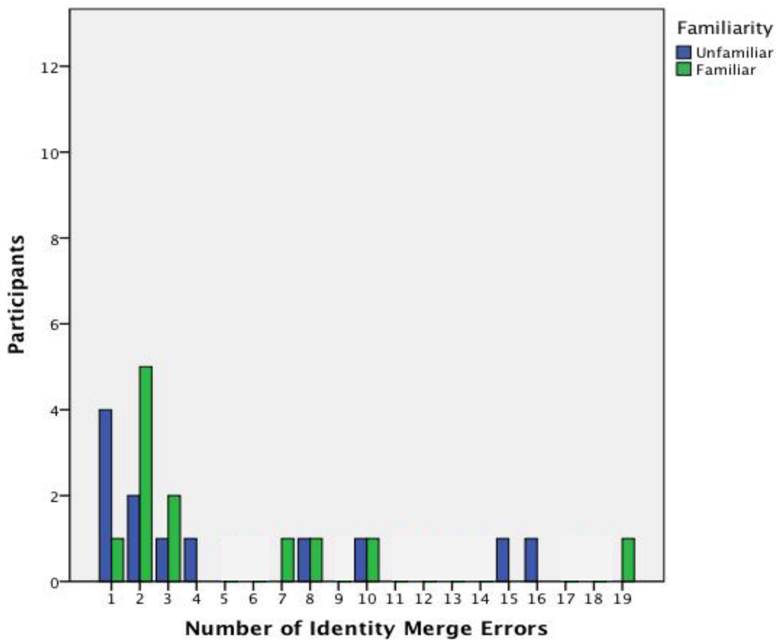


Figure 3.3. **Experiment 4: identity-merge errors.**
A histogram illustrating the range and frequency of identity-merge errors made by familiar (green) and unfamiliar (blue) participants in Experiment 4.

Confusion Matrix

As discussed, none of the unfamiliar participants reached the correct outcome of two identities present in the 40 images instead making both identity-split errors and identity-merge errors. All familiar participants made the correct number of identities but made similar identity-merge errors as the unfamiliar participants.

As in Experiments 1 – 3, to more closely examine participants' identification behaviour a confusion matrix was created. To observe the difference in familiar and unfamiliar identification two matrices were constructed: one for familiar participants (see Figure 3.4) and one for unfamiliar participants (see Figure 3.5). Images of Face 3 are listed 1-20 and images of Face 4 are listed 21-40. Every time one image was grouped with another image it was recorded to observe identification patterns amongst the images. The confusion matrices clearly show where images were frequently grouped together and where images were infrequently grouped together.

As 12 participants took part in each condition (familiar/unfamiliar) the number 12 indicates that data from 12 participants was used to create the matrix (as opposed to the number 40 as in Experiments 1, 2 and 3). The colour scale green (low) to red (high) was applied to illustrate frequencies and infrequencies of images categorisation. Therefore perfect performance would be all red cells in the top left and bottom right matrix quadrants with all green cells in the top right and bottom left quadrants.

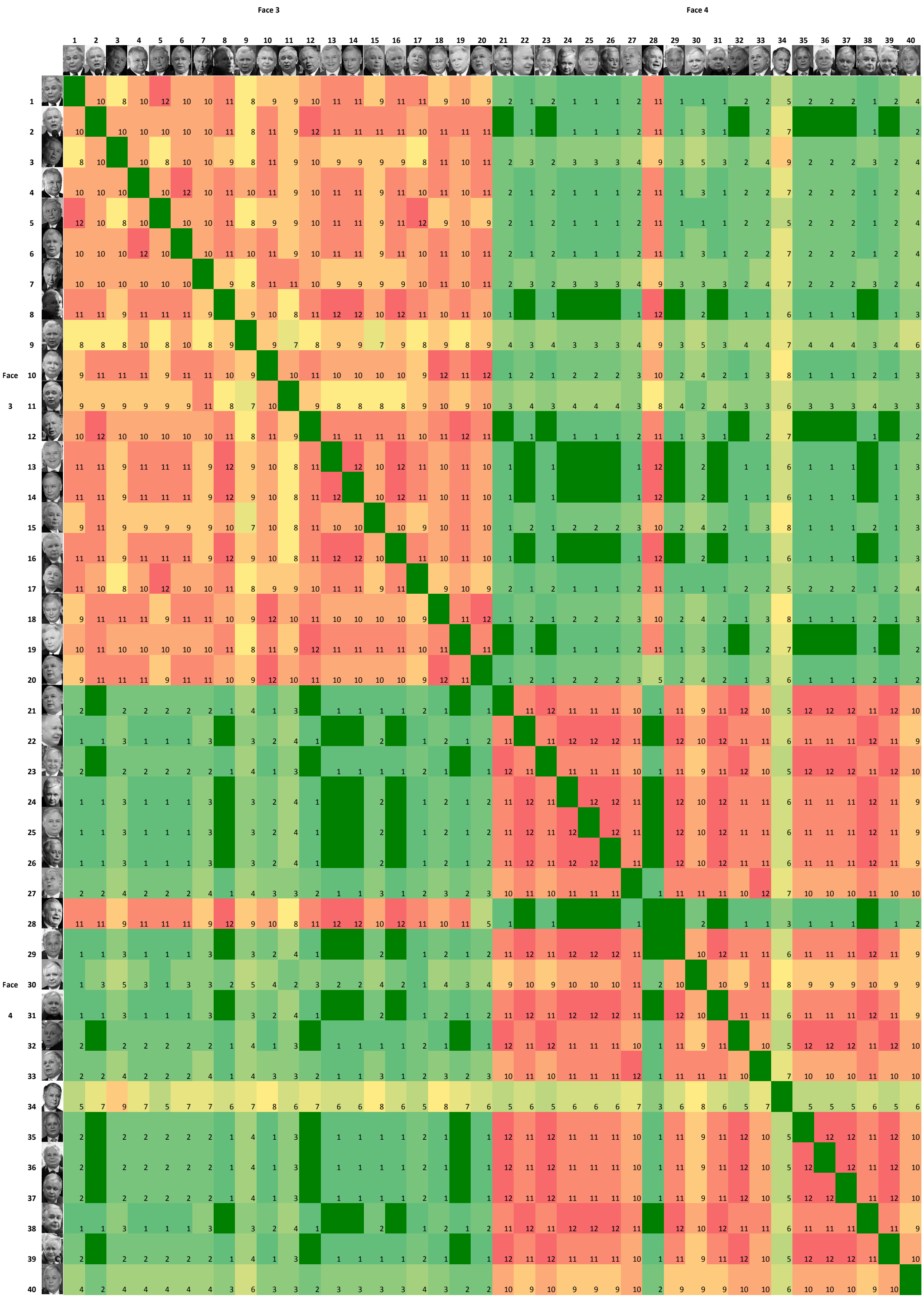


Figure 3.4. **Experiment 4: Familiar Participants Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.

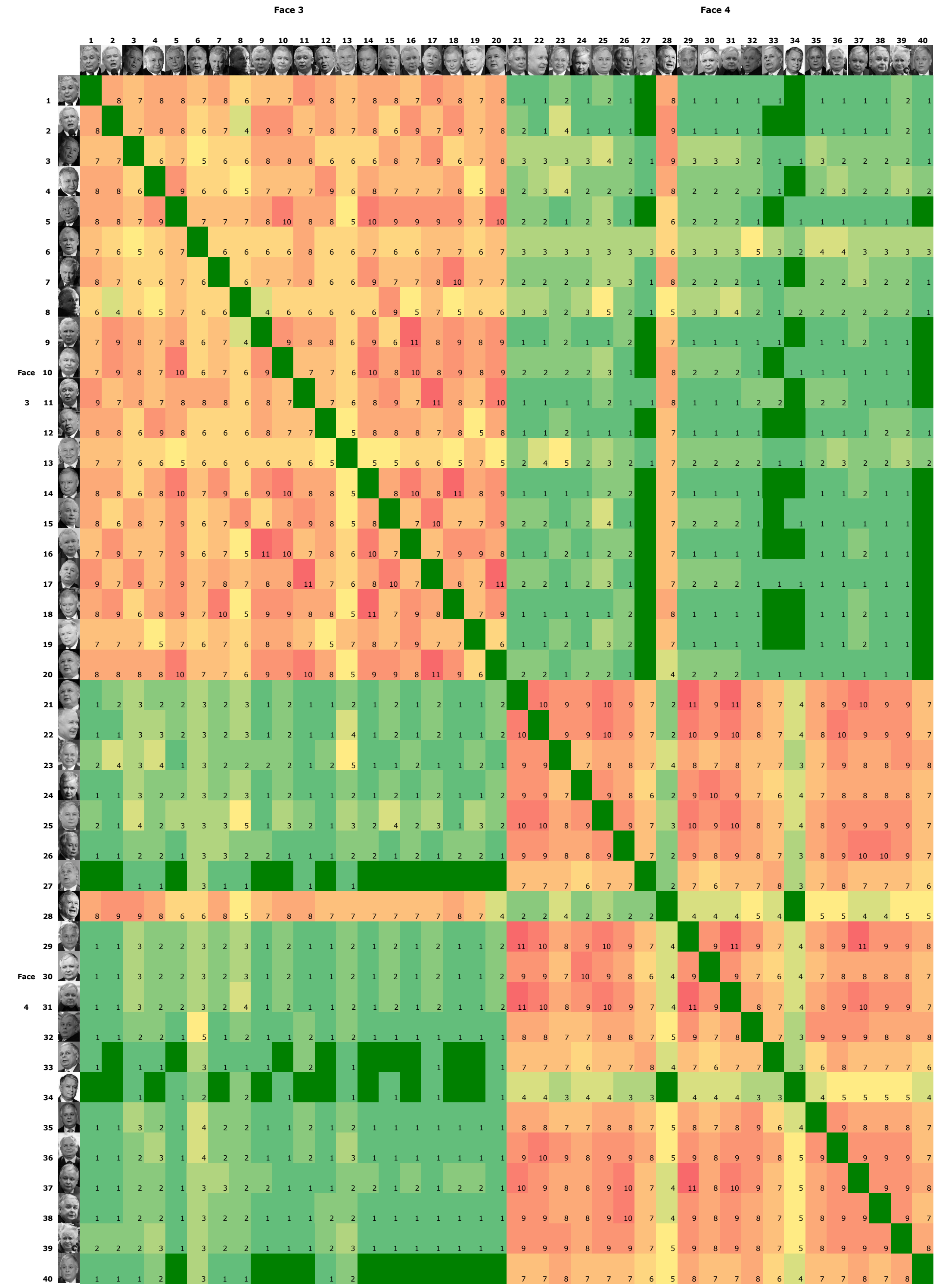


Figure 3.5. **Experiment 4: Unfamiliar Participants Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.

From the confusion matrix it is possible to visualise identification behaviour across familiar and unfamiliar participants. For example comparing the two matrices the more red cells in the familiar participants' confusion matrix reflects the smaller number of identities made by familiar participants. It is also possible to see the similar levels of identity-merge errors made by both groups of participants. Examples of participants' identification behaviour are shown in Figure 3.6. Interestingly it is also possible to see that for both familiar and unfamiliar participants the same images seemed to cause identification problems (see Figure 3.6).

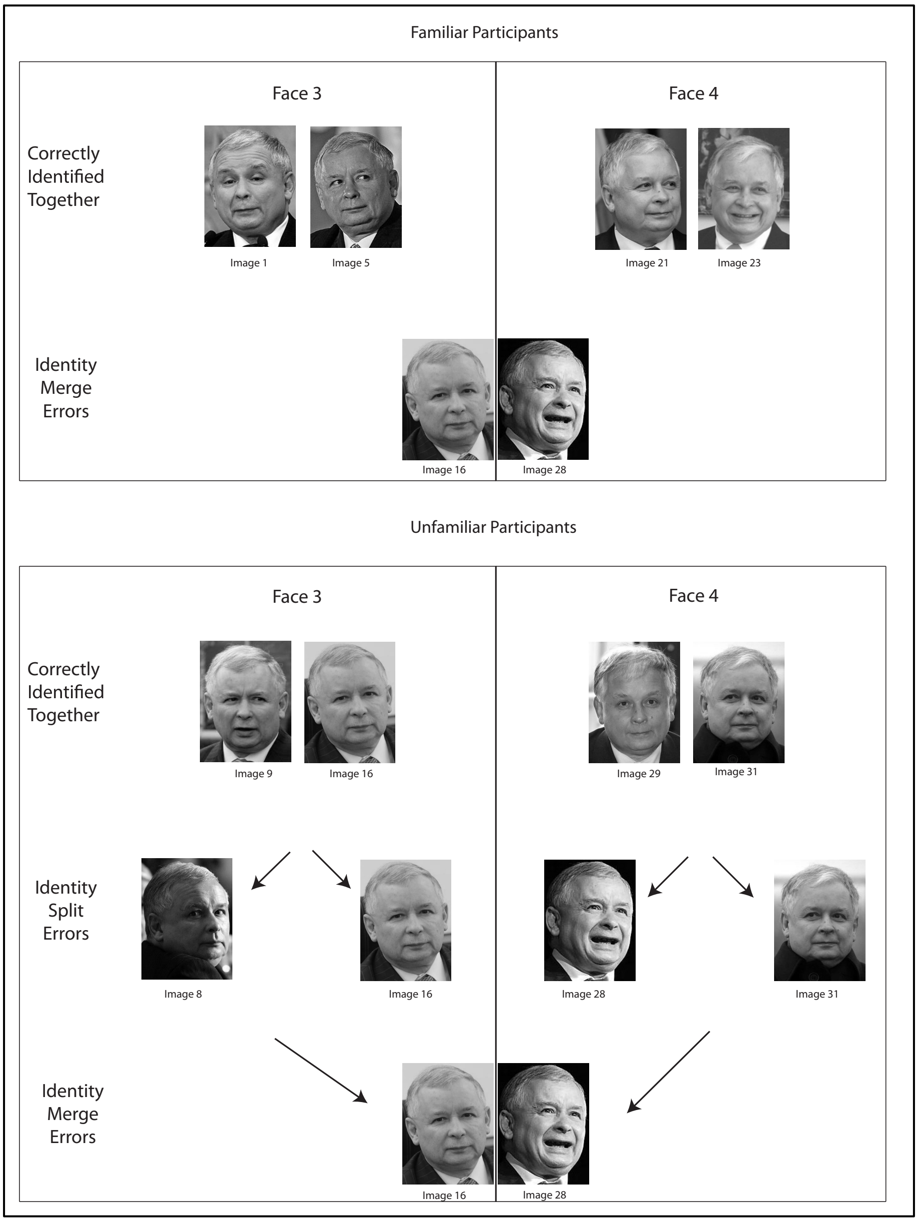


***Figure 3.6.* Experiment 4: Examples of identification errors.**
Examples of identification errors made by familiar and unfamiliar participants are shown separately with frequently correctly identified images in the top row, frequent identity-split errors in the second row (for unfamiliar participants only) and an example of an identity-merge error in the bottom row.

Discussion

In Experiment 4, images of identical twins were introduced to an image-sorting task to examine participants' identification with highly similar faces. The faces used were famous in Poland but unknown in the UK, this allowed for familiarity to be deliberately controlled as both Polish (familiar) and British (unfamiliar) participants could be recruited.

As expected, participants familiar with the faces reached the correct outcome sorting the images into two identities, while participants unfamiliar with the faces made a mean number of 6.1 identities. The number of identities made by unfamiliar participants is surprising: given the images are of identical twins, it could have been expected that participants would perceive the images as of one face. This finding adds to the support of the effect of within-person variability on identification, as even images of identical twins were perceived as too dissimilar to group together.

It could be suggested that the number of identities produced by unfamiliar participants is actually a task effect i.e. when asked to sort multiple images by identity participants may have felt this implied that multiple identities existed amongst the images. In defence of this, it is helpful to consider the way the images were sorted by unfamiliar participants. The images were sorted in much the same way as participants sorted images in Chapter 2, with a focus on basic facial characteristics variations such as expression and hairstyle. Specifically unfamiliar participants used facial cues such as the presence of moles on Face 4 to influence their identification behaviour. This suggests that participants were not being lead by the task instructions but were making visual decisions that they felt were accurate identifications.

Despite reaching the outcome of two identities, familiar participants made the same level of identity-merge errors as unfamiliar participants: all participants made a mean number of 5 identity-merge errors, i.e. wrongly identifying an image of Face 3 with Face 4. This suggests that familiar participants had the semantic information to be able to identify that two identities existed amongst the images but not the visual knowledge to accurately identify each image. It could also reflect the difficulty in distinguishing between the two identical twins.

3.3 Experiment 5: Free sorting with highly dissimilar faces

In Experiment 4 participants unfamiliar with the highly similar faces were very poor at identification, making 6.1 different identities. Participants familiar with highly similar faces were shown to reach the correct number of identities but made the same rate of identity-merge errors as unfamiliar participants. Therefore identification across within-person variability is difficult even when familiar with the highly similar faces. In Experiment 5 identification with highly dissimilar faces was tested by combining one face used in Experiments 1 - 3 with one face used in Experiment 4.

Method

Participants

Twenty-two participants took part in Experiment 5 in return for a small payment; 12 undergraduate students at the University of Glasgow recruited from the School of Psychology subject pool, and 10 members of the public in attendance at the 2011 British Science Festival in Bradford (7 male: 15 female, age range 18-61).

Unlike the previous experiments, none of the participants were familiar with the faces used in the stimuli. This was established by asking participants following the task whether they had recognised, or were familiar with, any of the faces they had seen. All participants indicated that they were not familiar with the faces.

Design and Stimuli

Comparing the stimuli used in Experiments 1 – 3 and in Experiment 4, Faces 1 and 2 are very dissimilar to Faces 3 and 4. The opportunity was therefore taken to create two image sets of two highly dissimilar faces by combining one face from Experiments 1 - 3 and one face from Experiment 4 (see Figure 3.7). As the images had previously been used in Experiments 1 – 4, no further stimuli preparation was required. Using the same "free sorting" design as Experiment 4, stimuli was counterbalanced so that participants were presented with either image set 1 or image set 2 of the stimuli.

Procedure

Participants were presented with either image set 1 or set 2 and asked to: "sort the images by identity so that images of the same face, or person, are grouped together". Participants were also informed that there was no time restriction although the trial time was recorded, and they were free to create as many or as few groups as they saw fit. Participants' outcome was recorded in addition to how the images were categorised.

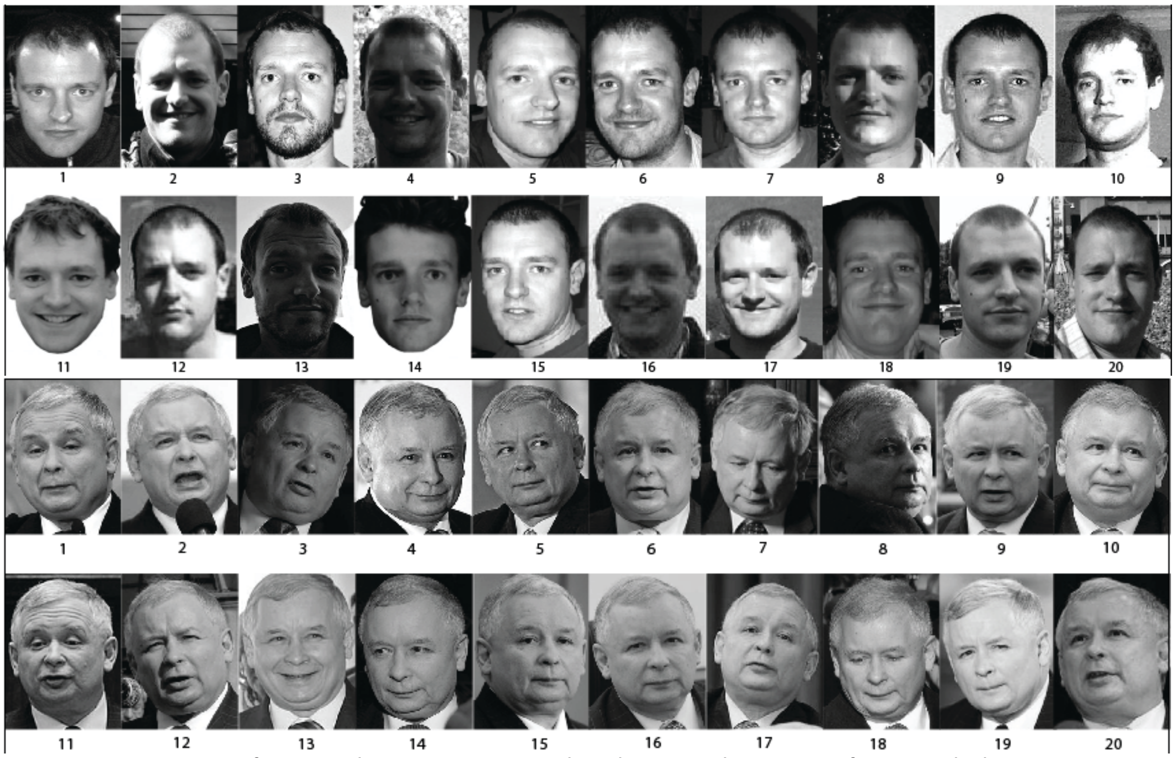


Image Set 1: 20 images of Face 1 in the top two rows (numbered 1 - 20) and 20 images of Face 3 in the bottom two rows (also numbered 1 - 20 but later renumbered to 21 - 40 for the confusion matrix).

Image Set 1: 20 images of Face 2 in the top two rows (numbered 21 - 40) and 20 images of Face 4 in the bottom two rows (also numbered 21 - 40).

***Figure 3.7.* Images used in Experiment 5.**

Two image sets were used in Experiment 5, Image Set 1 comprises of twenty images of Face 1 and Face 3 and Image Set 2 comprises of images of Face 2 and Face 4.

Results

Overall participants made a mean number of 4.7 identities (median 4; mode 3; range 2-12) instead of the two identities present. The result was analysed with a one-sample t-test and it was found that the mean number of identities made was significantly higher than the actual number of identities [t (21) = 5.27, $p <$ 0.01]. This result suggests that participants found it difficult to identify images of two different faces despite the two identities being very dissimilar.

Identity-split errors

Participants made no identity-merge errors illustrating that the faces were dissimilar enough to justify grouping them apart. Any errors participants did make are therefore identity-split errors, suggesting that the problem is grouping images of the same face together. As participants were given either image set 1 or 2, performance with each set is considered separately (see Figure 3.8) to see whether difficulties occurred with one face more than another.

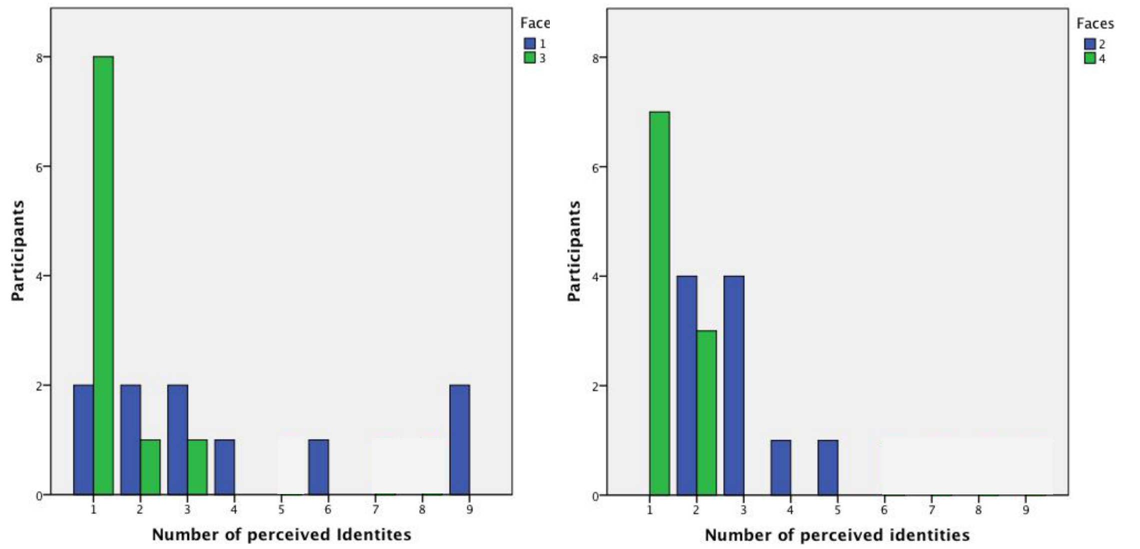


Figure 3.8. **Experiment 5: Number of perceived identities.**
Histograms illustrating the range and frequency of identities perceived by participants in Experiment 5 given either image set 1 (Face 1 and Face 3) or image set 2 (Face 2 and Face 4).

Image set 1: Face 1 and Face 3

Participants made a mean number of 4 identities for Face 1 (median 3; mode 1, 2 and 3; range 1 - 9) and a mean number of 1.3 identities for Face 3 (median 1; mode 1; range 1 - 3) (see Figure 3.8). A significant difference was found [t (18) = 2.76, $p <$ 0.05] indicating that significantly more identity-split errors were made for Face 1 than Face 3. This suggests that it was more difficult to group together images of Face 1 than images of Face 3.

Image set 2: Face 2 and Face 4

Participants made a mean number of 2.9 identities for Face 2 (median 3; mode 2 and 3; range 2 - 5) and a mean number of 1.3 identities for Face 4 (median 1; mode 1; range 1 - 2). A significant difference was found [t (9) = 4.31, $p < 0.01$] indicating that there is a significant difference between the identities made for Face 2 and Face 4. As with image set 1, this suggests that it was more difficult to group together images of Face 2 than images of Face 4.

Improvement in performance

Two of the 22 participants correctly identified the two identities present. As this is the first time in this study that participants have been able to accurately identify all images in a free sorting image task, it could be suggested that the dissimilarity of the faces helped improve identification performance. Furthermore 15 of the 22 participants correctly identified all images of one of the presented faces (Face 3 or Face 4 depending on the image set). This further suggests that the dissimilarity of the faces helped identification. From this it could be concluded that the wider the between-person variability appears, the less varied within-person variability appears but only for some faces (i.e. Face 3 or Face 4).

One way to analyse whether wider between-person variability improves identification performance is to compare the results of Experiment 5 to the results of Experiment 1 and Experiment 4. By comparing the number of identities made for each face in Experiment 5, with the number made for each face in either Experiment 1 or Experiment 4, any differences can be observed. As participants in Experiment 5 were all unfamiliar with the faces they identified, only data from participants unfamiliar with the faces will be used from Experiments 1 and 4.

Face 1

For example, in Experiment 1 participants made a mean number of 6.1 identities for Face 1 and in Experiment 5 participants made a mean number of 4 identities for Face 1. An independent-sample t-test found that this is not a significant difference [t (39) = 1.55, $p > 0.05$].

Face 2

In Experiment 1 participants made a mean number of 4.8 identities for Face 2 and in Experiment 5 participants made a mean number of 2.9 identities for Face 2. An independent-sample t-test found that this is a significant difference [t (39) = 2.78, $p < 0.01$].

Face 3

In Experiment 4 participants made a mean number of 4 identities for Face 3 and in Experiment 5 participants made a mean number of 1.3 identities for Face 3. An independent-sample t-test found that this is a significant difference [t (20) = 2.76, $p < 0.05$].

Face 4

In Experiment 4 participants made a mean number of 5 identities for Face 4 and in Experiment 5 participants made a mean number of 1.3 identities for Face 4. An independent-sample t-test found that this is a significant difference [t (20) = 3.11, $p < 0.01$].

From the t-tests it can be concluded that identification performance for Faces 2, 3 and 4 is better in Experiment 5 than in Experiments 1 and 4. This is an interesting finding as it suggests that dissimilarity between the faces in the image-sorting tasks, or the between-person variability of the faces, influences identification performance but only for some faces. Manipulating between-person variability in an image-sorting task results in improved performance for Faces 2, 3 and 4 but not for Face 1. This supports the idea discussed in Chapter 2, that improving awareness or learning of within-person variability is not something that can be achieved for all faces but rather awareness of within-person variability is achieved independently, face-by-face.

Confusion Matrix

For the 10 participants who took part at the British Science Festival data was not collected on how they categorised the images, therefore their data cannot be analysed further. For the remaining 12 participants full data was recorded (as in the previous experiments) and so their data can be analysed further. All remaining 12 participants were given image set 1 i.e. images of Face 1 and Face 3. From this point on, the remaining 12 participants will be referred to as participants.

To more closely examine participants' identification behaviour a confusion matrix was produced (see Figure 3.9). Images of Face 1 are listed 1 – 20 and images of Face 3 are listed 21 – 40. Every time one image was grouped with another image it was recorded to observe identification patterns amongst the images. As in Experiment 4, the number 12 indicates that data from 12 participants was used to create the matrix. Again the colour scale of green (low) to red (high) was applied; perfect performance would therefore show all red cells in the top left and bottom right quadrants with all green cells in the top right and bottom left quadrants.

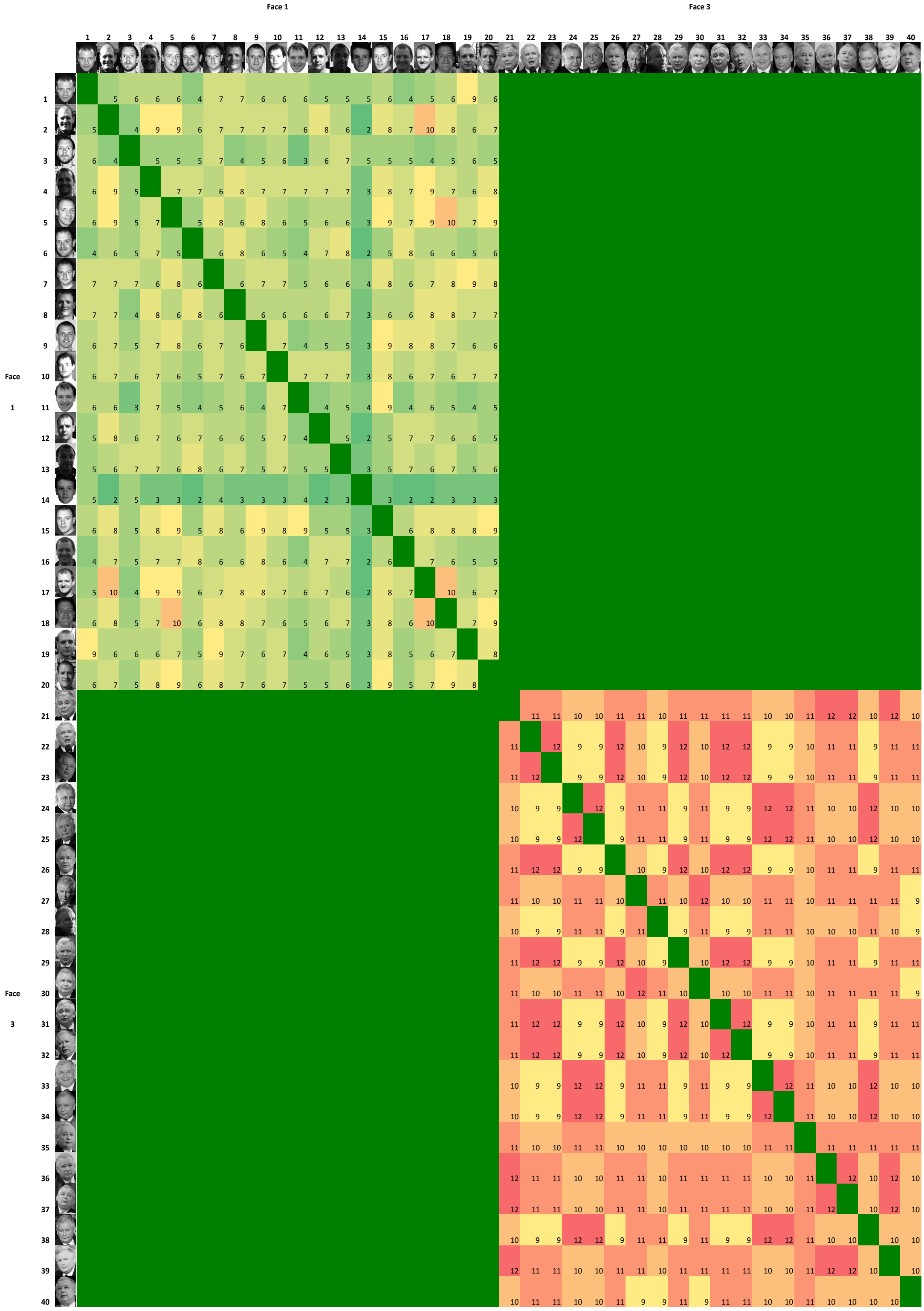


Figure 3.9. **Experiment 5: Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.

From the matrix the absence of identity-merge errors becomes clear in comparison to the matrices in Experiment 4, with all green cells in the top right and bottom left quadrants. It is also clear that images of Face 3 are identified together more frequently than images of Face 1, as more green cells are present in the top left quadrant and more red cells are present in the bottom right quadrant, resulting in less identity-split errors for Face 3. It is also possible to see identification image by image and so see the images that were frequently and infrequently grouped together for Face 1 and Face 3 (see Figure 3.10).

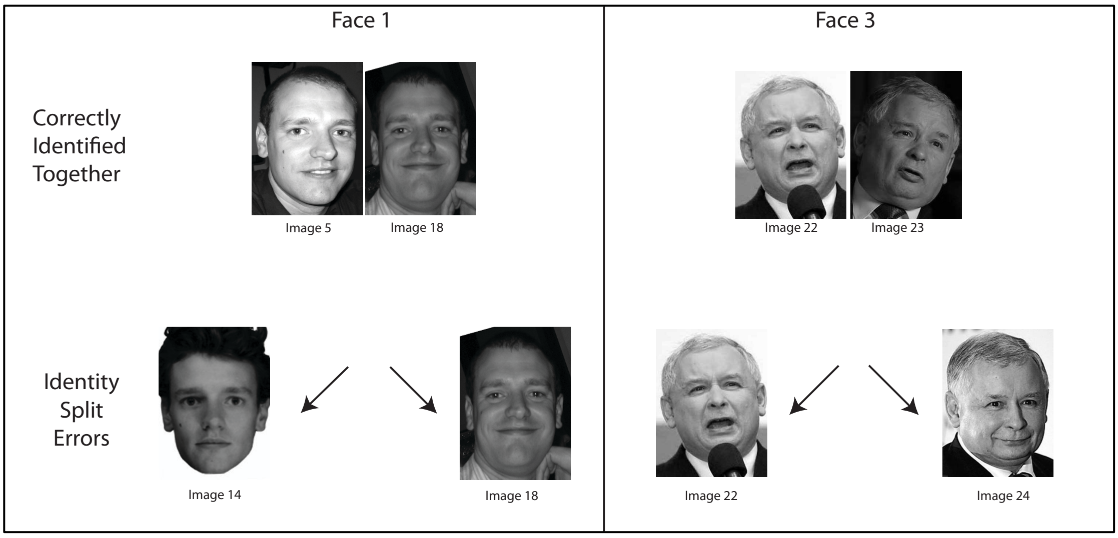


Figure 3.10. **Experiment 5: Examples of identification errors.**
Examples are shown of frequently correctly identified images in the top row and an example of an identity-split error in the bottom row (no identity-merge errors were made).

Discussion

It was proposed at the end of Chapter 2 that changes in the between-person variability of the faces in an image sorting task would effect the within-person variability of the faces. To investigate this idea and following on from the extreme of highly similar faces used in Experiment 4, images of highly dissimilar faces were used in Experiment 5. Unlike the previous experiments, none of the participants were familiar with the faces shown in the images.

In Experiment 5 participants made a mean number of 4.7 identities, less than the 6.1 identities made by unfamiliar participants in Experiment 4. This decrease in the number of identities produced (in comparison to all of the previous image-sorting tasks) suggests that having images of two very dissimilar faces helps to improve identification. Furthermore, 2 of the 22 participants reached the correct outcome and 15 of the 22 participants reached the correct outcome for *one* face. No identity-merge errors were made, reflecting that participants perceived images of the two faces to be too dissimilar to group together. This demonstrates that wider between-person variability helps to improve identification errors in merging different identities.

In addition, comparisons between Experiment 5 and Experiments 1 and 4 indicate that identification performance improved for Faces 2, 3 and 4 in Experiment 5 when highly dissimilar faces were introduced to an image-sorting task. This suggests that an increase in between-person variability can help to improve performance but not for all faces. This supports the idea that within-person variability is learnt on a face-by-face basis as opposed to being something that can be learnt and applied to all faces.

Analysing participants' identification separately for each face shows that less identity-split errors were made for Faces 3 and 4 than Faces 1 and 2. This suggests that increasing the between-person variability helped to improve identification performance for Faces 3 and 4 more than Faces 1 and 2. Alternatively this finding could simply be a reflection of the amount of within-person variability present in the images; for example the images of Face 1 vary more across age than the images of Face 3. Overall it could be concluded that the dissimilarity of the faces helped to improve performance across within-person variability for Faces 3 and 4 but not Faces 1 and 2. This suggests that awareness of within-person variability is not learnt equally across all faces but rather is dependent on the extent of within-person variability face by face.

3.4 Experiment 6: Free sorting with one face

In Experiments 4 and 5 the similarity of the faces used in the image sorting tasks were manipulated to explore whether the two faces used were affecting identification performance. To overcome the issue of similarity the logical development is to test identification performance with images of just one face. Therefore in Experiment 6 participants are given twenty images of just one face to identify in a "free sorting" design task.

Method

Participants

Twenty participants took part in Experiment 6 in return for a small payment; 10 undergraduate students at the University of Glasgow recruited from the School of Psychology subject pool, and 10 members of the public in attendance at the 2011 British Science Festival in Bradford (9 male: 11 female; age range 18 – 63 years).

Design and Stimuli

To explore identification with images of one face, the images used in Experiments 1 - 3 were divided to create two image sets: 20 images of Face 1 made up image set 1 and 20 images of Face 2 made up image set 2 (see Figure 3.11). As the images had previously been used in Experiments 1, 2 and 3 no further preparation was required. Using the same "free sorting" design as previous experiments stimuli was counterbalanced so that participants were presented with either image set 1 or image set 2.

As in Experiment 5, none of the participants were familiar with the faces used in the stimuli. This was established following the task, by asking participants to indicate with a "Yes/No" response whether they recognised or were familiar with the face shown to them. All participants indicated that they were not familiar with the face in their image set.

Procedure

Participants were given either image set 1 or set 2 and asked to: "sort the images by identity so that images of the same face, or person, are grouped together". Participants were also informed that there was no time restriction, although their trial time would be recorded, and that they were free to create as many or as few identities as they saw fit. Participants' outcome was recorded in addition to how the images were categorised.

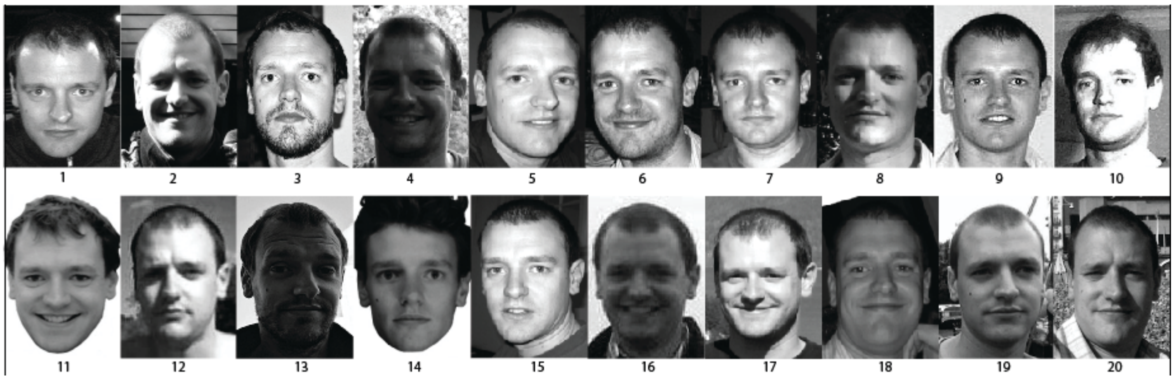


Image Set 1: 20 images of Face 1.

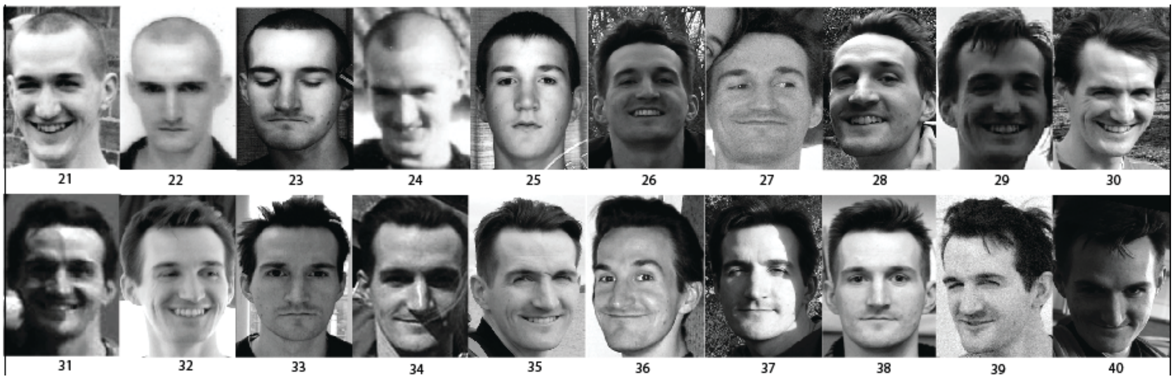


Image Set 2: 20 images of Face 2.

Figure 3.11. **Images used in Experiment 6.**
Two image sets were used in Experiment 6: Participants were given either twenty images of Face 1 shown in the top two rows or twenty images or Face 2 shown in the bottom two rows (numbering of the images is kept constant to the numbering in previous experiments).

Results

As only one identity existed amongst the images given to participants, only identity-split errors could be made. Participants made a mean number of 5.1 identities (median 5; mode 5; range 1 – 10) instead of the one identity actually present. The mean number of identities made by participants is significantly higher than the number of identities actually present [t (19) = 6.91, p < 0.01]. Of the 20 participants 2 participants reached the correct outcome of one identity present in the stimuli. This demonstrates that it is possible to accurately identify one person from multiple images.

To establish whether performance was worse for one face more than the other, performance for each face was considered separately (see Figure 3.12). Participants who were given images of Face 1 made a mean number of 4 identities (median 5; mode 5; range 1-9). Participants who were given images of Face 2 made a mean number of 4.2 identities (median 5.5; mode 6; range 1-10). No significant difference was found between the mean number of identities produced for Face 1 and Face 2 [t (18) = 0.16, p > 0.05], demonstrating that identifying multiple images of the same face is difficult across different faces.

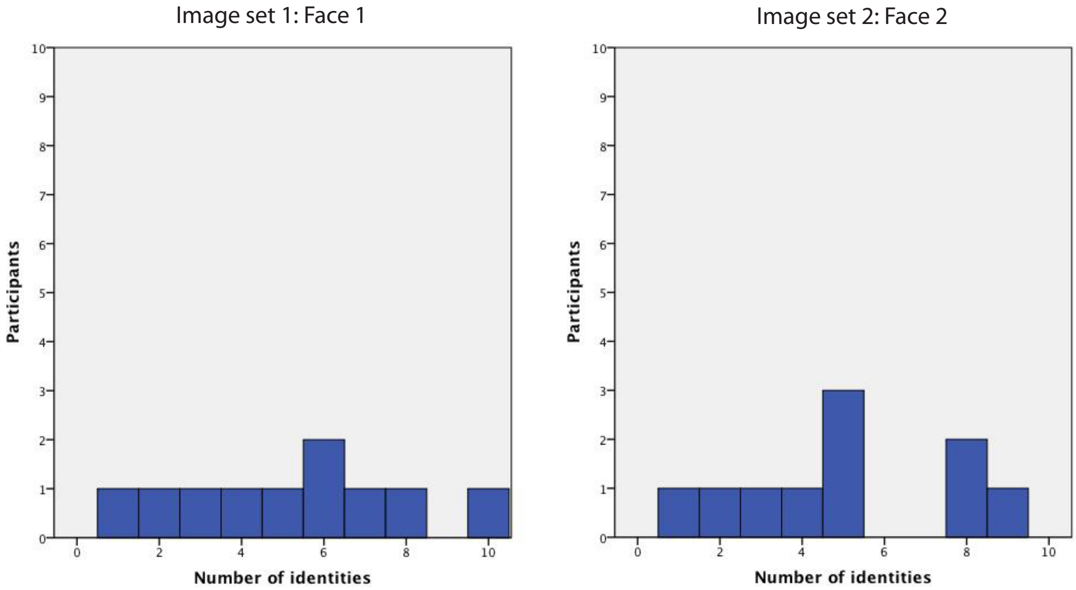
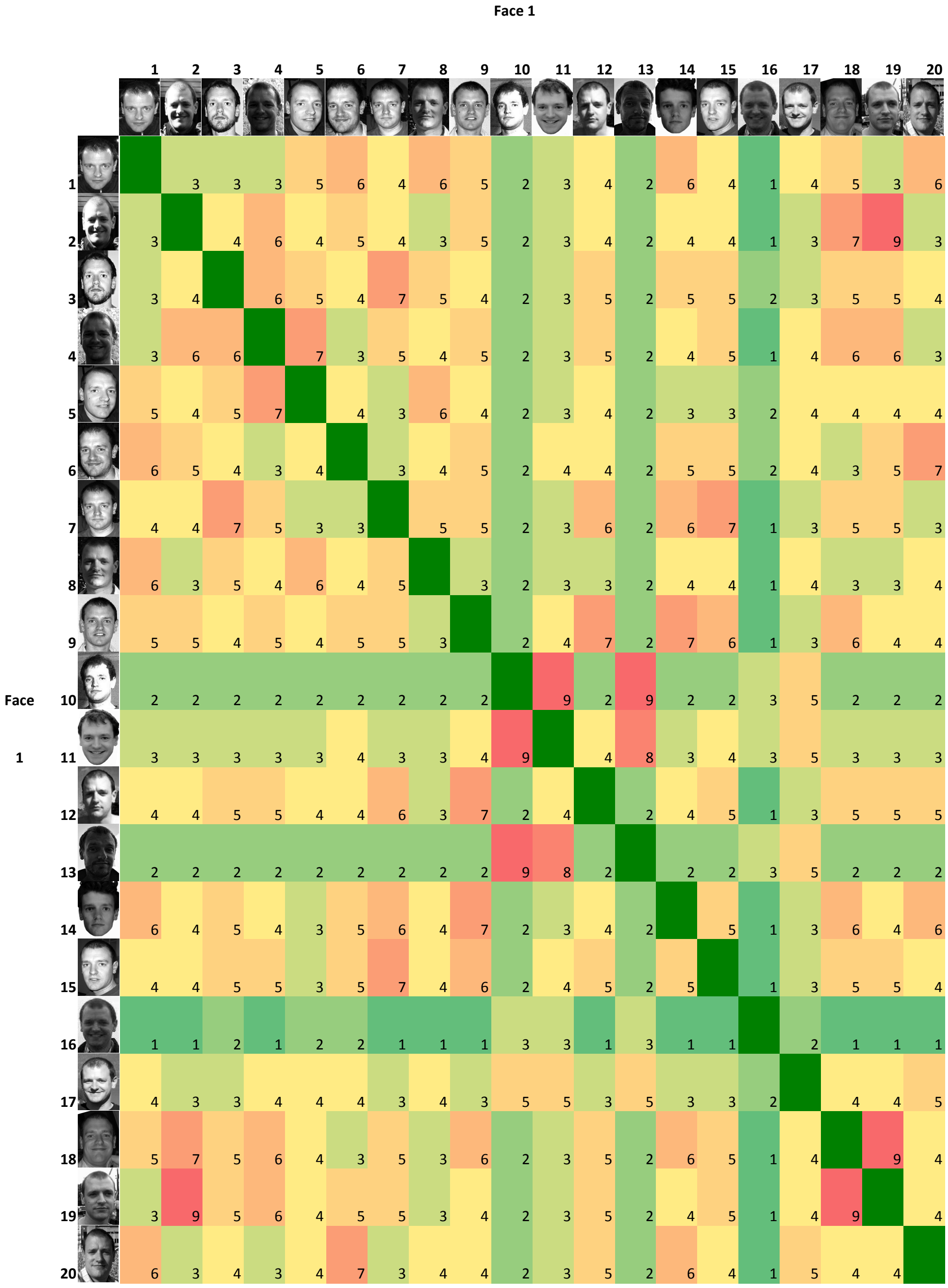


Figure 3.12. **Experiment 6: Number of perceived identities.**
Histograms illustrating the range and frequency of identities made by participants in Experiment 6, given either image set 1 (Face 1) or image set 2 (Face 2).

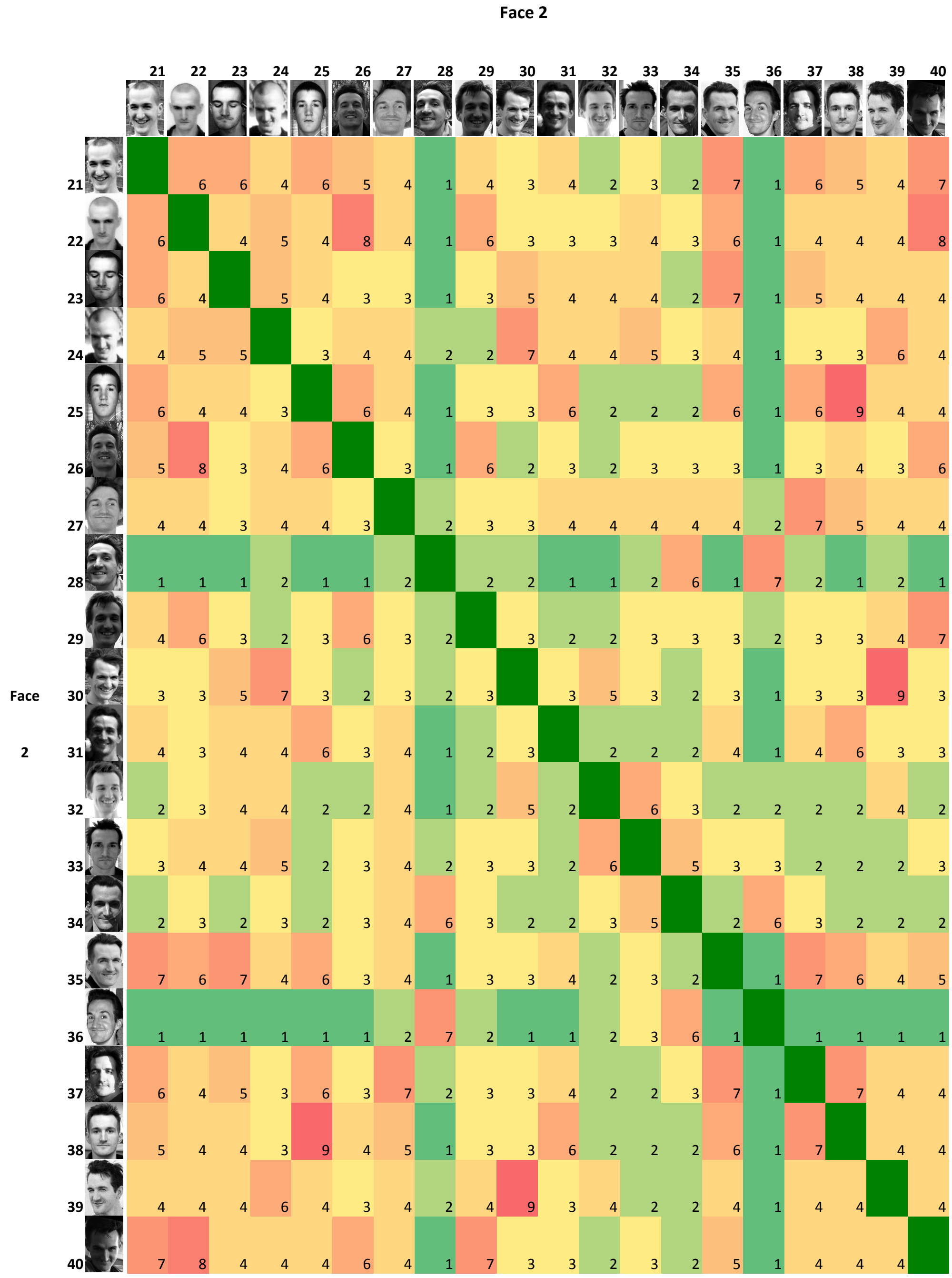
Confusion Matrix

Two matrices were created to illustrate the identification behaviour of participants presented with images of Face 1 or Face 2 separately (see Figure 3.13 and 3.14). Every time one image was grouped with another image it was recorded to observe identification patterns amongst the images. As two groups of ten participants identified images of each face, the number 10 in a matrix cell indicates that data from 10 participants was used to create the matrix. The colour scale green (low) to red (high) was again applied; therefore perfect performance would be represented with an all red matrix.



Face 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		3	3	3	5	6	4	6	5	2	3	4	2	6	4	1	4	5	3	6
2	3		4	6	4	5	4	3	5	2	3	4	2	4	4	1	3	7	9	3
3	3	4		6	5	4	7	5	4	2	3	5	2	5	5	2	3	5	5	4
4	3	6	6		7	3	5	4	5	2	3	5	2	4	5	1	4	6	6	3
5	5	4	5	7		4	3	6	4	2	3	4	2	3	3	2	4	4	4	4
6	6	5	4	3	4		3	4	5	2	4	4	2	5	5	2	4	3	5	7
7	4	4	7	5	3	3		5	5	2	3	6	2	6	7	1	3	5	5	3
8	6	3	5	4	6	4	5		3	2	3	3	2	4	4	1	4	3	3	4
9	5	5	4	5	4	5	5	3		2	4	7	2	7	6	1	3	6	4	4
10	2	2	2	2	2	2	2	2	2		9	2	9	2	2	3	5	2	2	2
11	3	3	3	3	3	4	3	3	4	9		4	8	3	4	3	5	3	3	3
12	4	4	5	5	4	4	6	3	7	2	4		2	4	5	1	3	5	5	5
13	2	2	2	2	2	2	2	2	2	9	8	2		2	2	3	5	2	2	2
14	6	4	5	4	3	5	6	4	7	2	3	4	2		5	1	3	6	4	6
15	4	4	5	5	3	5	7	4	6	2	4	5	2	5		1	3	5	5	4
16	1	1	2	1	2	2	1	1	1	3	3	1	3	1	1		2	1	1	1
17	4	3	3	4	4	4	3	4	3	5	5	3	5	3	3	2		4	4	5
18	5	7	5	6	4	3	5	3	6	2	3	5	2	6	5	1	4		9	4
19	3	9	5	6	4	5	5	3	4	2	3	5	2	4	5	1	4	9		4
20	6	3	4	3	4	7	3	4	4	2	3	5	2	6	4	1	5	4	4	

Figure 3.13. **Experiment 6: Face 1 Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.



Face 2	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
21		6	6	4	6	5	4	1	4	3	4	2	3	2	7	1	6	5	4	7
22	6		4	5	4	8	4	1	6	3	3	3	4	3	6	1	4	4	4	8
23	6	4		5	4	3	3	1	3	5	4	4	4	2	7	1	5	4	4	4
24	4	5	5		3	4	4	2	2	7	4	4	5	3	4	1	3	3	6	4
25	6	4	4	3		6	4	1	3	3	6	2	2	2	6	1	6	9	4	4
26	5	8	3	4	6		3	1	6	2	3	2	3	3	3	1	3	4	3	6
27	4	4	3	4	4	3		2	3	3	4	4	4	4	4	2	7	5	4	4
28	1	1	1	2	1	1	2		2	2	1	1	2	6	1	7	2	1	2	1
29	4	6	3	2	3	6	3	2		3	2	2	3	3	3	2	3	3	4	7
30	3	3	5	7	3	2	3	2	3		3	5	3	2	3	1	3	3	9	3
31	4	3	4	4	6	3	4	1	2	3		2	2	2	4	1	4	6	3	3
32	2	3	4	4	2	2	4	1	2	5	2		6	3	2	2	2	2	4	2
33	3	4	4	5	2	3	4	2	3	3	2	6		5	3	3	2	2	2	3
34	2	3	2	3	2	3	4	6	3	2	2	3	5		2	6	3	2	2	2
35	7	6	7	4	6	3	4	1	3	3	4	2	3	2		1	7	6	4	5
36	1	1	1	1	1	1	2	7	2	1	1	2	3	6	1		1	1	1	1
37	6	4	5	3	6	3	7	2	3	3	4	2	2	3	7	1		7	4	4
38	5	4	4	3	9	4	5	1	3	3	6	2	2	2	6	1	7		4	4
39	4	4	4	6	4	3	4	2	4	9	3	4	2	2	4	1	4	4		4
40	7	8	4	4	4	6	4	1	7	3	3	2	3	2	5	1	4	4	4	

Figure 3.14. **Experiment 6: Face 2 Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.

From the matrices most notable perhaps is that there is a wide variability in how images are identified, represented by a range of shades of red-to-green cells. In both matrices there are very few red cells, indicating the frequency of identity-split errors. It is also clear to see which images caused identification problems across the majority of participants, as their cells are predominantly green, for example images 10, 13 and 16 of Face 1 and images 28 and 36 of Face 2. Shown in Figures 3.15 and 3.16 are examples of participants' frequent correct identifications and frequent identity-split errors.

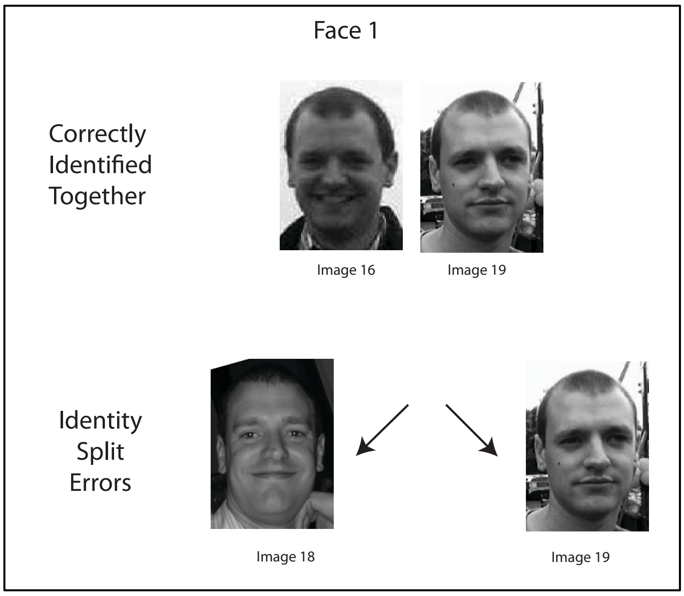


***Figure 3.15*. Experiment 6: Examples of identification behaviour for Face 1.**
Examples are shown of frequently correctly identified images in the top row and frequent identity split-errors in the bottom row (no identity-merge errors could be made).

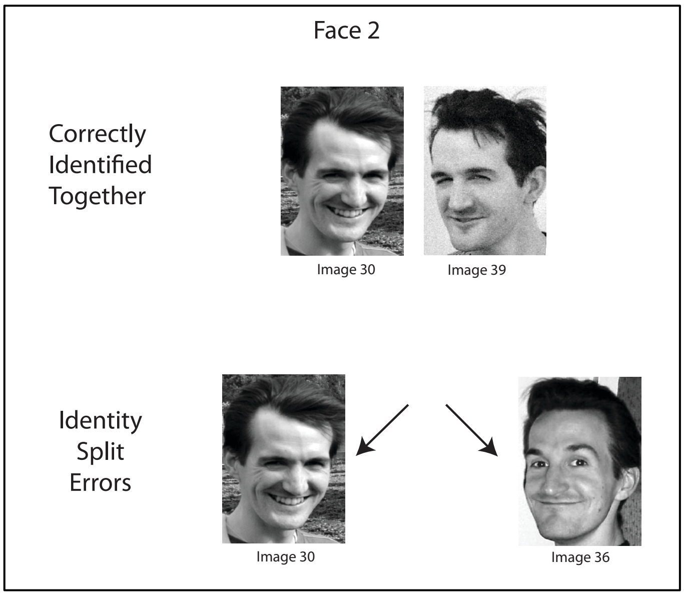


***Figure 3.16*. Experiment 6: Examples of identification behaviour for Face 2.**
Examples are shown of frequently correctly identified images in the top row and frequent identity-split errors in the bottom row (no identity-merge errors could be made).

Discussion

Following the manipulations of similarity in Experiments 4 and 5, a simpler method was proposed to measure whether the two identities shown in the images were causing participants poor performance: conduct an image-sorting task with multiple images of just one face. Participants were therefore asked to sort 20 images of either Face 1 or Face 2.

Without distractors of images of similar or dissimilar faces, participants were surprisingly bad at identifying multiple images of just one person. Participants made a mean number of 5.1 different identities from the 20 images of one face. This result clearly indicates that the effect of within-person variability on identification found in Experiments 1 – 5, also occurs with just one identity. This finding illustrates that the difficulties of identification across within-person variability are robust.

It could be suggested that this result is due to a task effect i.e. given the instructions to sort the images by identity could imply to participants that more than one identity exists. However, in defence of this suggestion it is clear that correct performance is possible as 2 of the 20 participants reached the correct outcome of one identity. Furthermore participants often followed up the basic instruction of “sort these images by identity or person”, by asking if there could be any number of identities present. In reply, participants were informed that this was correct and that they could make as many or as few identities as they saw fit. Therefore this lends support to the result not being caused by a task effect, but rather participants often genuinely struggled to identify the images accurately.

3.5 Experiment 7: Free sorting with Multi-PIE

In Experiment 6 it was shown that participants struggled to accurately identify multiple images of just one face. In Experiment 7, participants performance with multiple images of one person is examined using images taken from the Multi-PIE database created by Gross, Matthews, Cohn, Kanade and Baker (2010). The database was created to encompass within-person variability across the dimensions: pose, illumination and expression, with the images taken across four sessions over a period of 6 months. Developing on from studies which use one or two images of each face, controlled by factors such as expression and pose (e.g. Bruce 1982), the Multi-PIE database aims to successfully represent within-person variability by taken 100+ images of each face used. By using images from the database in a free sorting task it can be explored whether the images do capture realistic within-person variability by comparing participants' performance to participants' performance in Experiment 6.

Previous experiments that have taken multiple images of the same face, typically two images, capture the images across a very short period of time (e.g. Burton et al 1999). Similarly the images taken for the Multi-PIE database were captured over a limited period of time, up to six months. The images captured in the database are therefore unlikely to meet the same variability as the images of Face 1 and 2, used in Experiment 6. From this, and given the controlled conditions the images were captured in, it can be hypothesised that participants will perform better at identification of the faces from the Multi-PIE database.

Method

Participants

Twenty undergraduate students recruited from the School of Psychology subject pool participated in Experiment 7 in return for a small payment (5 male: 15 female, age range 18 - 24).

Design and Stimuli

Two faces were selected at random from the Multi-PIE database (Gross et al., 2010). From a large selection of images, taken over four sessions and across different poses, illuminations and expressions (PIE), 20 images of each face were selected at random for the stimuli. The faces will be referred to as Face 5 and Face 6. The images were cropped to exclude the image background as all images had been taken in the same location. The images were resized to the standard passport photo size (45mm x 35mm), printed in black and white, and laminated to create two sets of 20 images for participants to sort (see Figure 3.17). Using the same free sorting design as in previous experiments, the stimuli was counterbalanced so that participants were presented with either image set 1 or image set 2.

As in Experiment 5 and 6, none of the participants were familiar with the faces used in the stimuli. This was established following the task, by asking participants to indicate with a "Yes/No" response whether they recognised or were familiar with the face shown to them. All participants indicated that they were not familiar with the face in their image set.

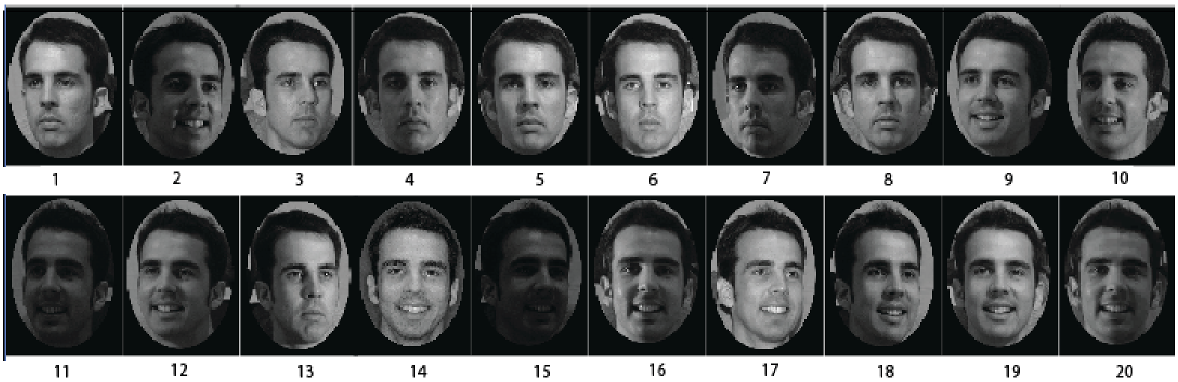


Image Set 1: 20 images of Face 5 (numbered 1 - 20).

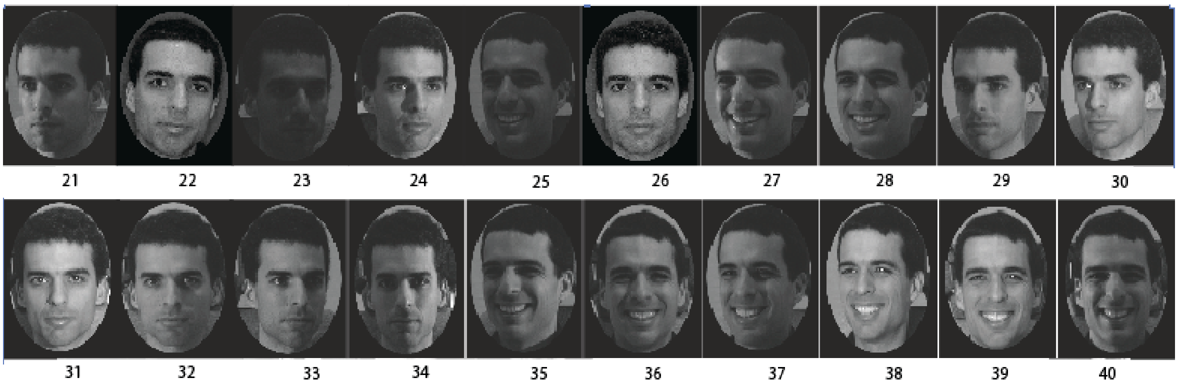


Image Set 2: 20 images of Face 6 (numbered 21 - 40).

***Figure 3.17.* Images used in Experiment 7.**
Two image sets were used in Experiment 7: Participants were given either set 1, twenty images of Face 5 shown in the top two rows or set 2, twenty images or Face 6 shown in the bottom two rows.

Procedure

Participants were presented with either image set 1 or set 2 and asked to: "sort the images by identity so that images of the same face, or person, are grouped together". Participants were also informed that there was no time restriction, although their trial time would be recorded, and that they were free to create as many or as few identities as they saw fit. Participants' outcome was recorded in addition to how the images were categorised.

Results

Overall participants sorted the images into a mean number of 3.3 identities (median 3; mode 2 and 5; range 1-6). The mean number of identities made by participants is significantly higher than the number of identities actually present [t (19) = 3.44, $p < 0.01$]. This indicates that participants struggled to identify images of faces taken from the Multi-PIE database (Gross et al., 2010). However three out of the 20 participants reached the correct outcome of one identity. This suggests that it is possible to accurately identify one person from multiple images taken from the Multi-PIE database.

From the images used in Experiments 6 and 7 (see Figures 3.11 and 3.17) it is clear that the within-person variability captured in Face 5 and Face 6 is much more controlled than the within-person variability captured in the ambient images of Face 1 and Face 2. To test whether the Multi-PIE images adequately capture within-person variability, participants' performance in Experiment 6 and 7 were compared. In Experiment 6 participants made a mean number of 5.1 identities and in Experiment 7 participants made a mean number of 3.3 identities. This difference is significantly different with participants in Experiment 6 creating more identities than participants in Experiment 7 [t (38) = 2.55, $p < 0.05$]. This difference indicates that the images in the two experiments are different; from these results it could be suggested that the images from the Multi-PIE database do not adequately represent within-person variability in the same way that the ambient images in Experiment 6. This results in significantly less identity-split errors made by participants in Experiment 7.

To establish whether poor performance was caused by one face more than the other, performance for each face was considered separately (see Figure 3.18). Participants who were given images of Face 5 sorted the images into a mean number of 3.4 identities (median 3.5; mode 5; range 1-6). Participants who were given images of Face 6 sorted the images into a mean number of 3.2 identities (median 3; mode 2 and 3; range 1-6). No significant difference was found between the mean number of identities produced for Face 5 and Face 6 [t (18) = 0.76, $p > 0.05$] demonstrating that identifying multiple images of the same face is difficult across different faces.

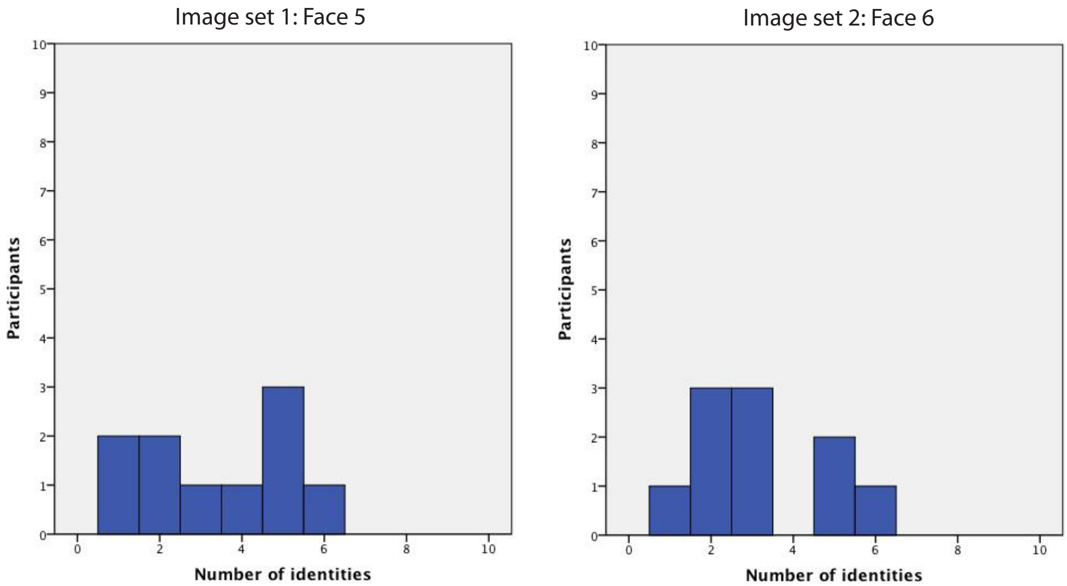
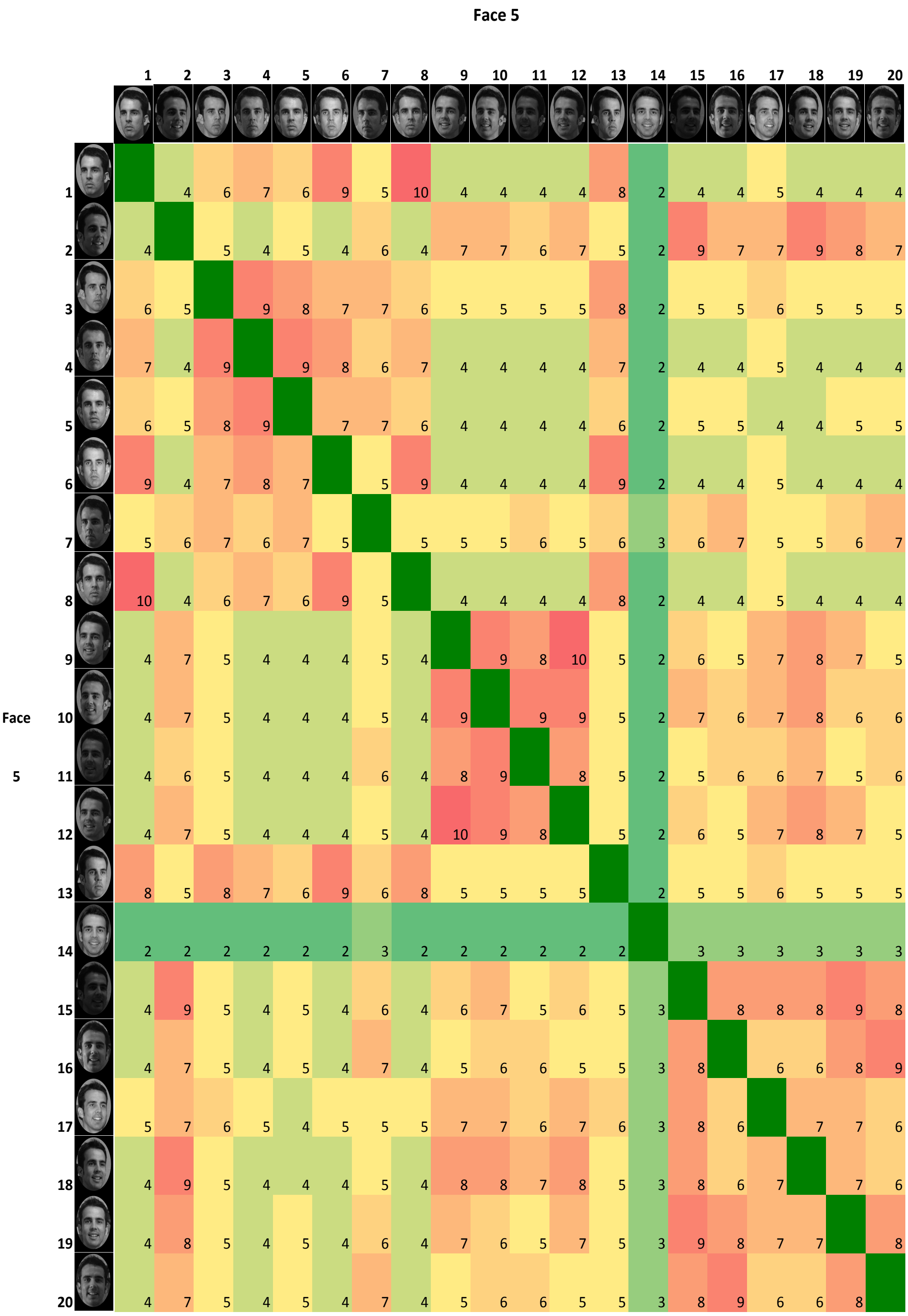


Figure 3.18. **Experiment 7: Number of perceived identities.**
Histograms illustrating the range and frequency of identities made by participants in Experiment 7, given either image set 1 (Face 5) or image set 2 (Face 6).

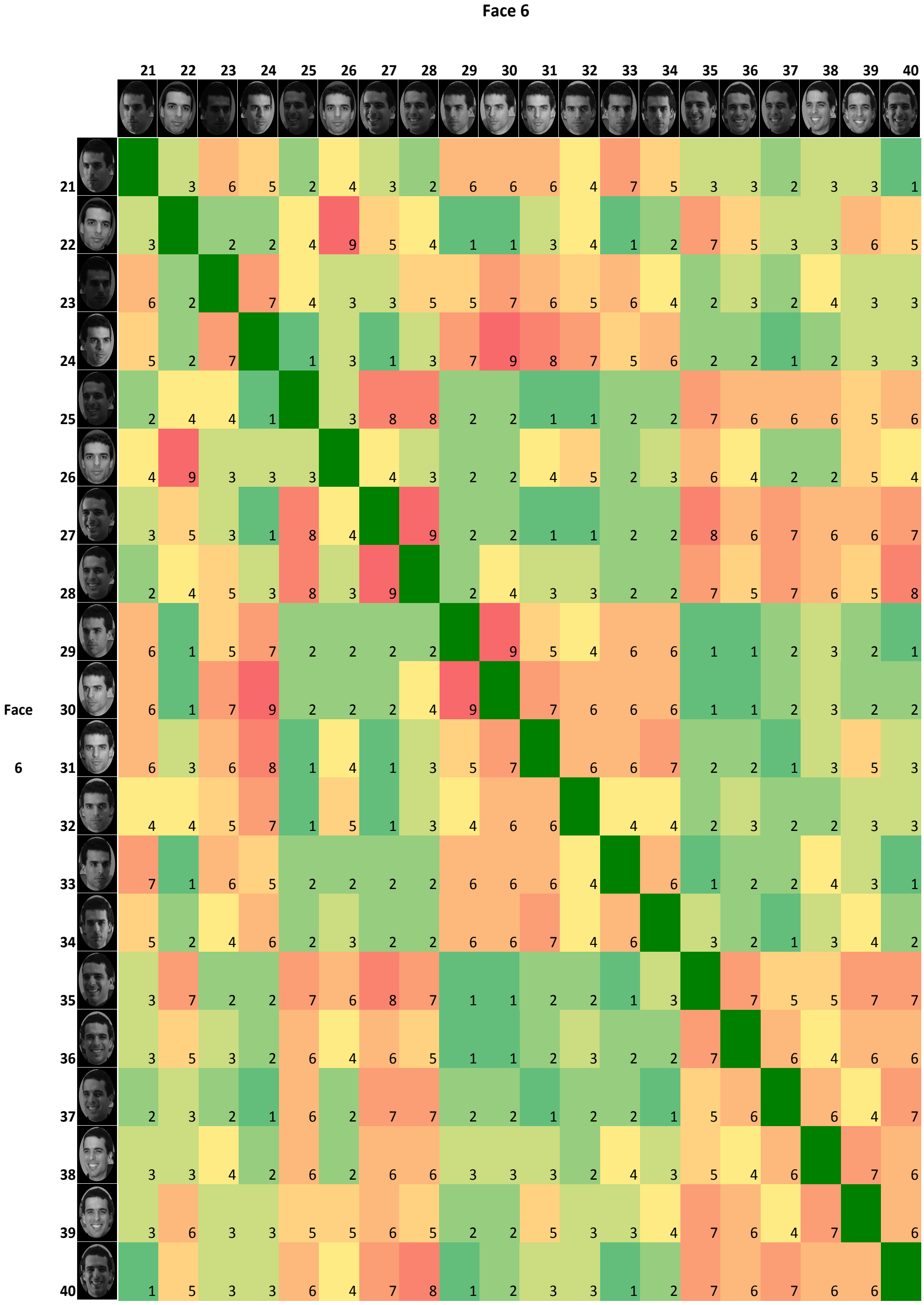
Confusion Matrix

Two matrices were created to illustrate the identification behaviour of participants presented with images of Face 5 or Face 6 separately (see Figure 3.19 and 3.20). Every time one image was grouped with another image it was recorded to observe identification patterns amongst the images. As two groups of 10 participants identified images of each face, the number 10 in a matrix cell indicates that data from 10 participants was used to create the matrix. The colour scale green (low) to red (high) was again applied therefore perfect performance would be represented with an all red matrix.



Face 5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		4	6	7	6	9	5	10	4	4	4	4	8	2	4	4	5	4	4	4
2	4		5	4	5	4	6	4	7	7	6	7	5	2	9	7	7	9	8	7
3	6	5		9	8	7	7	6	5	5	5	5	8	2	5	5	6	5	5	5
4	7	4	9		9	8	6	7	4	4	4	4	7	2	4	4	5	4	4	4
5	6	5	8	9		7	7	6	4	4	4	4	6	2	5	5	4	4	5	5
6	9	4	7	8	7		5	9	4	4	4	4	9	2	4	4	5	4	4	4
7	5	6	7	6	7	5		5	5	5	6	5	6	3	6	7	5	5	6	7
8	10	4	6	7	6	9	5		4	4	4	4	8	2	4	4	5	4	4	4
9	4	7	5	4	4	4	5	4		9	8	10	5	2	6	5	7	8	7	5
10	4	7	5	4	4	4	5	4	9		9	9	5	2	7	6	7	8	6	6
11	4	6	5	4	4	4	6	4	8	9		8	5	2	5	6	6	7	5	6
12	4	7	5	4	4	4	5	4	10	9	8		5	2	6	5	7	8	7	5
13	8	5	8	7	6	9	6	8	5	5	5	5		2	5	5	6	5	5	5
14	2	2	2	2	2	2	3	2	2	2	2	2	2		3	3	3	3	3	3
15	4	9	5	4	5	4	6	4	6	7	5	6	5	3		8	8	8	9	8
16	4	7	5	4	5	4	7	4	5	6	6	5	5	3	8		6	6	8	9
17	5	7	6	5	4	5	5	5	7	7	6	7	6	3	8	6		7	7	6
18	4	9	5	4	4	4	5	4	8	8	7	8	5	3	8	6	7		7	6
19	4	8	5	4	5	4	6	4	7	6	5	7	5	3	9	8	7	7		8
20	4	7	5	4	5	4	7	4	5	6	6	5	5	3	8	9	6	6	8	

Figure 3.19. **Experiment 7: Face 5 Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.



Face 6	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
21		3	6	5	2	4	3	2	6	6	6	4	7	5	3	3	2	3	3	1
22	3		2	2	4	9	5	4	1	1	3	4	1	2	7	5	3	3	6	5
23	6	2		7	4	3	3	5	5	7	6	5	6	4	2	3	2	4	3	3
24	5	2	7		1	3	1	3	7	9	8	7	5	6	2	2	1	2	3	3
25	2	4	4	1		3	8	8	2	2	1	1	2	2	7	6	6	6	5	6
26	4	9	3	3	3		4	3	2	2	4	5	2	3	6	4	2	2	5	4
27	3	5	3	1	8	4		9	2	2	1	1	2	2	8	6	7	6	6	7
28	2	4	5	3	8	3	9		2	4	3	3	2	2	7	5	7	6	5	8
29	6	1	5	7	2	2	2	2		9	5	4	6	6	1	1	2	3	2	1
30	6	1	7	9	2	2	2	4	9		7	6	6	6	1	1	2	3	2	2
31	6	3	6	8	1	4	1	3	5	7		6	6	7	2	2	1	3	5	3
32	4	4	5	7	1	5	1	3	4	6	6		4	4	2	3	2	2	3	3
33	7	1	6	5	2	2	2	2	6	6	6	4		6	1	2	2	4	3	1
34	5	2	4	6	2	3	2	2	6	6	7	4	6		3	2	1	3	4	2
35	3	7	2	2	7	6	8	7	1	1	2	2	1	3		7	5	5	7	7
36	3	5	3	2	6	4	6	5	1	1	2	3	2	2	7		6	4	6	6
37	2	3	2	1	6	2	7	7	2	2	1	2	2	1	5	6		6	4	7
38	3	3	4	2	6	2	6	6	3	3	3	2	4	3	5	4	6		7	6
39	3	6	3	3	5	5	6	5	2	2	5	3	3	4	7	6	4	7		6
40	1	5	3	3	6	4	7	8	1	2	3	3	1	2	7	6	7	6	6	

Figure 3.20. **Experiment 7: Face 6 Confusion Matrix.**
Every time participants categorised one image with another image it was recorded in the matrix to illustrate the identification behaviour of participants. Infrequent categorisation is represented at the green end of the colour scale, with frequent categorisation represented in reds.

From the two matrices it is clear that a wide range of variability in how images were identified; in both matrices there are few red cells, instead there is a range of colours reflecting the range of frequencies by which images were sorted together. Difficulties with certain images can be observed for example it is clear that image 14 of Face 5 was difficult for participants to group with other images. Overall there seems to be more of a 'block' pattern emerging from the two matrices; a block pattern occurs when a small group of images are infrequently sorted with another small group of images. For example for Face 6 images 29 to 34 images 35 to 37 are infrequently grouped together resulting in a green 'block' on the matrix. This blocking pattern is a clear illustration that identification behaviour is strongly led by variations in expression and pose. This suggests that problems in within-person identification can occur across variations as simple as a smiling expression versus a neutral expression. Shown in Figure 3.21 and 3.22 are examples of participants' frequent correct identifications and frequent identity split errors.

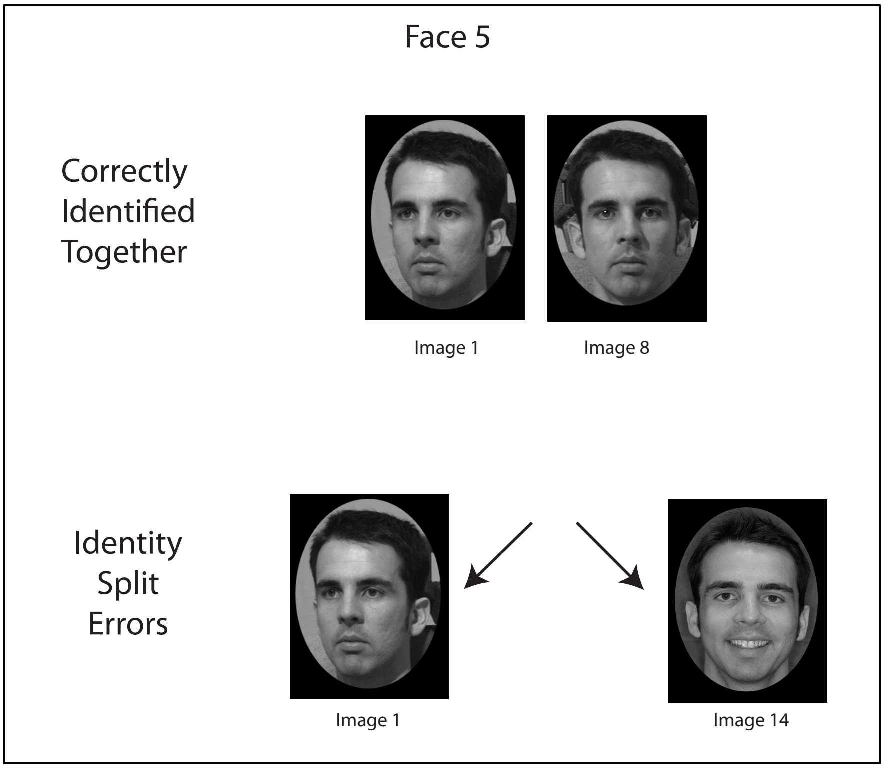


Figure 3.21. **Experiment 7: Examples of identification behaviour for Face 5.**
Examples are shown of frequently correctly identified images in the top row and frequent identity-split errors in the bottom row (no identity-merge errors could be made).

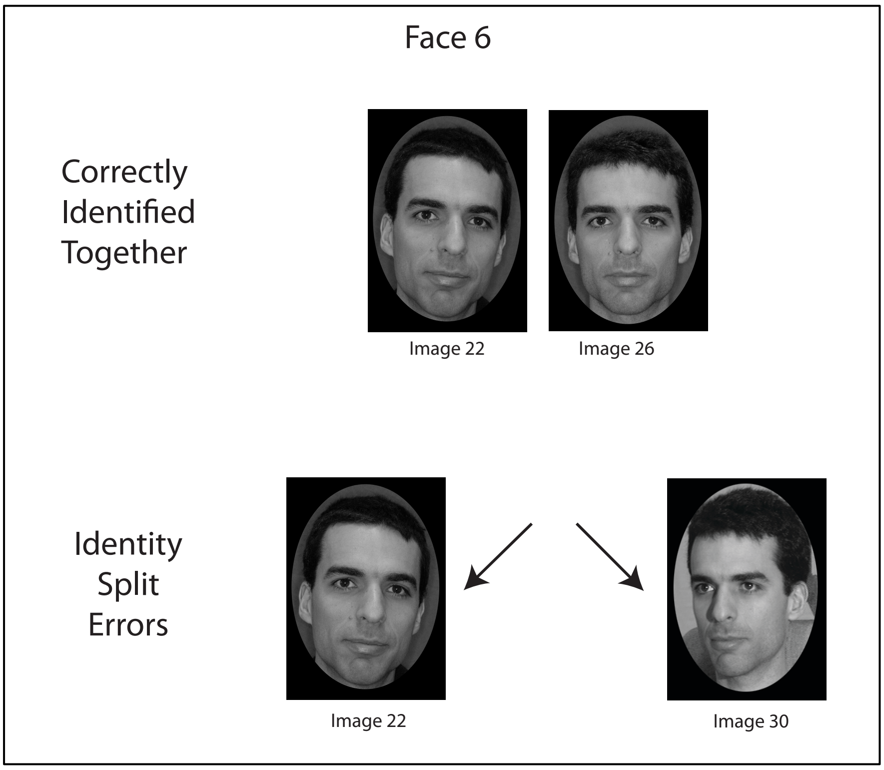


Figure 3.22. **Experiment 7: Examples of identification behaviour for Face 6.**
Examples are shown of frequently correctly identified images in the top row and frequent identity-split errors in the bottom row (no identity-merge errors could be made).

Discussion

In Experiment 7 participants made a mean number of 3.3 identities when asked to identify 20 images of just one face. This supports the findings of Experiment 6 and shows that the same identification difficulties occur across within-person variability of one face as they do across within-person variability of two faces.

These findings illustrate that the Multi-PIE database (Gross et al., 2010) does portray some within-person variability in images, however it could be argued that the images taken from the Multi-PIE database do not incorporate the full extent of within-person variability. In comparison to the results of Experiment 6, participants in Experiment 7 made significantly fewer identities. This suggests that it was easier for participants in Experiment 7 to identify across multiple images of one face. From this it could be concluded that the Multi-PIE images do not capture the same within-person variability as the images used in Experiment 6, therefore within-person variability extends beyond the dimensions controlled for in the Multi-PIE database; pose, illumination and expression.

In addition, comparing the images of Faces 5 and 6 to Faces 1 and 2 there is a clear difference in time; the ambient images of Faces 1 and 2 are captured over a longer time period than the images of Face 5 and 6. This suggests that time is a key factor in within-person variability. This point may seem obvious, however it poses a great challenge in research as to capture the levels of within-person variability that occurs in real-life, images need to be gathered over a longer period of time. To overcome this problem instead of striving to capture realistic within-person variability in laboratory settings, researchers need to explore other methods of gathering data, for example, taking advantage of the growing trend of documenting daily life on personal websites and online profiles. This would be an interesting future development; collecting image sets that capture a person's appearance over several years.

3.6 General Discussion

In Experiment 4 participants familiar with highly similar faces were able to accurately identify two identities amongst the 40 images, while participants unfamiliar with the same faces made a mean number of 6.1 identities from the 40 images. Despite reaching the correct number of identities, participants familiar with the faces made similar levels of identity-merge errors as participants unfamiliar with the faces. Therefore participants familiar with the highly similar faces were aware that the images should be identified into two faces but were unable to accurately identify each image.

In Experiment 5 participants were asked to identify images of two highly dissimilar faces. Two of the 22 participants reached the correct outcome and 15 of the 22 participants reached the correct outcome for one of the identities. Furthermore, comparing the results of Experiment 5 with the results of Experiments 1 and 4 found that identification performance had improved in Experiment 5 for Faces 2, 3 and 4. This suggests that increasing between-person variability (through using dissimilar faces) can help improve identification across within-person variability but only to a certain extent and not for all faces.

In Experiment 6 participants were asked to identify multiple images of just one face. Participants made a mean number of 5.1 identities, significantly higher than the one identity actually present. This finding illustrates that the within-person variability effect on identification found in the previous experiments is robust, and is not the result of presenting participants with many images of similar looking faces.

In Experiment 7 participants were again asked to identify images of just one face, this time faces from the Multi-PIE database were used (Gross et al., 2010). This enabled the findings of Experiment 6 to be extended across other faces and also allowed for the variability measured in the Multi-PIE database to be compared to the variability represented in the images used in Experiment 6. Participants made a mean number of 3.3 different identities from the images of just one face. In comparison to the previous image sorting experiments, this finding indicates that the variability captured in the Multi-PIE database is not as realistic a representation of within-person variability as, for example, the images of Face 1 and Face 2.

It could be suggested that poor performance in Experiments 4 – 7 could be a task effect i.e. reflecting the instructions given to participants. When asked to “sort the images by identity”, participants may have felt this implied that multiple identities existed amongst the images. This is an important point to consider, however, given that throughout the experiments a small number of participants were able to reached the correct outcome this would suggest that identification behaviour was not influenced by the task instructions. Furthermore, participants typically used visual information when identifying the images, for example differences in hairstyle or facial hair, and age variations, suggesting again that identification was not influenced by the instructions but led by the visual information available to participants. As proposed in Chapter 2, a simple way to ensure that task instructions are not influencing participants’ behaviour would be to ask participants following the task whether they felt the instructions influenced their final outcome. This feedback could then be used to improve the instructions if necessary.

Another limitation is that, as in Chapter 2, a maximum of only two faces were used as stimuli in Experiments 4 – 7, therefore generalising conclusions about identification based on these findings should be done so carefully as two faces are not an adequate representation of the wider population. To overcome this in future identification research, within-person variability should be considered across a larger number of faces. A possible future experiment could simply be based on the image sorting tasks laid out in Chapters 2 and 3, with the additional introduction of new faces.

The findings of this chapter extend the findings of Chapter 2 by demonstrating that within-person variability can cause problems in identification in different faces. Experiments 4 and 5 have shown that similarity of the faces used in image sorting tasks can have an influence on participants' performance with familiar and unfamiliar participants struggling to correctly identify highly similar faces while participants performance improves with highly dissimilar faces. Interestingly this improvement observed in Experiment 5 applies only to images of one of the faces presented to participants. This finding supports the concept outlined in Chapter 2 that awareness of within-person variability occurs face by face as opposed to being applicable across all faces. Experiments 6 and 7 further extend the findings of Chapter 2 by demonstrating the robust effect of within-person variability on identification in image sorting tasks with images of one face.

In addition to supporting the findings of Chapter 2, Experiments 4 – 7 develops the findings from Jenkins et al., (2011) to illustrate that within-person variability occurs across different faces and has similar effects on identification. Jenkins et al., (2011) demonstrated that when participants were familiar with a face, they were able to accurately identify across within-person variability. This supports previous research that reports high performance for identification of familiar faces (e.g. Clutterbuck & Johnston 2004; Bruce et al., 2001). In Experiment 4 this familiarity effect is examined with highly similar faces and it can be concluded that familiar participants were able to correctly identify two faces but often made mistakes in identifying one face as the other. This finding suggests that familiarity is beneficial in identifying across within-person variability to an extent but perhaps highly similar faces cause difficulties even when we are familiar with the people.

The findings of this chapter underline the need to consider realistic within-person variability. Participants in Experiments 6 and 7 were both asked to identify twenty images of one face, with the difference of the images participants were given. In Experiment 6 the ambient images of either face as used in Experiments 1 – 3 were presented to participants while in Experiment 7 participants were given images of faces taken from the Multi-PIE database created by Gross et al., (2010). The aim of the database was to capture a full representation of within-person variability across the controlled dimensions: pose, illumination and expression (PIE), to use in the advancement of face recognition algorithms.

It was proposed that if the database accurately captured realistic within-person variability, the results from Experiments 6 and 7 would be comparable. Participants in Experiment 7 did struggle to reach the correct number of one identity present in the images, however the mean number of identities produced by participants was significantly lower than the mean number of identities created by participants in Experiment 6. This indicates that the Multi-PIE database has achieved images of within-person variability to a limited extent and that realistic variability incorporates more than just variations in pose, illumination and expression.

This comparison of Experiments 6 and 7 clearly demonstrate that controlled variability across dimensions such as pose, lighting and facial expression are limited in their representation of natural within-person variability. Moreover, capturing the images over a limited time frame demonstrates the large impact of time on within-person variability. This is an important finding as it indicates that previous research that has used similarly controlled methods of exploring within-person variability are also limited (e.g. Bruce 1982; Bruce et al., 1999). Therefore the approach to using varying images is somewhat flawed in previous research and images of more natural within-person variability are needed.

In addition to illustrating the robustness and extent of within-person variability, Chapters 2 and 3 have clearly shown that one image is not an adequate representation of a face. From Experiment 7 it is also shown that images taken under controlled experimental settings are not an adequate representation of a face. To explore realistic within-person variability there needs to be a change in the use of images in face identification research. For example, instead of attempting to collect images that capture variability within the time frame of a lab-based experiment, researchers should utilise other sources of images such as online personal profiles. With the involvement of volunteers willing to contribute a range of their own personal images, stimuli can develop to naturally capture within-person variability. This would be an exciting development in future face processing research.

Overall it is clear that within-person variability is greater than previously represented in face recognition research. It can also be concluded that image-sorting tasks are a simple concept that can provide detailed insight into difficulties in identification across within-person variability. Given the large and robust effect of within-person variability in face identification, the question arises of whether within-person variability can affect other areas of face processing and to what extent. To address this, within-person variability will now be explored in the perception of personality from faces.

Chapter Four

Within-Person Variability in Social Dimensions

4.1 Experiment 8: Within-person variability in social dimension

Introduction

In Chapters 2 and 3 within-person variability is shown to have a big effect on face identification. Identity is not the only signal that we can perceive in a face; facial appearance also plays a significant role in the perceptions we make about a person's character. It is therefore important that we explore whether within-person variability affects perceived facial perceptions.

Research has shown that people can quickly form impressions of a person's personality from their facial appearance (Todorov, Mandisodza, Goren & Hall 2005; Willis & Todorov 2006). For example in a study from Willis and Todorov (2006) it was found that first impressions made after just a 100-millisecond exposure to a face were highly correlated with impressions made without a time constraint. There is little evidence to suggest these perceptions are an accurate impression of the person, however these impressions are reliable and consistent across perceivers (Berry & Finch Wero 1993; Todorov, Said, Engell & Oosterhof 2008). Furthermore these perceptions can influence social outcomes such as our behaviour or judgments towards the perceived person (Oosterhof & Todorov 2008; Willis & Todorov 2006).

For example perceptions made based on facial appearance have been shown to relate to voting behaviour in political elections (Ballew and Todorov 2007; Antonakis & Dalgas 2009). In a study from Todorov, Mandisodza, Goren and Hall (2005) participants were asked to rate images of unfamiliar candidates from previous US Senate elections for the trait competence. Overall 71.6% of participants rated the successful candidate as more competent than the losing candidate. Perceptions of dominance have also been shown to predict success in military careers in a study from Mueller and Mazur (1996). Participants were asked to rate graduation photographs of army cadets for the trait dominance. Results showed that dominance predicted military rank 20 years later, with higher dominance resulting in higher promotions.

Face perception research has also identified personality traits that can reliably be perceived in facial appearance. In a study from Oosterhof and Todorov (2008) two independent dimensions, trustworthiness and dominance, were identified as being adequate to characterise face evaluation. Oosterhof and Todorov (2008) reached this conclusion by conducting principal components analysis (PCA) of neutral expression faces that had been rated for several trait dimensions. The first two principle components accounted for the majority of variance in the faces. Based on the trait ratings, Oosterhof and Todorov interpreted these principle components as trustworthiness and dominance. They then illustrated that perceptions such as threat can be expressed through manipulation of faces in a 2D face space along the dimensions trustworthiness and dominance (2008). This finding of the prominence trustworthiness and dominance has in facial perception supports research that shows the importance of trustworthiness and dominance in social perception (e.g. Wiggins 1979).

In contrast research has also found the personality traits warmth and competence to be universally reliable social dimensions (Fiske, Cuddy & Glick 2007; Judd, James-Hawkins, Yzerbyt & Kashima 2005). The perceptions of both traits are thought to be developed on an evolutionary basis: The dimension warmth being perceived from the judgment of whether a new person or group of people are likely to harm you, and the dimension competence being perceived as the judgment of whether the person or group of people have the ability to carry out harmful intentions (Fiske et al., 2007). Warmth and competence have been shown to be robust social judgments that can be traced back to fundamental perceptions of personality as presented by Asch (1946). Some facial perception research has investigated the effect of warmth and career success (Livingston & Pearce 2009) however the majority of research explores the perceptions of warmth and competence in individual and group behaviour as opposed to facial perceptions.

Given the effect facial perceptions can have on judgment behaviour and decision-making, it is important to have a complete understanding of face perception. As shown in identification tasks in Chapters 2 and 3, multiple images capturing within-person variability can have a significant effect on identification behaviour. Therefore one image is not an adequate representation of a person's face; an image offers only a limited view of a face.

Stimuli used in face perception research typically fall into one of two categories: using one image to represent a person or face, or using a computer-generated image to represent a face. A person's face can vary across different measures in the short and long term. Therefore to effectively represent a face multiple images that encompass within-person variability must be used. Research using one image to represent a face often explores between-person variability as opposed to variability that can occur within one person's face.

Research that uses computer-generated or computer manipulated faces similarly overlooks the concept of within-person variability. For example in the previously discussed study from Oosterhof and Todorov (2008) computer-generated 2D faces were used, that could be varied in a multidimensional face space for dimensions such as trustworthiness. By manipulating the trustworthiness of a face, a new face could be created. This technique allows for an unlimited source of faces to be created however the variability this technique addresses is very controlled. It does begin to address the facial features that are key to varying facial appearance for social dimensions but only between-persons. As in Experiments 1 – 7, ambient images of faces need to be obtained to gain an understanding of realistic within-person variability in face perception.

In particular the introduction of within-person variability can explore whether perceptions of social dimensions in faces are constant. For example does a face have one consistent level of trustworthiness or does the perception of trustworthiness vary dependent on the image of a face. If the perception of social dimensions in faces is based on an underlying facial structure, such as placement of facial features, it could be expected that within-person variability will not have an effect on the perception of a face. However if perceptions are based on variability such as expression changes then it could be expected that perceptions would vary dependent on an image of a person.

In this chapter within-person variability in the social dimensions trustworthiness and dominance are examined. The specific selection of the perceptions of the traits trustworthiness and dominance is based on previous research into the most reliably perceived traits from facial appearance (Oosterhof & Todorov 2008). Previous research has shown in depth that personality dimensions can be perceived through facial appearance. However previous research often uses just one image to represent a face or digitally manipulates face images to investigate the extent to which personality is perceived facially. This may not give an accurate representation of a face. Here, participants rated multiple images of 80 different faces for the traits trustworthiness and dominance.

The effect of familiarity will also be explored in this chapter, as participants were either familiar or unfamiliar with the faces rated. To achieve this familiarity divide with all faces, the experiment was conducted in two locations: the UK and Australia, and the famous people used in the stimuli were gathered from groups of celebrities well known solely in the UK or Australia. The images were therefore constant in both conditions, alternatively making up the familiar or unfamiliar faces.

Familiarity may effect personality perceptions as impressions may already have been formed for familiar faces. Impressions may have been formed from having previously seen images of the person or seeing them in their job role such as a television presenter. If the previously formed impressions influence personality perceptions it is hypothesised that significant differences will occur between ratings of familiar and unfamiliar faces (in either direction) dependent on the identity. In addition familiarity may affect the variability of personality perceptions. Again, if impressions have already been formed about a person, different images of the same person may be rated as similar for trustworthiness and dominance. Therefore significantly less variability will occur within-person for images of familiar faces.

It is hypothesised that ratings of trustworthiness and dominance will show a wide range of within-person variability. It is also hypothesised that this wide range of variability will be evident in face overlapping between-persons. Finally, it is hypothesised that familiar ratings will be significantly different and significantly less varied than unfamiliar ratings across both traits.

Method

Participants

Forty participants took part in this experiment in return for a small payment: 20 Australian undergraduate students from the University of New South Wales and 20 British undergraduate students from the University of Glasgow (age: 18 – 45 years; 12 males: 28 females).

Design and Stimuli

The stimuli comprised of 12 images of 80 different faces (see Figure 4.1 for an example of 12 images of one face). The faces comprised of 40 UK famous people and 40 Australian famous people (see Table A.1, Appendix A for list of names), therefore by keeping the images constant and conducting the experiment in the UK and Australia, ratings for both familiar and unfamiliar faces could be examined.

The images were collected through an Internet image search. The criteria for the images included locating high-quality images, displaying the full face without any obscuring objects, and without any other visible figures or faces. The images were then edited to show only the face (see Figure 4.1 as an example), and were uniformly sized to 250 x 375 pixels.

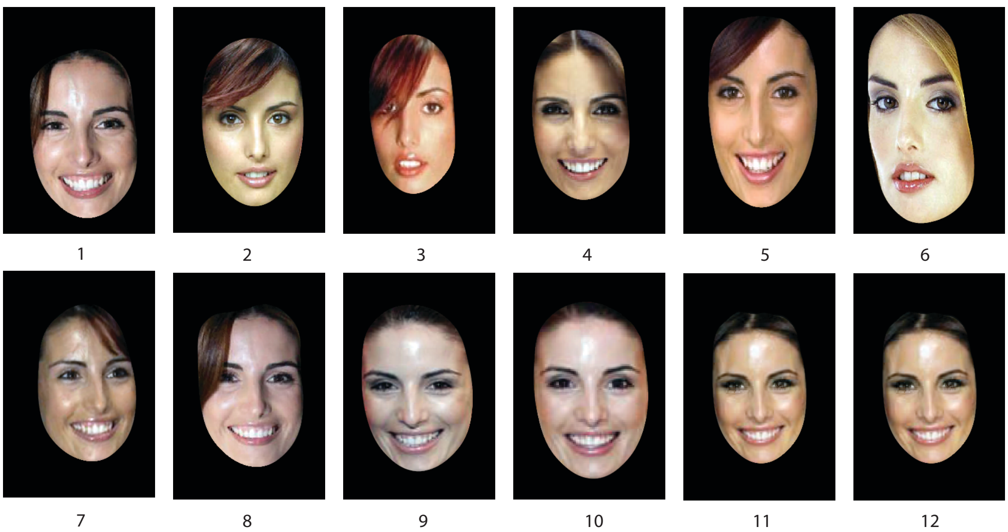


Figure 4.1. **Example of images used in Experiment 8.**
Illustrating the typical variability illustrated in the 12 images of each face shown in Experiment 8.

The images were then inserted into a computer task using experiment-building software, LiveCode. The task was programmed to repeat the instructions given to participants in text form, with a “Start Experiment” button available at the bottom of the screen for participants to press when ready to begin the task. Following this, the images would appear individually on the screen, sized 6 x 9 cms, above a seven-button scale of 1 – 7. To move onto the next image, participants were required to press a button on the scale.

Participants were allocated into either condition A and B: 20 participants in condition A viewed images 1-6; 20 participants in condition B viewed images 7-12. Presenting each participant with 6 images as oppose to the full set of 12 images, allowed for repetitiveness and participants’ fatigue to be reduced.

Procedure

Participants were informed that the experiment involved rating faces for the personality traits trustworthiness and dominance, and the definition of each trait was clarified (see Figure A.5 in Appendix A for trait definitions). It was explained that the images would appear on the computer screen above a button scale of 1 – 7. Participants were asked to indicate how trustworthy or dominant they perceived the image shown on the screen to be on a seven-point scale (1 = not at all and 7 = very). Following their response, the image would leave the screen and another image would appear. Participants were asked to ensure they were rating each image individually, as opposed to each identity.

Participants were made aware that there was no time limit and to move onto the next image they were required to make a rating. At three equal intervals throughout the task participants were presented with ‘break’ screens allowing for them to take an optional break before continuing with the task.

Familiarity was a key element of this experiment and so images of people famous in two different locations were used, and the experiment conducted in these two locations: the UK (at the University of Glasgow) and in Australia (at the University of New South Wales). To ensure that the famous faces selected were actually familiar or unfamiliar to the right participants, a short additional task was given to participants at the end of the experiment.

Participants were shown an image of each face again along with the name of the famous person, on the computer screen, and were asked to indicate with a “Yes/No” response whether they were familiar with each person, by clicking on either a “Yes” or “No” button below the image. Participants’ responses could then be analysed and if any of the appointed familiar faces were “unfamiliar” or any of the appointed unfamiliar faces were “familiar”, data could then be excluded. All participants were familiar with the appointed familiar faces and unfamiliar with the appointed unfamiliar faces in their locations, reflecting that the stimuli adequately captured images of familiar and unfamiliar faces.

Results

To visualise variability within images of each face scatterplots were created. Firstly all trustworthy and dominant ratings from all participants were averaged to create a trustworthy rating and a dominance rating for each image. The mean ratings were then plotted by identity in a scatterplot with trustworthiness ratings along the x-axis and dominance ratings along the y-axis. To give an example of the variability found, Figure 4.2 below shows a selection of faces (see Figures A.1, A.2, A.3 and A.4, Appendix A for scatterplots of all faces).

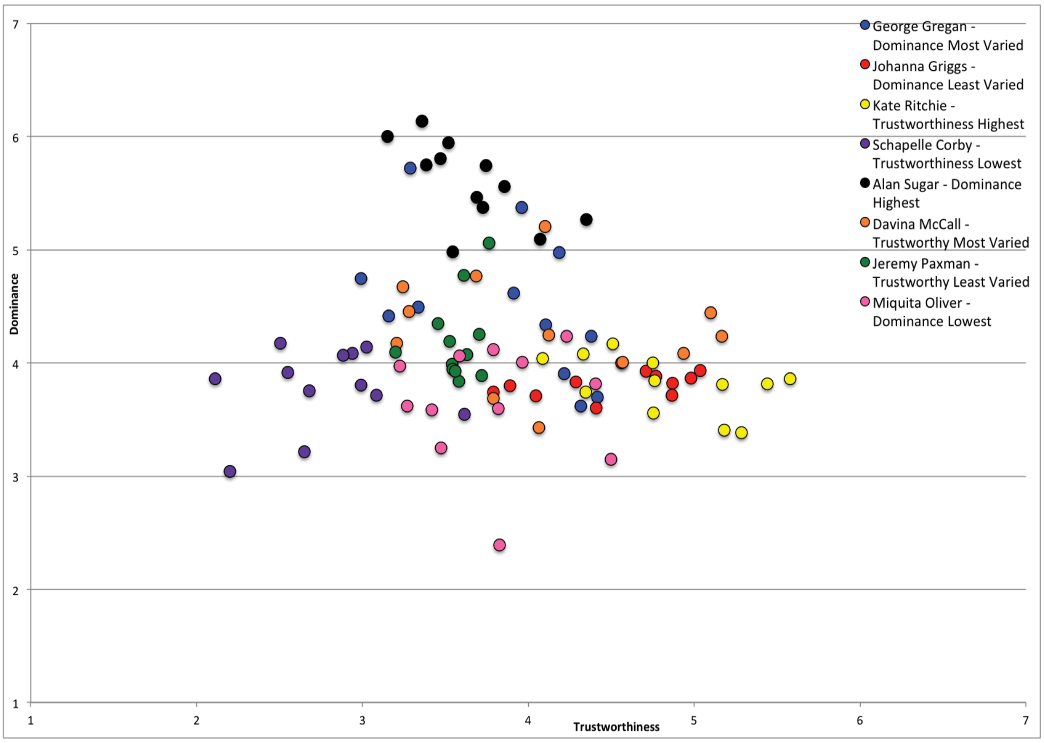


Figure 4.2. **Experiment 8: Examples of variability in Trustworthy and Dominance ratings.** A scatterplot showing faces that illustrate the highest and lowest rated face for each dimension as well as the most and least varied face for each dimension, with trustworthiness ratings along the x-axis and dominance ratings along the y-axis.

Each group of coloured data points represents a face, with each data point representing each of the 12 images rated by participants. By visualising the data in this form the wide variability of ratings within-person is shown: variability of ratings within-person is so wide that it causes overlapping between-persons. It is also possible to visualise the images represented by the data points to see the within-person variability that occurs between the images. For example, in Figure 4.3 the face rated the most variability for dominance is shown with the images rated most and least dominant.

Figure 4.3. **Experiment 8: Variability in Dominance.**
Two images of the same face (Face 10) illustrating the widest range of variability for ratings of dominance.

A further way of visualising the image ratings is to plot the images as data points in the same scatterplot showing the trustworthy and dominance ratings for a face. For example in Figure 4.4 all images of Face 10 plotted by the mean trustworthy and dominance ratings given to each image. Visualising the images in this way allows for observations to be made as to why images were rated differently, and why images were rated low or high for trustworthiness or dominance. For example the images rated highest for trustworthiness all show smiling expressions while the images rated lowest for trustworthiness show head poses and eye gazes looking down. Furthermore the images rated highest for trustworthiness are rated lowest for dominance suggesting that a smiling expression indicates high trustworthiness but low dominance.

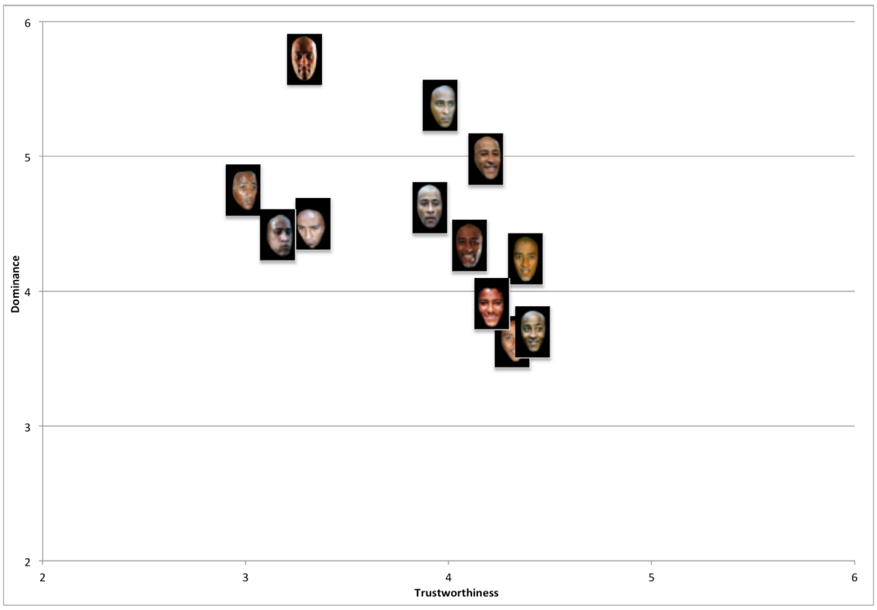
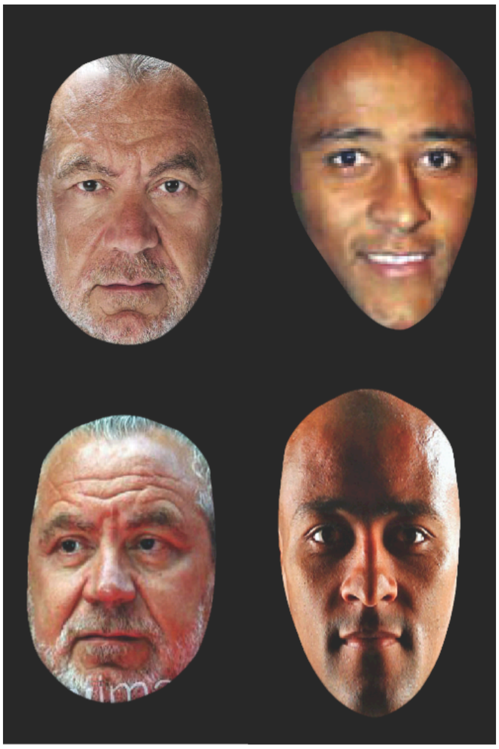


Figure 4.4. **Experiment 8: Variability of one face.**
Scatterplot of trustworthiness and dominance ratings given to images of Face 10, with data points represented by the appropriate image.

Overlapping Between-Person

A wide range of between-person variability also occurs with this data as there is a lot of overlap between different images of different faces. Overlapping occurs when a rating of one face falls between the minimum and maximum rating of another face. This is a intriguing finding as it demonstrates that faces can be shown as more or less trustworthy or dominant than each other dependent on the image used to represent each face. For example shown in Figure 4.5 are images of two faces (Face 71 and Face 38) that illustrate an image reversal for trustworthiness, and images of two faces (Face 41 and Face 10) that illustrate an image reversal for dominance. Therefore the face that is perceived as most trustworthy or dominant depends on the image of the face shown. This demonstrates that facial perceptions for traits such as trustworthiness and dominance are not stable, therefore are not determined solely by anatomical invariants, such as bone structure, but also by more transient aspects of the image.

Trustworthiness



Dominance

***Figure 4.5*. Experiment 8: Reversal of perceptions.**
Examples of reversal in perceptions of trustworthiness and dominance: The images on the left are of one face and the images on the right are of another face. The image on the left on the top row is rated as higher than the image on the right, while the image on the right of the bottom row is rated lower than the image on the left.

To illustrate the frequency with which trustworthiness and dominance ratings for images overlap between-person the range of trustworthy and dominant ratings for each face was calculated and then compared with the range of ratings for every other face. Every occurrence of one face's range of ratings overlapping with another face's range of ratings could then be indicate in a matrix. This was calculated for both trustworthiness and dominance separately (see Figures 4.6 and 4.7) and a colour scale was then applied so that green cells indicated when two faces overlap, and red cells indicates no overlapping.

Almost all of the faces overlap: 85% of faces trustworthiness ratings overlap and 75% of faces dominance ratings overlap. From the matrices it is possible to make different conclusions about the variability of ratings for images. For example comparing the two matrices, less overlap occurs for dominance (75% overlap) than trustworthiness (85% overlap). This suggests that there is more between-person variability for the trait dominance than trustworthiness. It is also possible to see the faces that are rated in the extreme of trustworthiness or dominance as they are infrequently overlapped by other faces (represented in red matrix cells). This occurs predominantly for dominance suggesting that some faces are more likely to be rated as particularly high or low for dominance than trustworthiness.

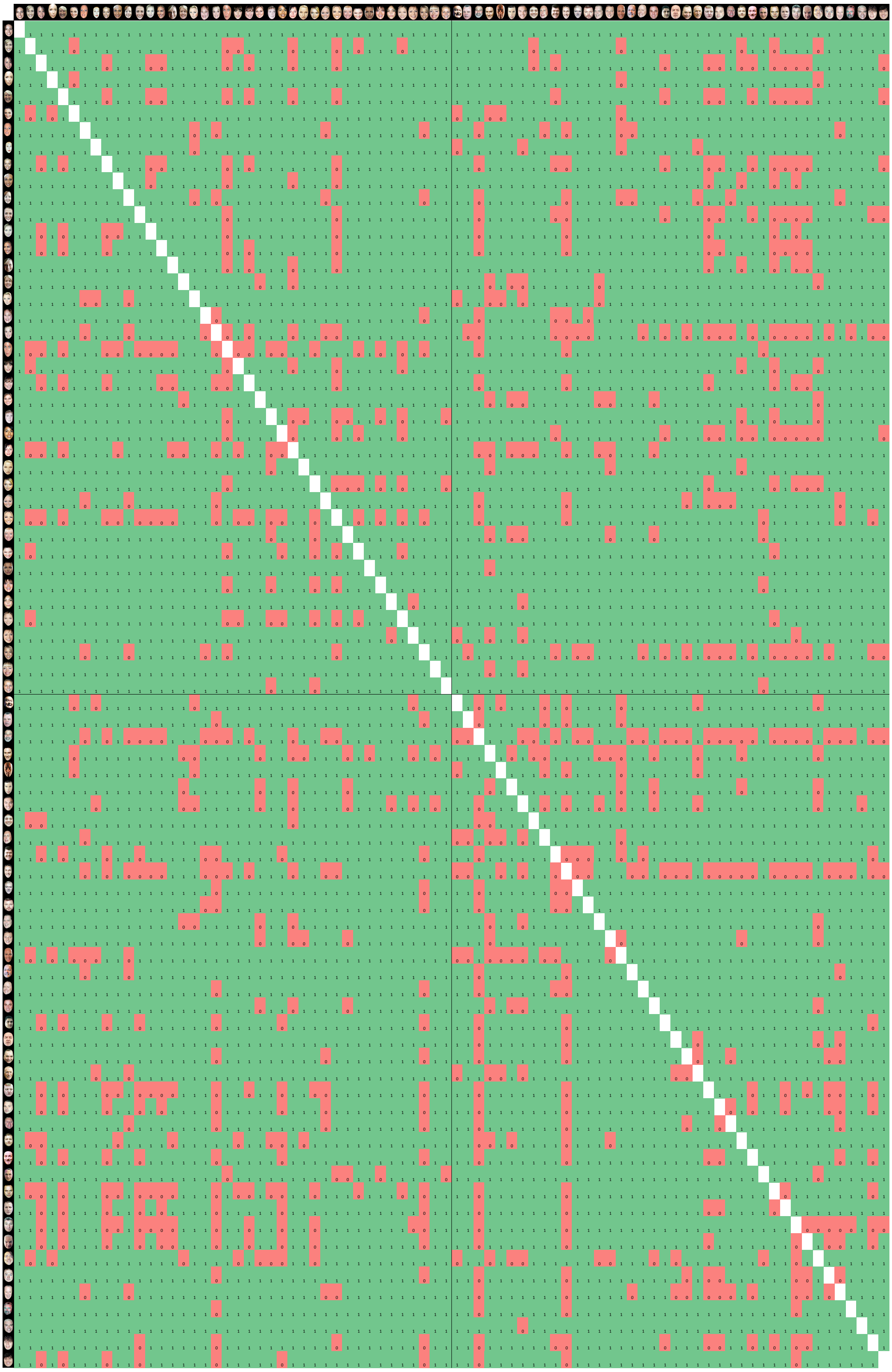


Figure 4.6. **Experiment 8: Trustworthy Overlap Matrix.**
Matrix illustrating every time the trustworthiness ratings for one face overlaps with trustworthiness ratings for another face. Faces that overlap are shown in green, and faces that don’t overlap are shown in red.

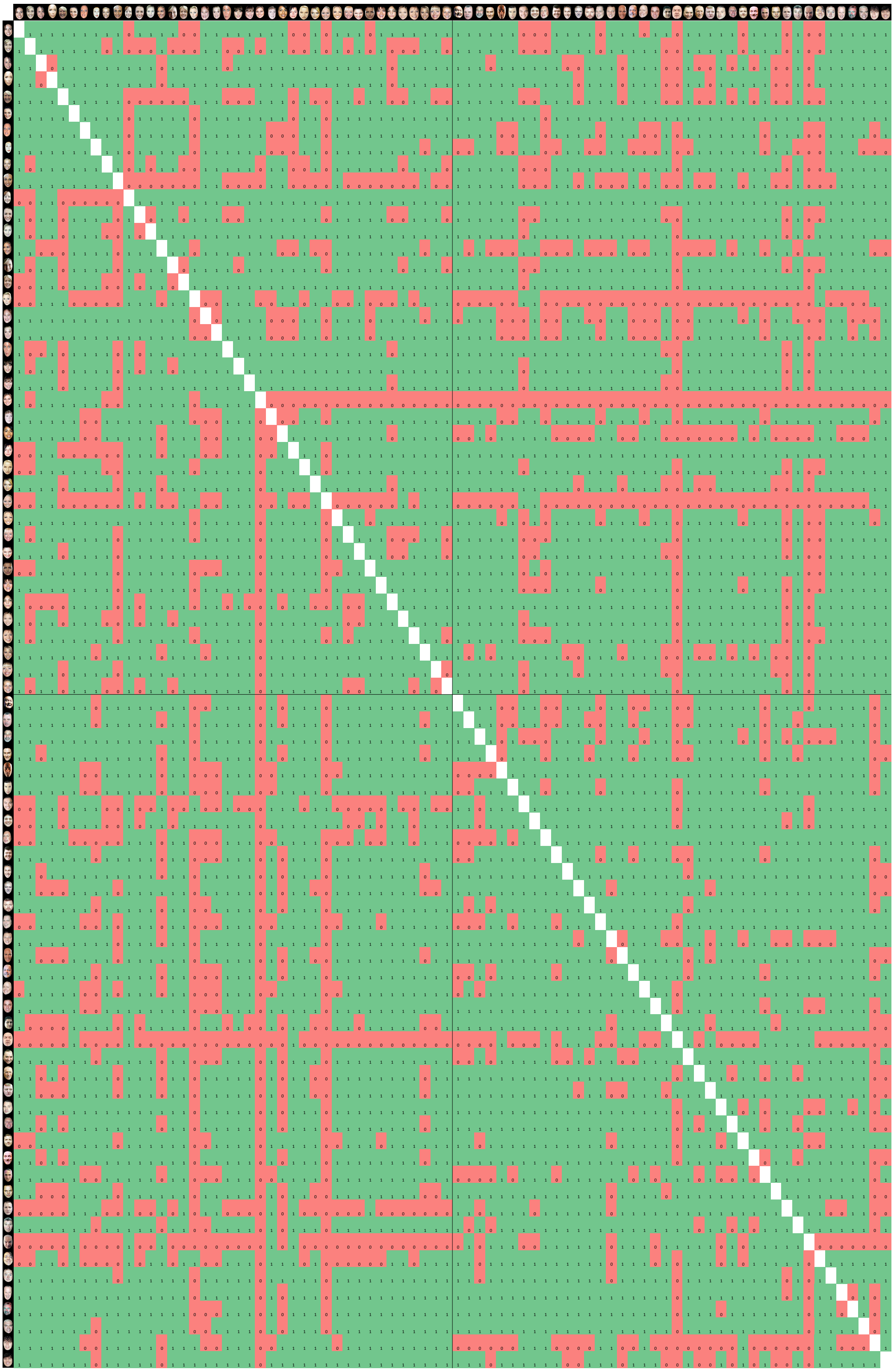


Figure 4.7. **Experiment 8: Dominance Overlap Matrix.**
Matrix illustrating every time the dominance ratings for one face overlaps with the dominance ratings for another face. Faces that overlap are shown in green, and faces that don't overlap are shown in red.

Familiarity

By conducting the experiment in two different locations it was possible to manipulate familiarity by using images of both British and Australian famous faces as stimuli. The images remained constant therefore every image served in the familiar and unfamiliar condition. Previously all ratings for images were combined. To examine familiarity, the ratings were divided into familiar ratings (ratings given to a face familiar to participants) and unfamiliar ratings (ratings given to a face unfamiliar to participants).

First the familiar and unfamiliar ratings for each face were visually compared in a boxplot independently for trustworthiness and dominance (see Figures 4.8, 4.9, 4.10 and 4.11). From the boxplots it is possible to visualise participants' rating behaviour. Most notably the high and low variability that can occur by face is clearly illustrated. This range of high and low variability occurs for both trustworthiness and dominance ratings suggesting that variability is not trait dependent.

Ratings for familiar and unfamiliar faces are also shown, with unfamiliar ratings represented in blue and familiar ratings represented in green. Again there is a lot of variability in the comparison between familiar and unfamiliar ratings and different patterns of ratings can be visualised. For example faces are frequently rated similarly and dissimilarly across familiarity. Large and small ranges in variability also occur across familiarity.

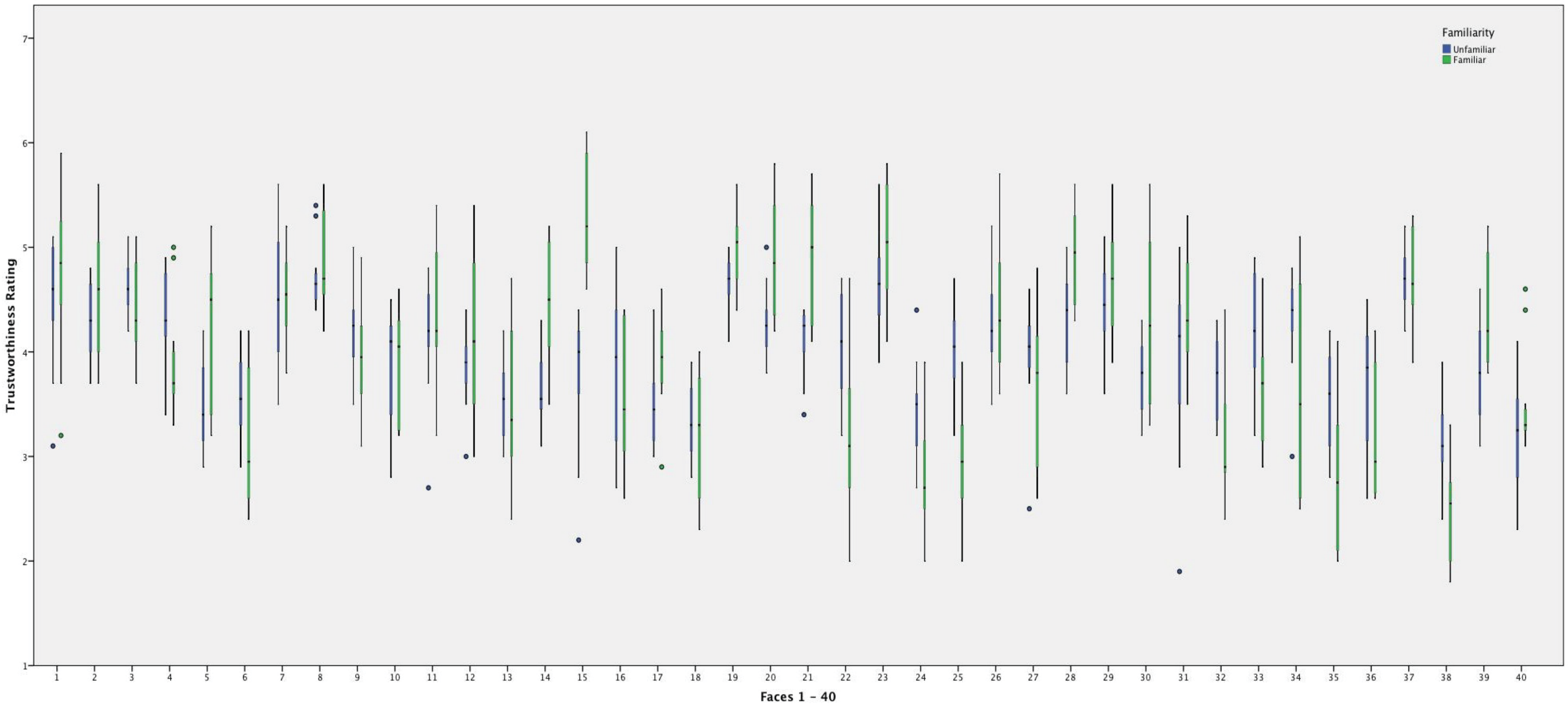


Figure 4.8. **Experiment 8: Trustworthy ratings by familiarity (Faces 1-40).**
A boxplot illustrating the trustworthy ratings for faces 1 - 40, showing ratings from participants familiar with the face in green and ratings from participants unfamiliar with the faces in blue.

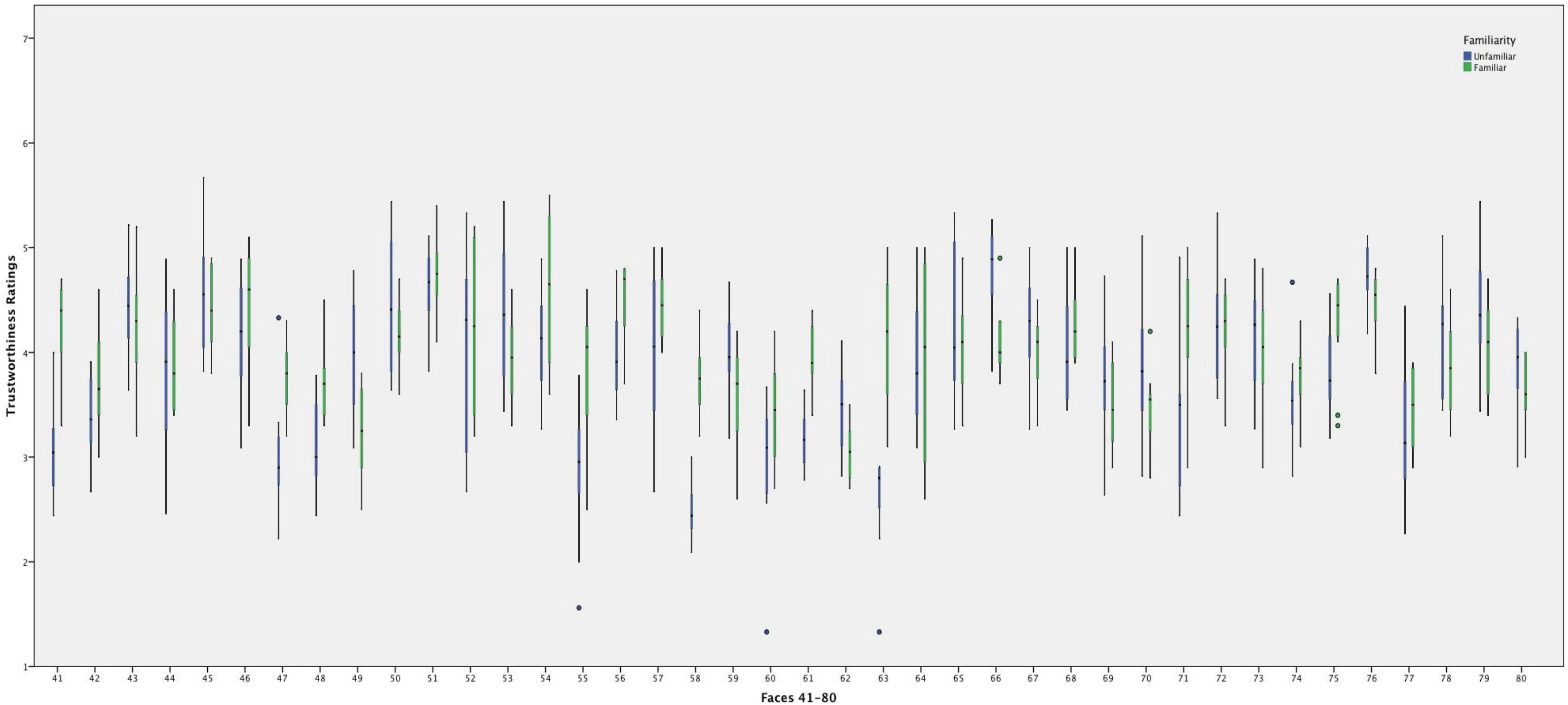


Figure 4.9. **Experiment 8: Trustworthy ratings by familiarity (Faces 41-80).**
A boxplot illustrating the trustworthy ratings for faces 41 - 80, showing ratings from participants familiar with the face in green and ratings from participants unfamiliar with the faces in blue.

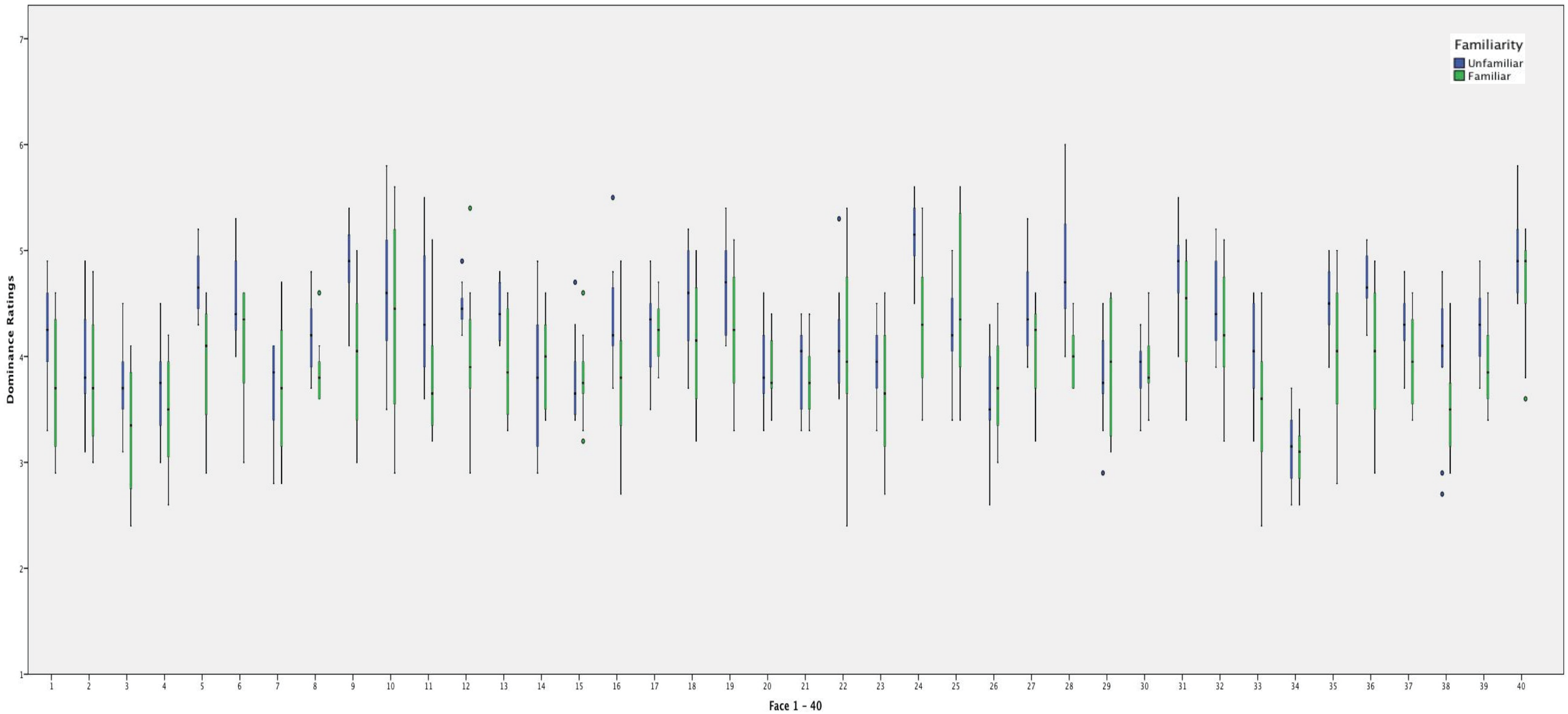


Figure 4.10. **Experiment 8: Dominance ratings by familiarity (Faces 1-40).**
A boxplot illustrating the dominance ratings for faces 1 - 40, showing ratings from participants familiar with the face in green and ratings from participants unfamiliar with the faces in blue.

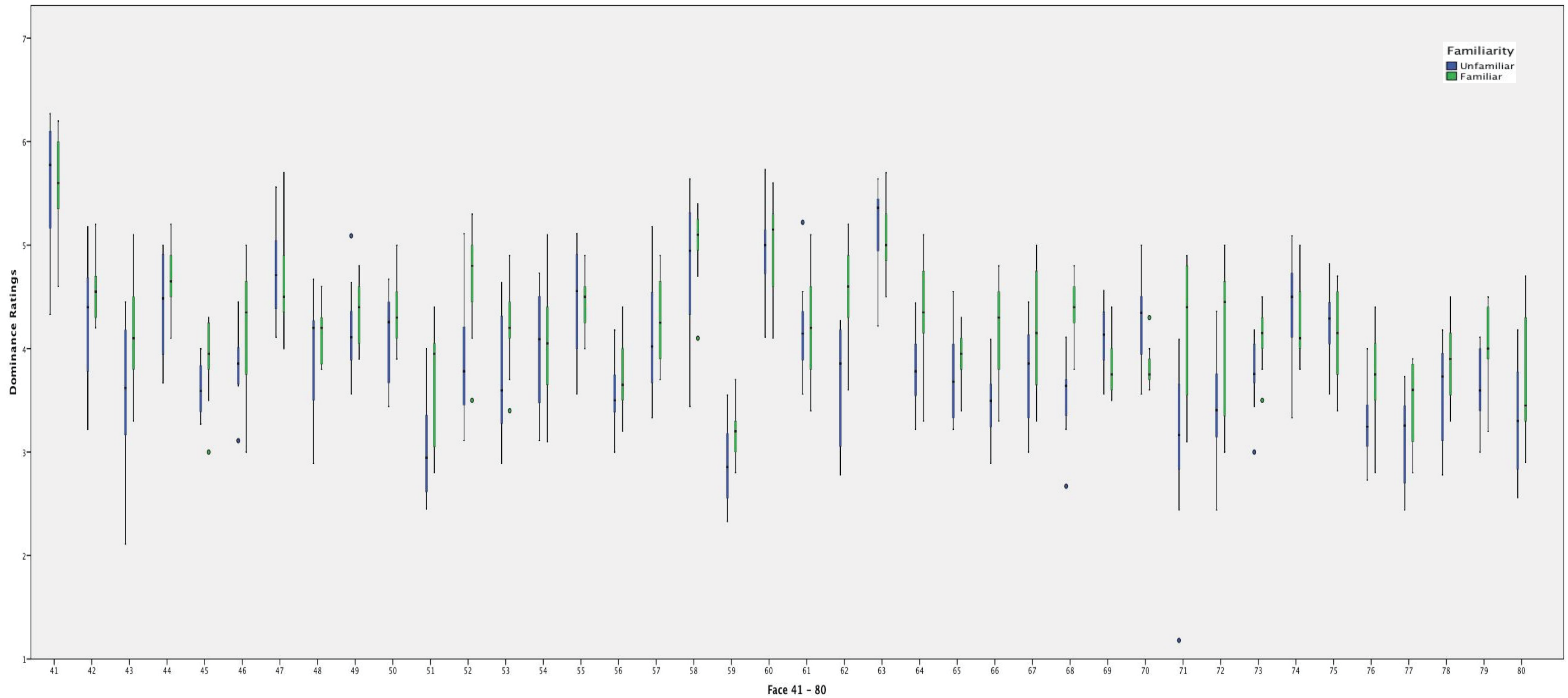


Figure 4.11. **Experiment 8: Dominance ratings by familiarity (Faces 41-80).**
A boxplot illustrating the dominance ratings for faces 41 - 80, showing ratings from participants familiar with the face in green and ratings from participants unfamiliar with the faces in blue.

It was hypothesised that familiar ratings will be significantly different than unfamiliar ratings as previously formed impressions would influence perceptions of familiar faces. To explore whether familiar ratings are significantly different than unfamiliar ratings paired t-tests were conducted for each face, 160 t-tests in total: 80 comparing trustworthiness ratings and 80 comparing dominance ratings.

It was also hypothesised that familiar ratings would be significantly less varied than unfamiliar ratings as previously formed impressions would lead to similar ratings being made for images of familiar faces. To explore whether familiarity ratings are significantly less varied than unfamiliar ratings F-tests were conducted for each face, 160 t-tests in total: 80 comparing trustworthiness ratings and 80 comparing dominance ratings.

Are trustworthiness ratings significantly different for images of familiar and unfamiliar faces?

Given the high number of paired sample t-tests conducted, it would be expected that 4 out of the 80 tests would reach significance by chance. From the t-tests, significant differences were found between familiar and unfamiliar ratings for 42 of the 80 faces rated (see Table 4.1). This high number of significant outcomes suggests that being familiar with a face does have a significant effect on the trustworthiness ratings attributed to images of familiar faces.

Of the 42 faces rated significantly different 21 are rated as more trustworthy when they are familiar to participants and 21 are rated as more trustworthy when they are unfamiliar to participants. Therefore familiarity does not equate to overall being more or less trustworthy, rather the ratings are dependent on the identity of the face.

For example, familiar participants rated images of Face 49 as significantly less Trustworthy than unfamiliar participants. Face 49 represents a celebrity recently revealed by the media to have been unfaithful in their marriage. In comparison familiar participant rated images of Face 54 as significantly more Trustworthy than unfamiliar participants. Face 54 represents a popular family-friendly TV presenter.

Considering the mean differences between ratings when a face is familiar or unfamiliar to participants (regardless of whether the difference is statistically significant) shows that overall 36 of the 80 faces are rated as more trustworthy if they are familiar to participants. This supports the t-test findings that being familiar or unfamiliar does not increase or decrease perceptions of trustworthiness but rather trustworthiness is dependent on the identity of a face. If you are familiar with a person and believe them to be trustworthy this will be reflected in their trustworthiness ratings, and if you are familiar with a person and believe them to be Untrustworthy this too will be reflected in their trustworthiness ratings.

***Table 4.1.* Experiment 8: Trustworthy ratings by familiarity.**
Table of paired sample t-test results for 80 faces, significant differences between familiar and unfamiliar ratings indicated with *.

Face	t	p	*significant	Face	t	p	*significant
1	-2.38	0.04	*	41	7.89	0.00	*
2	-1.34	0.21		42	2.07	0.06	
3	1.77	0.10		43	-1.28	0.23	
4	2.86	0.02	*	44	0.44	0.67	
5	-2.77	0.02	*	45	-0.64	0.53	
6	2.94	0.01	*	46	3.23	0.01	*
7	0.04	0.97		47	3.35	0.01	*
8	-0.85	0.41		48	5.31	0.00	*
9	1.26	0.23		49	-4.17	0.00	*
10	-0.68	0.51		50	-1.60	0.14	
11	-1.10	0.29		51	0.97	0.35	
12	-1.29	0.23		52	0.79	0.45	
13	0.03	0.98		53	-2.07	0.06	
14	-3.76	0.00	*	54	2.31	0.04	*
15	-9.28	0.00	*	55	3.38	0.01	*
16	1.14	0.28		56	3.82	0.00	*
17	-4.37	0.00	*	57	1.60	0.14	
18	0.85	0.42		58	14.21	0.00	*
19	-2.38	0.04	*	59	-3.27	0.01	*
20	-3.48	0.01	*	60	2.11	0.06	
21	-6.29	0.00	*	61	5.11	0.00	*
22	2.84	0.02	*	62	-3.26	0.01	*
23	-3.33	0.01	*	63	5.67	0.00	*
24	4.06	0.00	*	64	-0.01	0.99	
25	6.06	0.00	*	65	-1.06	0.31	
26	-1.08	0.31		66	-5.18	0.00	*
27	1.38	0.20		67	-2.70	0.02	*
28	-6.00	0.00	*	68	1.96	0.08	
29	-1.36	0.20		69	-1.49	0.16	
30	-1.99	0.07		70	-2.30	0.04	*
31	-2.80	0.02	*	71	2.95	0.01	*
32	3.66	0.00	*	72	-0.07	0.95	
33	5.17	0.00	*	73	-1.03	0.32	
34	2.08	0.06		74	1.85	0.09	
35	3.60	0.00	*	75	3.05	0.01	*
36	2.51	0.03	*	76	-2.59	0.03	*
37	-0.35	0.74		77	1.53	0.15	
38	6.44	0.00	*	78	-1.89	0.09	
39	-4.57	0.00	*	79	-2.30	0.04	*
40	-1.24	0.24		80	-1.27	0.23	

Are dominance ratings significantly different for familiar and unfamiliar faces?

Again given the high number of paired sample t-tests conducted, it would be expected that 4 out of the 80 tests would reach significance by chance. From the t-tests, significant differences were found between familiar and unfamiliar ratings for 27 of the 80 faces rated (see Table 4.2). This high number of significant outcomes suggests that familiarity does have a significant effect on ratings of dominance.

Of the 27 faces rated significantly different, 25 faces were rated as more dominant by familiar participants. This suggests that familiar faces are perceived as more dominant than unfamiliar faces. Perceiving familiar faces as more dominant may again be related to preconceived impressions of a person. When rating images of familiar faces further information about that person may come to mind, causing the image to have a more dominant impression that an image of an unfamiliar face. Alternatively this could be a sampling effect, a result of using famous faces who are typically more dominant given their profession.

Considering the mean differences between ratings when a face is familiar or unfamiliar to participants (regardless of whether the difference is statistically significant) shows that overall 66 of the 80 faces are rated as more dominant when they are familiar to participants. This supports the t-test findings that faces are perceived as more dominant if they are familiar to participants.

Table 4.2. **Experiment 8: Dominance ratings by familiarity.**
Table of paired sample t-test results for 80 faces, significant differences between familiar and unfamiliar ratings indicated with *.

Face	t	p	*significant	Face	t	p	*significant
1	1.46	0.17		41	0.07	0.94	
2	0.65	0.53		42	2.07	0.06	
3	1.71	0.11		43	1.53	0.15	
4	0.91	0.38		44	1.53	0.15	
5	4.36	0.00	*	45	2.33	0.04	*
6	2.17	0.05		46	1.42	0.18	
7	0.03	0.97		47	-0.90	0.39	
8	2.91	0.01	*	48	0.64	0.53	
9	4.22	0.00	*	49	1.42	0.18	
10	0.69	0.50		50	2.32	0.04	*
11	2.27	0.04	*	51	2.53	0.03	*
12	2.54	0.03	*	52	5.58	0.00	*
13	2.81	0.02	*	53	2.07	0.06	
14	-1.23	0.24		54	0.19	0.85	
15	-0.20	0.84		55	-0.12	0.91	
16	3.35	0.01	*	56	1.37	0.20	
17	0.09	0.93		57	0.63	0.54	
18	1.72	0.11		58	1.04	0.32	
19	1.66	0.13		59	1.86	0.09	
20	0.23	0.83		60	0.06	0.96	
21	0.55	0.60		61	0.20	0.84	
22	0.26	0.80		62	3.38	0.01	*
23	1.04	0.32		63	-0.71	0.49	
24	4.71	0.00	*	64	3.18	0.01	*
25	-0.91	0.38		65	1.31	0.22	
26	-0.70	0.50		66	3.50	0.00	*
27	3.06	0.01	*	67	1.80	0.10	
28	3.92	0.00	*	68	6.46	0.00	*
29	-0.24	0.82		69	-2.30	0.04	*
30	-0.14	0.89		70	-3.53	0.00	*
31	2.15	0.06		71	3.78	0.00	*
32	1.07	0.31		72	2.39	0.04	*
33	2.07	0.06		73	2.68	0.02	*
34	0.57	0.58		74	-0.58	0.58	
35	1.62	0.13		75	-0.63	0.54	
36	3.66	0.00	*	76	2.59	0.03	*
37	1.65	0.13		77	2.06	0.06	
38	1.91	0.08		78	1.50	0.16	
39	2.58	0.03	*	79	3.54	0.00	*
40	1.40	0.19		80	1.37	0.20	

Are Trustworthy ratings significantly more varied for familiar or unfamiliar faces?

From the F-tests, significant differences in variance were found between familiar and unfamiliar ratings for 6 out of the 80 faces (see Table 4.3). Out of the 6 faces, all were rated as more varied by unfamiliar participants. This could suggest that within-person variability occurs more with unfamiliar faces, however given the number of F-tests performed, this low outcome of significant differences could be expected just from chance.

Table 4.3. **Experiment 8: Trustworthy variability by familiarity.**
Table of F-test results for 80 faces, significant differences between familiar and unfamiliar ratings indicated with *.

Face	F	p	sig	Face	F	p	sig
41	0.95	0.93		1	0.62	0.44	
42	1.58	0.46		2	0.32	0.07	
43	1.48	0.53		3	0.30	0.06	
44	0.31	0.07		4	0.58	0.38	
45	0.48	0.24		5	0.36	0.11	
46	1.04	0.95		6	0.34	0.09	
47	0.44	0.19		7	2.65	0.12	
48	0.79	0.71		8	0.42	0.16	
49	0.61	0.42		9	0.92	0.89	
50	0.23	0.02	*	10	1.04	0.95	
51	0.86	0.81		11	0.64	0.47	
52	0.82	0.75		12	0.19	0.01	*
53	0.33	0.08		13	0.24	0.03	*
54	1.99	0.27		14	0.34	0.09	
55	1.07	0.92		15	1.46	0.54	
56	0.72	0.60		16	1.30	0.67	
57	0.21	0.02	*	17	0.82	0.74	
58	1.56	0.47		18	0.33	0.08	
59	1.24	0.72		19	0.49	0.25	
60	0.59	0.39		20	0.32	0.07	
61	1.35	0.63		21	0.29	0.05	
62	0.52	0.30		22	0.43	0.18	
63	1.84	0.33		23	0.60	0.41	
64	2.22	0.20		24	0.77	0.67	
65	0.49	0.25		25	0.67	0.52	
66	0.54	0.32		26	0.51	0.28	
67	0.48	0.24		27	0.53	0.31	
68	0.41	0.16		28	0.78	0.69	
69	0.54	0.32		29	0.54	0.33	
70	0.39	0.13		30	0.17	0.01	*
71	0.74	0.63		31	2.23	0.20	
72	0.54	0.32		32	0.38	0.12	
73	0.98	0.98		33	1.02	0.97	
74	0.48	0.23		34	0.22	0.02	*
75	1.33	0.65		35	0.42	0.17	
76	1.12	0.86		36	0.88	0.84	
77	0.34	0.09		37	0.35	0.09	
78	0.78	0.69		38	0.91	0.88	
79	0.66	0.50		39	0.84	0.77	
80	0.70	0.56		40	1.27	0.70	

Are dominance ratings significantly more varied for familiar or unfamiliar faces?

Again given the high number of F-tests conducted, it would be expected that 4 out of the 80 tests would reach significance by chance. From the F-tests, significant differences in variance were found between familiar and unfamiliar ratings for 13 out of the 80 faces (see Table 4.4). Therefore for some faces significant differences in variance of dominance ratings occurred between familiar and unfamiliar participants. Of the 13 faces, 11 faces had significantly more variance in ratings from unfamiliar participants than familiar participants. This suggests that for some faces perceptions of dominance were significantly more varied if the face was unfamiliar.

Table 4.4. **Experiment 8: Dominance variability by familiarity.**
Table of F-test results for 80 faces, significant differences between familiar and unfamiliar ratings indicated with *.

Face	F	p	sig
1	0.57	0.37	
2	0.73	0.61	
3	0.45	0.20	
4	0.68	0.53	
5	0.23	0.02	*
6	0.56	0.35	
7	0.55	0.33	
8	1.32	0.65	
9	0.39	0.13	
10	0.52	0.29	
11	1.08	0.90	
12	0.09	0.00	*
13	0.28	0.05	*
14	2.76	0.11	
15	1.31	0.67	
16	0.49	0.25	
17	1.68	0.40	
18	0.65	0.49	
19	0.53	0.31	
20	1.65	0.42	
21	1.60	0.45	
22	0.31	0.07	
23	0.31	0.06	
24	0.27	0.04	*
25	0.28	0.05	*
26	1.30	0.67	
27	0.84	0.78	
28	4.33	0.02	*
29	0.50	0.27	
30	0.94	0.91	
31	0.56	0.35	
32	0.59	0.40	
33	0.60	0.41	
34	1.52	0.50	
35	0.20	0.01	*
36	0.22	0.02	*
37	0.48	0.24	
38	1.78	0.35	
39	0.90	0.86	
40	0.53	0.31	

Face	F	p	sig
41	0.51	0.28	
42	0.22	0.02	*
43	0.56	0.35	
44	0.39	0.13	
45	2.01	0.26	
46	3.81	0.04	*
47	0.99	0.99	
48	0.25	0.03	*
49	0.51	0.28	
50	0.55	0.34	
51	1.47	0.53	
52	0.77	0.67	
53	0.48	0.23	
54	0.93	0.91	
55	0.31	0.06	
56	1.05	0.94	
57	0.46	0.22	
58	0.28	0.04	*
59	0.35	0.10	
60	1.17	0.80	
61	1.54	0.49	
62	0.62	0.44	
63	0.65	0.49	
64	1.89	0.30	
65	0.32	0.07	
66	1.53	0.49	
67	1.47	0.53	
68	0.57	0.37	
69	0.91	0.88	
70	0.19	0.01	*
71	0.79	0.71	
72	1.56	0.47	
73	0.68	0.54	
74	0.54	0.31	
75	1.75	0.37	
76	1.67	0.41	
77	0.94	0.92	
78	0.57	0.36	
79	0.96	0.95	
80	1.14	0.83	

Discussion

Within-person variability

The findings of Experiment 8 clearly demonstrate that within-person variability does affect perceptions of personality. This is an important finding as within-person variability has not been previous examined in face perception research. Previous research has typically focused on perceptions made between-persons, for example Todorov et al., (2005) compared facial perceptions of competence across different political candidates. An important conclusion to take from the findings of Experiment 8 is that within-person variability exceeds between-person variability. This is best demonstrated by considering the ratings of images of different faces that overlap. As shown in Figure 4.6 and 4.7

overlapping of ratings of different faces occurs over 75% of the time; this illustrates the similarities of ratings between-person, in contrast ratings are more variable within-person.

Overlapping can also demonstrate the extent of within-person variability, for example by simply selecting a different image of a face personality perceptions can be reversed between-person (see Figure 4.5 for an example). This reversal of perceptions through changing images is a clear illustration of the importance of considering within-person variability: perceptions of trustworthiness and dominance are dependent on the image of a face as oppose to dependent on a face. This finding has important implications for future research and the use of images in experimental and real life settings.

For example, a large quantity of previous research has focused on different behaviours, attitudes and outcomes that can be influenced by the perception of personality from facial appearance (Todorov et al., 2005, 2008; Mueller and Mazur 1996). From these results it can be suggested that by altering your facial appearance, or appearance in an image representative of you, perceptions about your personality can be changed. This could be highly beneficial in a number of real-life scenarios such as the images people choose to represent themselves from work identification cards, to passport photographs and even images put on personal websites or networking sites.

The lack of focus on within-person variability in face perception research comes back to the practice of using one image to represent a face. Experiment 8 has shown that within-person variability influences participants' perceptions therefore isolating just one image to represent a face can be misleading and inaccurate. One image will not give a stable representation of a person's facial appearance, or lead to a stable perception of personality. To gain a better understanding of the perception of personality from facial appearance and how these perceptions shape our behaviour multiple images should be used to represent a person's face. The conclusion that within-person variability has to be considered in future research therefore goes hand in hand with the use of multiple images to adequately represent a person's face.

It is also interesting to note that the extent of within-person variability can differ from one face to the next, with some faces varying widely across the traits trustworthiness and dominance while other faces have more consistency in their ratings. This reflects the facial appearance of each face: some faces in the stimuli have more within-person variability than others. This is similar to the finding in Chapters 2 and 3 that some faces are easier to identify across within-person variability than others. It would be interesting to explore how different faces differ for within-person variability as it could help build a better understanding of how within-person variability is processed.

Trustworthiness and dominance

The personality traits trustworthiness and dominance were selected for participants to rate because previous research had illustrated that these two characteristics are both easily recognisable from facial appearance and account for the most variability between-person (Oosterhof & Todorov 2008). The results of Experiment 8 support this finding and also show that trustworthiness and dominance perceptions are highly variable within-person.

The findings of Experiment 8 illustrate that perceptions of personality are not based solely on anatomical invariants of a face, such as bone structure, but also based on transient facial characteristics, such as eye gaze, and image characteristics such as lighting. This suggests that research that manipulated anatomical changes, such as sexual dimorphism and face symmetry, to explore the different effects it has on perceptions is not encompassing all forms of face variability (Little & Hancock 2002; Little et al., 2012). In Experiment 8 perceptions of trustworthiness and dominance varied greatly without alterations being made to the structure of each face, demonstrating that perceptions can be influenced by transient factors.

Further analysis of the ratings also illustrates differences that occur between trustworthiness and dominance ratings. For example, comparing the overlapping matrices (Figures 4.6 and 4.7) ratings of trustworthiness overlap more than ratings of dominance. This suggests that dominance ratings are more likely to be particularly low or high on the 7-point scale resulting in more between-person variability for dominance.

When examining the images behind the ratings patterns behind the perceptions can be observed (see Figure 4.4). For example a particularly smiley image may be rated higher for trustworthiness than an image of the same face not smiling and looking down. This suggests that expression and eye gaze are important in perceiving trustworthiness. This finding supports previous research that has shown eye gaze to be key in the processing of social judgments (Willis, Palermo & Burke 2011). Meanwhile a forward facing image with a neutral expression may be rated as higher for dominance than an image of the same face smiling. This suggests that head position and expression are important in perceiving dominance. An interesting follow-up study would be to perform an image analysis on the same stimuli to observe whether patterns emerge to explain why certain images are perceived as more or less trustworthy or dominant.

Familiarity

In Experiment 8 familiarity was hypothesised to influence personality perceptions in two ways: participants would rate different images of a familiar face as less variable and images of faces would be rated different by participants familiar and unfamiliar with the faces. It was predicted that less variability would occur between ratings of images of familiar faces as participants would have a previously formed impression of the person leading them to base their perceptions on this formed impression. If all of their ratings were based on the same impression then less variability would occur.

From the F-test analysis it was shown that variability in trustworthiness ratings were similar across familiarity. This is an interesting finding as it shows that these perceptions are not influenced by previous impression made about the person. It could be concluded that additional information known about a familiar face isn't as influential as might be expected as personality perceptions vary by image. This finding also suggests that research claiming first impressions of a face are lasting may not present an accurate conclusion (Willis & Todorov 2006). Instead it could suggest that first impressions of an *image* are lasting but overall impressions of a person's personality can vary dependent on the image shown to represent the person.

The F-test analysis of dominance ratings shows that variability is significantly different for 13 out of the 80 faces. This finding does not suggest an overall difference in variability between familiar and unfamiliar ratings but does suggest that dominance is likely to be a more variable perception. This is reflected in the overlapping matrices where less overlap is shown for the trait dominance than the trait trustworthiness. Overall the difference in variance of dominance ratings is fairly equal between familiar and unfamiliar faces with ratings for 43 out of the 80 faces being more variable when the face is familiar to participants and ratings for the remaining 37 faces more variable when the face is unfamiliar to participants. The F-test analysis underlines the point made above that perceptions of faces are not stable but depend greatly on the image shown of a face.

It was also predicted that ratings would be significantly different when a face is familiar to participants. When participants see an image of a familiar face it may trigger their previously learnt information about that person. This information could then influence their ratings. In comparison when participants see an unfamiliar face the only information they have about the person is the image, therefore their perceptions are based solely on how the person looks in each image.

From the paired sample t-test analysis it was shown that ratings of trustworthiness were significantly different between familiarity for 42 of the 80 faces. Of the significant results there is an even split of whether familiar or unfamiliar faces are rated as significantly more trustworthy i.e. 21 familiar faces were rated as more trustworthy and 21 unfamiliar faces were rated as significantly more trustworthy. This illustrates that being familiar does not solely increase or decrease a perception of trustworthiness but rather previously formed impressions of a familiar face will influence trustworthiness ratings whether the impression is positive (trustworthy) or negative (untrustworthy).

The paired sample t-test analysis found that ratings of dominance were significantly different between familiarity for 27 of the 80 faces. Of the 27 faces, 25 faces were rated as more dominant when they were familiar to participants. This suggests that facial perceptions of dominance will increase when a face is familiar. The exception occurs with 2 faces that are rated as lower in dominance when they are familiar to participants. The two faces represent two famous people who are known for a submissive nature. This suggests that familiarity often amounts to being perceived as more dominant unless the person is widely thought of as a particularly submissive person.

Overall the effect of familiarity on personality perceptions is quite mixed. Previously formed impressions generated from being familiar with a person, appear to have a large influence on both trustworthy and dominance ratings. The effect of familiarity on differences of variability is less clear. Significant differences in variability between familiar and unfamiliar ratings do exist for dominance but only for 13 out of the 80 faces suggesting that it is not an overall effect but rather could be explained by individual identity variability. The number of significant differences in variability between familiar and unfamiliar ratings for trustworthiness is very low suggesting that variability for trustworthiness is similar when a face is familiar or unfamiliar to participants. The variability findings are of particular interest as they underline the importance of within-person variability; within-person variability can occur with both familiar and unfamiliar faces suggesting that it is a robust concept.

Overall the findings of Experiment 8 have shown that perceptions of trustworthiness and dominance can vary within-persons. This finding underlines the importance of within-person variability in face processing and demonstrates the need to consider within-person variability as an important concept in face perception.

Chapter Five

Within-Person Variability in Facial Attractiveness

5.1 Introduction

In Chapter 4, within-person variability has been shown to affect personality perception with different images of the same face varying for the personality traits trustworthiness and dominance. This has shown that face perceptions are not constant but vary depending on different images of a face. This finding draws into question whether within-person variability affects other perceptions that are often made from facial appearances. For example research has shown perceptions of attractiveness can be made after minimal exposure to a face (Olson & Marshuetz 2005). If within-person variability affects attractiveness, this would suggest that the perception of attractiveness is not a constant perception but can be manipulated dependent on the image shown to represent a face.

Research has shown facial attractiveness can influence many different areas in life. For example facial attractiveness has long been shown to be a significantly important factor in partner selection (Walster, Aronson, Abrahams & Rottmann 1966; Kalick, Zebrowitz, Langlois, & Johnson 1998), as well as having a significant effect on everyday social behaviours and interactions such as decision-making and personality perceptions (Langlois, Kalakanis, Rubenstein, Larson, Hallam & Smoot 2000; Watkins & Johnston 2000).

Facial attractiveness has been shown to affect peoples' perceptions in a range of different contexts. For example attractiveness has a positive effect on judgments of work competency from school age, to university, to adults in the workplace (Clifford & Walster 1973; Landy & Sigall 1974; Cash & Kilcullen 1985). More recently in a study from Watkins and Johnston (2000) facial attractiveness was shown to have a positive effect on employability. Participants were asked to read two different job applications of varying quality with images of either an attractive female or an average female face attached. The results found that when applications were of poorer quality an attached image of an attractive face had a positive effect on participants' decisions of employability over an attached image of an unattractive face.

Facial attractiveness has also been shown to influence political election outcomes (Sigelman, Sigelman, Thomas & Ribich 1986; Rosar, Klein & Beckers

2008; Lenz & Lawson 2011). For example in a study from Rosar, Klein and Beckers (2008) attractiveness influenced voters' behaviour with attractive political candidates receiving higher polls. In addition, the more attractive the candidates, the more people turnout to vote. Facial attractiveness has even been shown to influence jury decision-making, for example in a meta-analysis study of jury research from Mazzella and Feingold (1994) it was found that if a defendant was physically attractive they were more positively perceived by jurors and so less likely to be perceived as guilty. These findings illustrate the extent to which facial attractiveness can affect important political and criminal decision-making, and so highlight the importance and impact facial attractiveness can have.

As well as it's impact on decision-making and perceptions in professional, political and criminal settings, research has shown facial attractiveness can play an important role in social environments. For example in a study from Dion, Berschied and Walster (1972) the stereotype "What is beautiful is good" was examined with participants rating images of attractive, unattractive and average looking faces for different personality and life traits. The results showed that attractiveness has a significant effect on perceptions of personality and lifestyle with attractive faces rated higher for more desirable social personality characteristics as well as being rated more likely to experience success in their working and personal life. More recently this stereotype has continued to be found with attractive people being thought of as more positive in regards to their personality and intelligence, as well as their behaviour and other peoples' behaviour towards them (Lorenzo, Biesanz & Human 2010; Langlois, Kalakanis, Rubenstein, Larson, Hallam & Smoot 2000).

These studies illustrate the substantial impact attractiveness can have on decision-making and perceptions. From these findings it could be suggested that improving facial appearance to appear more attractive would be beneficial in social and professional interactions. A recent study from Morrison, Morris and Bard (2013) addresses the possibility of being able to improve facial attractiveness by exploring differences in attractiveness perceptions across different facial expressions. Participants were asked to rate images of different models with the different facial expressions: anger, fear, disgust, surprise, sadness, happiness and a neutral expression. The results found that attractiveness ratings did differ between expressions, however the results also show that attractiveness ratings were identity dependent i.e. differed more significantly by identity of the model in the image. This lead to the conclusion that underlying facial attractiveness is more important than facial expression in perceptions of attractiveness, and that facial attractiveness is a stable quality that cannot be manipulated.

However as previously discussed, a person's facial appearance can change and vary across a wider range of characteristics than solely facial expression. Within-person variability is a concept that has been overlooked in facial attractiveness research; instead the focus has been on between-person variability. For example, Morrison et al.,'s study on improving attractiveness through facial expressions compared different facial expressions *between-person* (2013). A clear indication of between-person variability is when one image is used to represent a face. The previous literature discussed all adopted the use of one image for example Watkins and Johnston (2000) used one image of an attractive female and one image of an unattractive female to explore attractiveness influence in job applications, while Rosar et al., (2008) used a single image to represent political candidates. Therefore these studies overlook the possibility that attractiveness can change within-person and so alter perceptions of that person.

The concept of within-person variability is of particular importance when considering research that manipulates face characteristics to show faces that are perceived as more or less attractive (Coetzee, Re, Perrett, Tiddeman & Xiao 2011; Little & Hancock 2002; Little, Hockings, Apicella & Sousa 2012). For example, research often examines the perception of attractiveness from faces that have been computer-generated or manipulated to vary across structural characteristics such as face symmetry (Little et al., 2001). As previously discussed, within-person variability often exceeds variability controlled for with this technique, and generating faces in this way can limit the understanding of variability in real faces. Therefore it is important to consider realistic variability with real faces.

In a recent study from Jenkins, White, van Montfort and Burton (2011) within-person variability is explored in perceptions of facial attractiveness.

Participants' viewed 20 images of 20 different unfamiliar faces (10 males; 10 females) and were asked to indicate whether they thought the person in the image was attractive or not with a 'yes/no' response. Participants were made aware that the same face might appear in different images and it was emphasised that the judgment of attractiveness should be based on the specific image. Participants' responses could then be combined to illustrate how often an image was perceived as attractive. The results found that variability occurred between different images of the same face with participants often categorizing images of the same face as both attractive and unattractive. This occurred to the extent that within-person variability was greater than between-person variability: it was possible to reverse attractiveness preferences simply by changing the image representing a face.

In Experiment 9 within-person variability is explored in facial attractiveness with participants being asked to rate multiple images of different faces. This experiment develops on from the findings of Jenkins et al., (2011) by firstly asking participants to make a rating on a 7-point scale to get a more detailed impression of attractiveness perceptions, and also examining the effect of familiarity. It could be hypothesised that ratings of attractiveness will be less varied for familiar faces than unfamiliar, as participants may be influenced by previously formed impressions of familiar faces based on their personality or behaviour, or from previously seen images of them. However considering the results of Experiment 8 on trustworthiness perceptions, it is hypothesised that variability will occur with both ratings of familiar and unfamiliar faces. Based on the findings of Experiment 8, it is also hypothesised that ratings of familiar faces will be significantly different than ratings of unfamiliar faces.

5.2 Experiment 9: Within-person variability in facial attractiveness

Method

Participants

Forty-eight undergraduate students participate in this experiment in exchange for a small payment. Twenty-four of the participants were British students at the University of Glasgow and twenty-four of the participants were Austrian students at the University of Vienna.

Design and Stimuli

By conducting the experiment in two different locations (Glasgow and Vienna) it was possible to manipulate familiarity by using images of both British and Austrian famous faces as stimuli (see Table B.1, Appendix B for list of names). The images remained constant therefore every image served in the familiar and unfamiliar condition. This resulted in a repeated-measures design.

A list of ten British and ten Austrian male and female celebrities was compiled ensuring each person was familiar solely in their country. Twelve images of each face were collected through an online search, 240 images in total. High quality images were carefully selected, with the face clearly visible and not obstructed by other objects or people. The images were not further controlled for therefore the final twelve images represented changes that encompass within-person variability. The images were individually edited and cropped to show only the face (see Figure 5.1 for an example), and uniformly sized to 250 x 375 pixels.

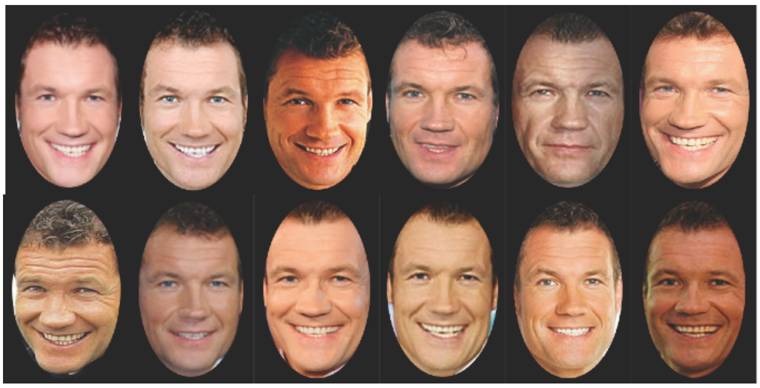


Figure 5.1. **Example of images used in Experiment 9.**
Twelve images of Face 1 illustrating the typical variability of images of each face shown in Experiment 9.

The images were then inserted into a computer task using experiment-building software, LiveCode. The task was programmed to repeat the instructions given to participants in text form, with a “Start Experiment” button available at the bottom of the screen for participants to press when ready to begin the task. Following this, the images would appear individually on the screen, sized 6 x 9 cms, above a seven-button scale of 1 – 7. To move onto the next image, participants were required to press a button on the scale.

Procedure

Participants were informed that the experiment involved rating faces for attractiveness. It was explained that the images would appear individually on the computer screen above a button scale of 1 – 7. Participants were asked to indicate how attractive they perceived the image shown on the screen to be on a seven-point scale (1 = not at all and 7 = very). Following their response, the image would leave the screen and another image would appear. Participants were made aware that there was no time limit and to move onto the next image they were required to make a rating. Participants were made aware that multiple images of the same faces would be presented and were asked to make their attractiveness ratings for each *image* as opposed to each *face*.

Familiarity was a key element of this experiment and so images of people famous in two different locations were used, and the experiment conducted in these two locations: the UK (at the University of Glasgow) and in Austria (at the University of Vienna). To ensure that the faces selected were familiar or unfamiliar to the right participants, a short task was given to participants at the end of the experiment. Participants were shown an image of each face again on a computer screen, along with the name of the famous person, and were asked to indicate with a “Yes/No” response whether they were familiar with each person, by clicking either a “Yes” or “No” button below the image.

Participants’ responses could then be analysed and if any of the appointed familiar faces were “unfamiliar” or any of the appointed unfamiliar faces were “familiar”, data could be excluded. All participants were familiar with the appointed familiar faces and unfamiliar with the appointed unfamiliar faces in their locations, reflecting that the stimuli adequately captured images of familiar and unfamiliar faces in each location.

Results

Attractiveness ratings were combined to create a mean attractiveness rating for each image of each face, when the faces were familiar and unfamiliar to participants. The ratings were then plotted in a boxplot (see Figure 5.2). It appears that images of familiar faces are rated as more attractive than images of unfamiliar faces.

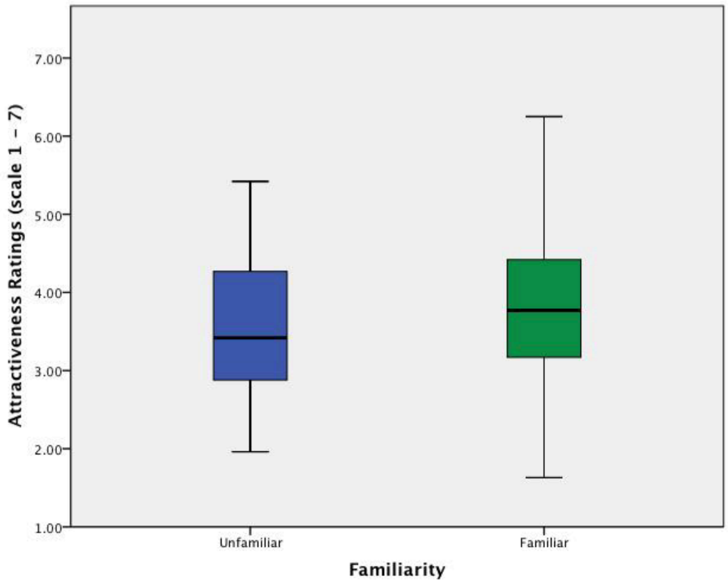


Figure 5.2. **Experiment 9: attractiveness ratings by familiarity.**
A boxplot illustrating the range of attractiveness ratings given to images of faces familiar and unfamiliar to participants.

The mean attractiveness rating for images of familiar faces is 3.78 and the mean attractiveness rating for images of unfamiliar faces is 3.54. A paired-sample t-test found that this difference was significant: images of faces were rated as significantly more attractive when they were familiar compared with when they were unfamiliar [t (239) = 7.28, $p < 0.01$].

To clearly visualise the variability between attractiveness ratings for different images of the same face, the mean attractiveness ratings for each image of each face were plotted independently in a boxplot with familiar ratings and unfamiliar ratings also shown independently (see Figure 5.3). This allows for the range of attractiveness ratings for each face to be clearly illustrated, between-person and within-person.

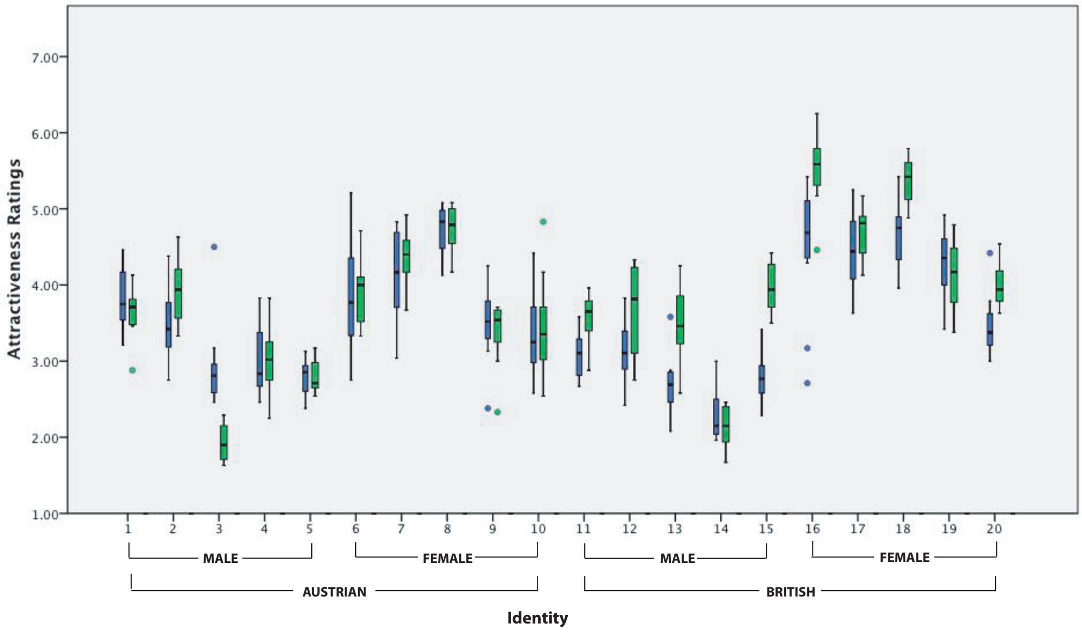


Figure 5.3. **Experiment 9: attractiveness ratings by identity.**
A boxplot illustrating the attractiveness ratings for each image of each face, with unfamiliar ratings shown in blue and familiar ratings shown in green.

Figure 5.3 visualises between-person variability with ratings varying widely across gender and by identity. This illustrates that there is a wide range of between-person variability in attractiveness perceptions. There are also a lot of differences in within-person variability by familiarity. It is possible to see where images have been rated similarly causing overlap between familiar and unfamiliar ratings (e.g. Face 2) and where the faces have been rated differently causing no overlap but instead a distinct range of attractiveness ratings (e.g. Face 3).

To explore whether these differences are significant, paired sample t-tests were conducted comparing the familiar and unfamiliar ratings for each face. From the t-tests significant differences were found between familiar and unfamiliar ratings for 10 out of the 20 faces (see Table 5.1), of these 10 faces 9 were rated as more attractive when the person was familiar to participants. This suggests that familiarity can be more attractive but not for all faces. This reflects the differences in the amount of within-person variability in different faces as described in Chapters 2 and 4.

Table 5.1. **Experiment 9: Effect of familiarity.**
Paired sample t-test results for all 20 faces, significant differences between familiar and unfamiliar attractiveness ratings indicated with a *.

Face	t	p	*significant
1	-1.55	0.14	
2	4.43	0.00	*
3	-32.84	0.00	*
4	-0.47	0.64	
5	-0.10	0.91	
6	0.00	1.00	
7	1.16	0.27	
8	0.66	0.52	
9	0.43	0.67	
10	1.97	0.07	
11	4.41	0.00	*
12	6.29	0.00	*
13	7.99	0.00	*
14	-1.31	0.21	
15	13.92	0.00	*
16	5.99	0.00	*
17	3.41	0.00	*
18	8.05	0.00	*
19	0.49	0.63	
20	9.47	0.00	*

To explore whether familiarity affects variability of attractiveness ratings, F-tests were conducted comparing the familiar and unfamiliar ratings for each face. No significant differences were found between the variability of ratings for familiar and unfamiliar faces (see Table 5.2). This indicates that similar to perceptions of trustworthiness in Experiment 8, there is no difference in the variability of attractiveness between familiar and unfamiliar ratings for each face.

Table 5.2. **Experiment 9: Effect of familiarity.**
F-test results for all 20 faces, significant differences between familiar and unfamiliar attractiveness ratings indicated with a *.

Face	F	p	*significant
1	0.69	0.55	
2	0.49	0.25	
3	1.04	0.94	
4	0.90	0.86	
5	0.59	0.40	
6	0.47	0.22	
7	0.57	0.36	
8	0.67	0.52	
9	0.79	0.70	
10	1.24	0.73	
11	1.26	0.70	
12	2.10	0.23	
13	2.74	0.10	
14	1.59	0.45	
15	1.63	0.43	
16	1.26	0.70	
17	0.66	0.50	
18	0.48	0.24	
19	1.18	0.78	
20	0.50	0.26	

Importantly despite previous research showing the difference in attractiveness between-persons, this data shows that the difference in attractiveness also depends on the image shown of each face. This finding is underlined by the ability to take images of faces and illustrate each face as the more attractive, or lesser attractive face. For example shown in Figure 5.4 are two different images of two faces, Face 9 on the left and Face 10 on the right. The mean unfamiliar attractiveness rating is higher for the top left image than the top right image, yet the mean unfamiliar attractiveness rating is higher for the bottom right image than the bottom left image. Therefore attractiveness judgements can be reversed given different images of the same person.

Figure 5.4. **Experiment 9: Reversal of attractiveness perceptions.**
An example of attractiveness ratings being reversed when shown different images of the same faces: The images of the left show Face 9 and the images on the right show Face 10, in the top row the image on the left was rated as more attractive than the image on the right. In the bottom row the image on the right was rated as more attractive than the image on the left.

Discussion

In Experiment 9 participants' perceptions of attractiveness for one face varies across different images of the face illustrating that within-person variability can have a significant effect on perceptions of facial attractiveness. The results also show that perceptions of attractiveness vary regardless of whether a face is familiar or unfamiliar to participants. This finding emphasises the effect images demonstrating within-person variability can have, as perceptions of attractiveness can vary widely even for a familiar face that may have previously been judged on characteristics such as attractiveness.

Paired sample t-tests comparing attractiveness ratings for images of familiar and unfamiliar faces found that significant differences occur for ten out of the 20 faces. Of the significant findings nine faces are rated as more attractive when they are familiar to participants. This suggests images of familiar faces are likely to be rated as more attractive than images of unfamiliar faces. This is supported by previous research that has found repeated exposure to a face or familiarity increases perceptions of likeability and favourable feelings towards the person (Zajonc 1968; Harrison 1969). Research has also shown that attractive faces can be perceived as more average and more familiar than unattractive faces (Langlois & Roggman 1990; Langlois, Roggman & Musselman 1994). The findings of Experiment 9 suggest that this can occur in reverse, with faces appearing more attractive when they are familiar.

Using images of familiar and unfamiliar faces also allows for the stereotype of beauty symbolising goodness (Dion et al., 1972) to be examined. If previous perceptions of familiar faces influence participants' judgments of attractiveness, this could result in negative impressions causing low attractiveness ratings and positive impression causing high attractiveness ratings. These perceptions may be judgments made about images of the person that were seen before the experiment, for example on television or in magazines. Alternatively these previous perceptions may be based on observing the person's personality or behaviour in their profession, for example watching a sports presenter on television.

This can be observed most clearly for participants familiar with Face 3 (see Figure 5.3). Face 3 represents an unpopular television celebrity; this is reflected in the attractiveness ratings given by participants who are familiar with him. In comparison to the ratings given by participants unfamiliar with Face 3, it would appear as though familiarity and previous knowledge have affected attractiveness perceptions. In this case the "What is beautiful is good" stereotype (as cited in Dion et al., 1972) is reversed, with negative person perceptions resulting in low attractiveness ratings.

The negative effect of familiarity for Face 3 is interesting as it raises the question of how quickly perceptions of attractiveness can change based on an additional piece of information about a person. For example if participants unfamiliar with Face 3 were informed that he was an unpopular person, would their attractiveness ratings reflect this information, or are negative perceptions gradually developed after multiple exposures to the person?

As shown in Figure 5.4 the effect of changing an image can actually reverse attractiveness preferences. This underlines the previous findings of the effect within-person variability can have on identification and personality judgments: perceptions of a face can change through the simple process of changing the image representing a face. Therefore it is possible to select an image to represent a face that will reverse basic person-specific preferences.

In addition to illustrating how much one face can vary in attractiveness, these findings emphasis the difference between *face* perception and *image* perception. Previous attractiveness research has typically used one image to represent a face (Tracy & Beall 2011; Principe & Langlois 2013) or digitally manipulated new faces from an original to explore different facial structures (Coetzee et al., 2011; Little et al., 2012). Therefore previous research has focused on face perceptions as oppose to perceptions of different images. The results in Experiment 9 illustrate that perceptions can vary significantly within-person. It is therefore important to use multiple images to represent one face in order to gain a more realistic impression of the attractiveness of the face.

Moreover these results highlight the importance to use images that capture within-person variability when comparing perceptions of attractiveness between-person as well as within-person. As shown above by simply changing the image that represents a face, attractiveness perceptions can be reversed. This suggests that comparing faces for attractiveness between-person is not reliable as variations in the images can have a significant effect on the perceived attractiveness.

5.3 Experiment 10: Anchoring effects in attractiveness perceptions

Introduction

Experiment 9 has shown that within-person variability exists in perceptions of facial attractiveness. In Experiment 10 within-person variability is developed further by examining whether perceptions can be influenced by images viewed prior to the attractiveness judgment. Specifically Experiment 10 examines whether viewing images of a face deliberately ordered by attractiveness before making an attractiveness judgment for that face affects perceptions of attractiveness.

In 1946 Asch explored impressions of personality based on presentations of information about a person. For example one experiment investigated how changing the order of verbal descriptions of a person affected first impressions. Participants heard a list of the same six characteristics ordered from positive to negative characteristics or from negative to positive characteristics. Participants were then asked to make a verbal summary of the person described. Asch found that participants' summaries varied dependent on the order the characteristics were presented: if the list began with a positive, the impression of the person was overall more positive than the impression formed if the list began with a negative characteristic. This suggests that the first characteristic is important in creating an impression of the described person. Interestingly Asch explored personality perceptions within-person. In this study a similar idea is examined with varying images of attractiveness.

The presentation of information prior to participants carrying out a task or judgment is also explored in anchoring effects research. Anchoring occurs when a judgment or estimation is made based on an initial value resulting in a biased answer (Tversky & Kahnemann 1974). For example, in an experiment from Tversky and Kahnemann (1974) participants were asked to estimate the percentage of African countries in the United Nations (UN). Prior to answering the participants watched as a "wheel of fortune" was spun, deliberately fixed to land on either the number 65 or 10. Participants were then asked to specify whether the number of African countries in the UN was greater or smaller than the number they had seen on the wheel, and to estimate the actual percentage by going higher or lower than the given number. The median estimation given by participants exposed to the number 65 was 45%, while the median estimation given by participants exposed to the number 10 was 25%. Therefore the 'anchor', the number presented on the wheel spin, significantly influenced participants decision-making.

Considering the concept of anchoring, it could be expected that when shown images of a person ordered by ascending attractiveness or descending attractiveness before being asked to rate a final image of the same person for attractiveness, participants will be influenced by the first image. Therefore it is hypothesised that participants will rate an image as more attractive if they were exposed to images shown in descending order of attractiveness than participants asked to rate the same image after exposure to images shown in ascending order of attractiveness.

Research has shown that long lasting impressions can be made from facial appearance (Willis & Todorov 2006). For example in an experiment from Willis and Todorov (2006), participants' ratings of attractiveness were correlated when participants were presented with face stimuli for a minimal exposure length of 100 milliseconds and an unrestricted length of exposure. Therefore it was concluded that first impressions are not only made surprisingly quickly but are also long lasting. Considering this lasting effect of first impressions, it was hypothesised that perceptions of unfamiliar faces would be influenced by the first image shown, while perceptions of familiar faces would not be influenced by the order of previously presented images. This was hypothesised as having seen images of familiar faces before participants were likely to have already made a first impression of the person. If first impressions are lasting participants' perceptions of familiar faces would not change with experimental manipulation.

Considering the concept of lasting first impressions, the hypotheses can be outlined: it is hypothesised that participants will rate an image of an *unfamiliar* faces as more attractive if they were exposed to images shown in descending order of attractiveness than participants asked to rate the same image after exposure to images shown in ascending order of attractiveness. And, it is hypothesised that attractiveness ratings of images of familiar faces would not be influenced by the order of previously presented images

Method

Participants

Sixty-four participants took part in Experiment 10, 32 Austrian undergraduate students from the University of Vienna and 32 British undergraduate students from the University of Glasgow.

Design and Stimuli

By conducting the experiment in two different locations (Glasgow and Vienna) it was possible to manipulate familiarity by using images of both British and Austrian famous faces as stimuli. The images remained constant therefore every image served in the familiar and unfamiliar condition. This resulted in a repeated-measures design.

The images rated for attractiveness in Experiment 9 were used to create the stimuli for Experiment 10. First, for each face the images were ordered from lowest to highest by the attractiveness ratings given in Experiment 9. The median two images were then removed and the image rated most attractive, least attractive and four images representing the range of attractiveness between the two extremes were selected. This allowed for the images to then be ordered into two sequences: “Ascending Attractiveness” and “Descending Attractiveness”. An example of one face’s set of six images is shown in Figure 5.5. As the images had been used as stimuli in Experiment 9, no further image preparation was required.

One of the two ‘median’ images removed from the original twelve was then randomly selected to act as the final image used as a test image that participants were required to view and rate for attractiveness. Only the unfamiliar ratings of attractiveness from Experiment 9 were used to select these image sets.

The images were then inserted into a computer task using experiment-building software, ePrime. The task was programmed to repeat the instructions given to participants in text form, with a “Start Experiment” button available at the bottom of the screen for participants to press when ready to begin the task. Following this, the images would appear individually on the screen, sized 6 x 9 cms, for 1 second before moving onto the next image in the sequence. The sequence would pause on the final 7th image, at which stage participants were asked to make an attractiveness rating on a scale of 1 – 7, indicating on a seven button-scale below the image. To move onto the next sequence, participants were required to press a button on the scale.

The two different types of sequences made up the two experimental conditions: Ascending Attractiveness and Descending Attractiveness. Participants were randomly assigned to either condition. As shown in Figure 5.5, the images participants viewed remained constant in both conditions: only the order of presentation changed. In addition the test image participants were asked to rate remained constant in both conditions.

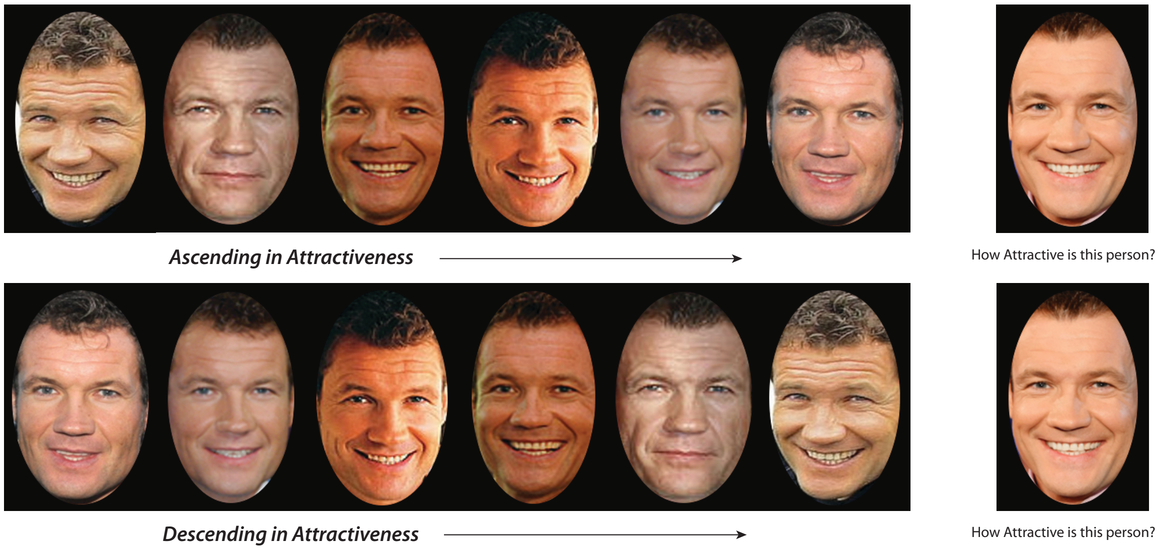


***Figure 5.5*. Experiment 10: Stimuli and design.**
The above sequences show the images of Face 1 in each condition Ascending Attractiveness and Descending Attractiveness (from left to right). The final image remained constant for all participants and was rated for attractiveness on a 7-point scale.

Procedure

Participants were instructed that they would see a series of six images of a face shown one after each other, at the rate of 1 second per image. Participants were instructed to look at each image shown, as they would be required to later make a judgment for each person. The sequence would then stop on a final image of the face (the median test image) and they would be asked to give a rating of how attractive they thought the face was using a 7-point scale. Participants were instructed to indicate their rating by clicking on one of the 7 buttons that made up the 7-point scale, positioned on the screen below the image. Following their response, the experiment would move onto the next sequence of images. Participants were presented with the images either ascending in order of attractiveness or descending in order of attractiveness. Participants were unaware of the ordering. All participants viewed and rated all stimuli therefore rated both familiar and unfamiliar faces.

Familiarity was key in this experiment and so following the experiment participants were presented with a list of the names of the famous faces, along with one image of each famous face, again shown on the computer screen. Participants were asked to indicate their familiarity with each person with a “Yes/No” response by clicking on either a “Yes” or “No” button below the image. Data from participants “unfamiliar” with the appointed familiar faces, or “familiar” with the appointed unfamiliar faces, could then be excluded from analysis. However all participants were “familiar” with all familiar faces, and “unfamiliar” with all unfamiliar faces.

Results

A 2 x 2 repeated measures ANOVA was conducted to analyse the factors of image order and familiarity. Image order has two levels: ascending in order of attractiveness and descending in order of attractiveness, and familiarity has two levels: familiar and unfamiliar faces. A summary of the mean attractiveness ratings for familiar and unfamiliar faces for each image order is shown in Figure 5.6 below.

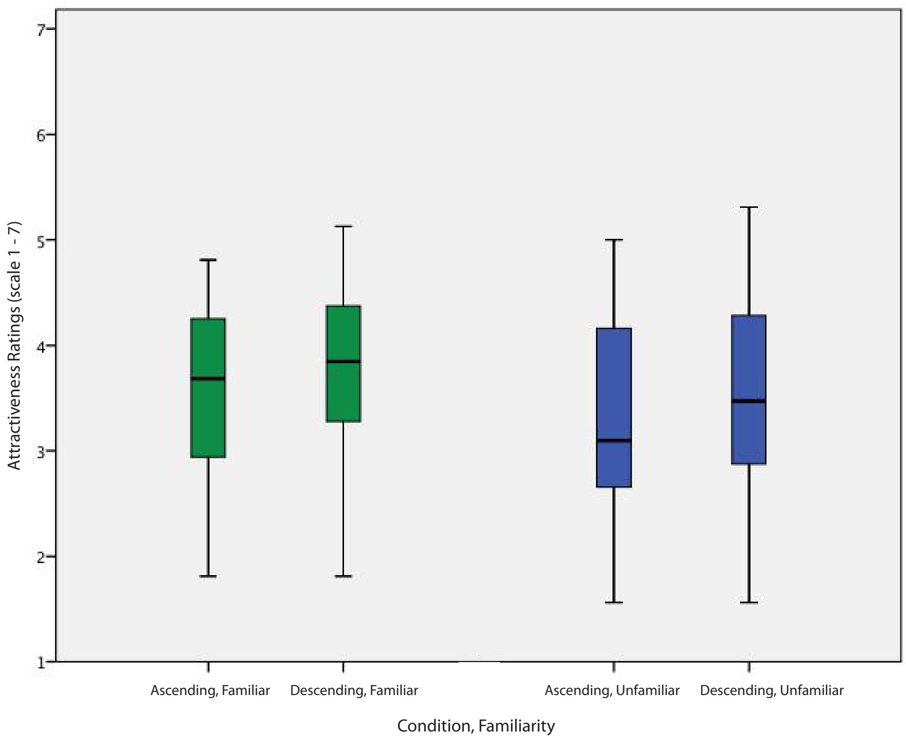


Figure 5.6. **Experiment 10: Summary of results.**
Boxplot illustrating the mean attractiveness ratings for familiar and unfamiliar faces in each condition: Ascending Attractiveness and Descending Attractiveness.

As shown in Figure 5.6, the mean attractiveness rating was higher for images descending in order of attractiveness than images ascending in order of attractiveness. In addition, attractiveness ratings for familiar faces appear to be higher than attractiveness ratings for unfamiliar faces for both image orders.

A significant main effect of image order was found: the mean attractiveness rating was significantly higher for images shown descending in attractiveness than images shown ascending in attractiveness [F (1, 19) = 5.22, $p < 0.05$, η_p^2 = 0.01]. There was no significant main effect of familiarity [F (1, 19) = 0.49, $p >$ 0.05, η_p^2 = 0.00] and no significant interaction between image order and familiarity [F (1, 19) = 0.00, $p > 0.05$, η_p^2 = 0.00].

Therefore viewing images of a face sequenced by attractiveness has a significant influence on the perception of attractiveness of that face. If the images of the face are presented in ascending order of attractiveness, the images will have a negative effect on the perception of attractiveness of that face. If the images are presented in descending order of attractiveness, the images will have a positive effect on the perception of attractiveness of that face. These results also show that the effect of viewing previous images influencing attractiveness perceptions remains constant regardless of familiarity with the face shown.

Discussion

The results of Experiment 10 further highlight the variability of attractiveness perceptions by demonstrating that a simple order effect can influence participants' ratings of attractiveness. By keeping the sequence images and the test image constant but changing the sequence order to either ascending or descending in attractiveness, can significantly effect the perceived attractiveness of the image. It is important to stress that the significant differences in perceived attractiveness occurred with the *same image*.

The findings of Experiment 10 can be closely compared to personality perception findings set out by Asch in 1946. Asch found that when given a list of descriptive characteristics ordered from positive to negative traits participants created an overall more positive impression of the person than when given a list of characteristics ordered from negative to positive traits. These results show a similar effect but with impressions of facial appearance as opposed to verbal character descriptions. If participants are shown the most attractive image of a person first, their attractiveness perception of the person is more positive than participants shown the least attractive image of a person first.

The effect of the ascending and descending attractiveness order could be explained as an anchoring effect as attractiveness ratings differ dependent on the first image viewed. Anchoring effects occur when estimations or judgments are made that are biased towards an initial piece of information. The findings of Experiment 10 are particularly closely comparable to an experiment from Tversky and Kahneman (1974). Participants were asked to quickly calculated an estimation of a multiple sum, the sum was presented in two orders: 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1, or 1 x 2 x 3 x 4 x 5 x 6 x 7 x 8. Participants' answers were significantly larger if they viewed the numbers in descending order than ascending order. Similarly the findings here show a significantly higher attractiveness rating if participants' viewed the images descending order in comparison to ascending order.

A surprising result is that participants' ratings of familiar faces are also affected by the sequence ordering of images prior to attractiveness judgments. It was hypothesised that the ordering of the images would not affect ratings of familiar faces: it was expected that perceptions of familiar faces would already have been made prior to the experiment. These perceptions may have been based on viewing images of the person or having seen the person in context of their familiarity, for example appearing on television. Therefore any experimental manipulation would not affect participants' perceptions of these familiar faces.

However previous research into anchoring effects has found similar results in different areas. For example, in a study from Mussweiler, Strack and Pfeiffer (2000) anchoring effects were found to occur even with expert participants. Participants classed as car experts, either mechanics or car dealers, were approached in a real life setting and asked to estimate the value of a used car. The researcher acting as the owner first gave his estimation of the car's value either 2,800DM or 5,000DM; this acted as the anchor. Despite their previous knowledge with the required task and with additional accurate details about the car, participants were still influenced by the anchor: participants estimated a higher value when initially presented with the estimation of 5,000DM than when presented with the owner's estimation of 2,800DM. This illustrates that anchoring effects can occur when familiar with a task or having prior information about the task. It could therefore be suggested that a similar effect is occurring here with the ratings for familiar images.

The results from the ratings of familiar faces is also interesting when considered in relation to previous research that suggests first impressions are lasting (Willis & Todorov 2006). If a first impression of attractiveness is lasting it would be expected that perceptions of familiar faces could not be manipulated. However in Experiment 10 participants' ratings of attractiveness are influenced by previous image order effects. This suggests that first impressions of a *face* are not long lasting but can be changed by changing an image representing a face. It could be possible that previous research, such as Willis and Todorov (2006), found long lasting first impressions of *images* rather than face or person impressions.

5.4 General Discussion

The results from Experiment 9 and 10 clearly illustrate the variability that can occur for perceptions of attractiveness within-person. Experiment 9 illustrates that when asked to rate multiple images of different faces participants' ratings can vary dependent on the image shown. Experiment 10 illustrates that participants' judgments of attractiveness of an image can vary dependent on a sequence of images of the same face shown prior to the attractiveness judgment. These findings supports the results of Experiment 8 and so together the findings begin to develop a new perspective of how much one person's facial appearance can vary and the effect that has on visual perceptions of their appearance and personality. It is therefore important that within-person variability is considered more in face research as it clearly plays a surprisingly influential role in facial perception.

These findings are also important as they illustrate the need to adequately represent a face through images. Typically face perception research uses one image to represent a face (Tracy & Beall 2011; Principe & Langlois 2013). However the findings here suggest that perceptions can vary and even be reversed by changing an image of a face. Therefore to accurately explore facial perceptions, multiple images should be used to represent one face. This underlines the need for future research to consider within-person variability, and to develop the use of multiple images to represent a face.

Research has demonstrated that many behaviours and judgments can be influenced by facial attractiveness (e.g. Watkins and Johnston 2000). The finding that attractiveness ratings of faces can vary and be manipulated is therefore very important. If changing facial appearance can improve the perception of attractiveness, this can be used to peoples' advantage in a range of different professional and social scenarios. It can also be explored to gain a more complete understanding of what is considered attractive both between-persons and within-persons.

The main conclusion from this chapter is that perceptions of facial attractiveness are not constant or stable but can vary dependent on a simple image change. Furthermore, participants' perceptions can be easily manipulated through the presentation of different images. This suggests that previous findings could be significantly different with the introduction of multiple images that capture within-person variability, and therefore calls for reinterpretation of research that focusing solely on between-person differences in facial attractiveness.

Chapter Six

Image Memory

6.1 Introduction

The previous chapters have illustrated the effect within-person variability can have on identification behaviour and perceptions of personality. Exploring face identification and face perception using images that encompass realistic within-person variability extends previous research in important ways: identification performance is surprisingly poor when within-person variability is taken into account, while personality perceptions vary widely when naturally varying images are rated by participants. Given the importance of within-person variability for these perceptual tasks, this chapter considers within-person variability in the context of face recognition memory.

Previous research into visual memory suggests that we are very good at remembering pictures, places and people over a long period of time (Standing 1973; Bahrick, Bahrick & Wittlinger 1975). Such experimental findings tie in with the personal experiences that we can remember things well from early life such as our teachers from schooldays. However previous research often confuses face memory with image memory therefore reporting good image memory as good face memory. This can lead to a confused impression of memory ability.

Previous studies of face recognition memory

In a study from Standing in 1973 memory for normal and vivid images was tested alongside memory for words. Standing tested participants' memory for stimuli two days following first exposure. Participants viewed between 20 and 10,000 pictures consisting of words, objects and different scenes. The pictures were divided into "vivid pictures" and "normal pictures": vivid pictures were defined as something unusual such as a photograph of a dog smoking a pipe, while a normal picture would be more typical such as a photograph of a dog. Two days following first exposure participants were shown up to 160 picture pairs, consisting of one previously viewed item and one new item. Their task in the test phase was to decided which picture they had seen before.

Accuracy on this task was relatively high, with low rates of recognition errors across participants. Memory for vivid pictures was found to be better than memory for normal pictures or words. For example, participants shown 1000 vivid pictures at first exposure were tested on 80 of the images and made a mean number of 4.8 recognition errors. In comparison participants shown 1000 normal pictures at first exposure were tested on 80 of the images and made a mean number of 9.2 recognition errors. For word memory recognition errors increased to 15.4 when participants were asked to remember 80 words from the 1000 shown at first exposure. Standing only tested memory up to 10,000 images for normal pictures. It was found that participants made a mean number of 27.2 recognition errors when asked to remember 160 images from the 10,000 normal pictures shown at first exposure. From this Standing (1973) concluded that recognition memory for pictures is not only very good but is potentially limitless, and also that pictures are remembered with more ease than words.

In a similar series of experiments, Shepard (1967) tested memory for words and sentences as well as pictures selected deliberately to be of high salience and low similarity to each other. Participants viewed 600 words, 612 sentences and 612 pictures at the encoding phase. Participants' memory was then tested by presenting 60 pairs of words, 68 pairs of sentences and 68 pairs of pictures, and asking participants to indicate which item in the pair they had seen at encoding. Shepard's experiments examined delay between exposure and test in more detail: memory for pictures only was tested over different lengths of delay. Participants were either tested following a delay of 2 hours, 3 days, 7 days or 120 days.

Participants' memory accuracy for words was 88.4% and for sentences was 89%. Participants' accuracy for pictures following a delay of two hours was near perfect at 99%. This dropped to 92% accuracy following a delay of three days, 87% following a week's delay and down to just above chance level following three months at 57% accuracy. The findings illustrate that memory for pictures was better than memory for words and sentences. However memory for pictures reduces to near chance performance after three months. This suggests that there is a decay rate for pictures that results in poor memory following an extended delay.

Shepard (1967) and Standing (1973) explored picture memory, which included pictures of faces but did not focus solely on face memory. In a series of experiments from Goldstein and Chance (1970), memory for faces was compared to memory for non-face stimuli. Goldstein and Chance (1970) argued that, as stimuli that we encounter daily, faces are familiar stimuli relative to the comparison stimuli of inkblots and snowflakes. It was hypothesized that the comparison of memory accuracy for faces versus inkblots and snowflakes would shed light on the role of familiarity in determining memory for particular items.

Participants were shown either 14 images of faces, inkblots or snowflakes in the initial presentation and were told that there would later be a memory task based on these images. Different time delays followed the initial exposure of images: immediately after the initial exposure phase or 2 days following the initial exposure phase. In the memory phase, participants were presented with 84 images, including the original 14 faces, inkblots or snowflakes they had seen, and were asked to indicate for each image whether they had seen that image before. Images in the test phase were presented individually and participants were made aware that all of the original stimuli would be present in the test phase.

Memory for inkblots was close to chance at 51% for participants tested after no delay, dropping to 42.5% for participants following the delay of 2 days. Participants' memory for snowflakes was worse at 36.5% accuracy with no delay and 30% accuracy after the delay of 2 days. In comparison participants' memory for faces was better: participants' memory for faces was 71.5% with no delay and 71.5% following the delay.

Although the performance for face recognition was higher than for participant groups given inkblots and snow crystals as stimuli, it was poorer than the recognition results from the previous two studies, (Standing 1973; Shepard 1967). Goldstein and Chance (1970) focused on images of faces whereas the previous two studies (Standing 1973; Shepard 1967) explore more general picture memory. This suggests that faces are harder to remember. The results from Goldstein and Chance (1970) illustrate that participants' memory for unfamiliar faces was poor but what about memory for familiar faces?

Bahrick, Bahrick and Wittlinger (1975) explored memory for faces that were once highly familiar to participants by using images taken from their high school yearbooks. Participants were asked to identify images of classmates from groups of five images containing one target image. Participants did not have to name their old classmate, just simply remember which face was of an old classmate. Participants were tested at various intervals between 3 months and 570 months (47 years) following their high school graduation.

Participants tested 3 months after graduation were 90% accurate in their memory performance. This finding remains consistent with delays of 9 months to 34 years, dropping only for the final delay length to 71% accuracy after a period of 47 years following graduation. This finding suggests that memory for highly familiar faces is good and can be long lasting. In comparison to participants' performance for faces in Goldstein and Chance's study (1970), memory for familiar faces appears to be very good even after long periods of time.

More recently Brady, Konkle, Alvarez and Oliva (2008) explored to what extent we are able to remember detail in recognition memory. Brady et al., presented participants with images of 2,500 everyday objects such as a loaf of bread. Following a short delay of ten minutes participants memory was tested by presenting 300 image pairs of one previously seen image and one new image, and asked participants to remember which image they had seen before. The image pairs were made up of a target image paired with either an image of a "novel" object, an image of a similar "exemplar" object, or an image of the same object varied in some way, defined as a "state" image.

Participants' performance in the three test conditions was high; memory accuracy in novel image pairs was 92%, for similar image pairs was 88% and for varied image pairs was 87% (see Figure 6.1). These findings illustrate that participants were able to distinguish objects they had seen when paired with images of different objects, very similar objects and in images of the same object varied across one dimension such as pose.

This study extends recognition memory to illustrate that participants can successfully retain detailed memories of a large number of images. Interestingly participants were able to distinguish between varied images of the same object. This suggests that participants were aware of within-object variations and were able to remember what they had seen before and the variations that they hadn't seen before. In the context of within-person variability in faces, this would suggest that in an image memory task with faces participants would be able to distinguish between different images of the same face.

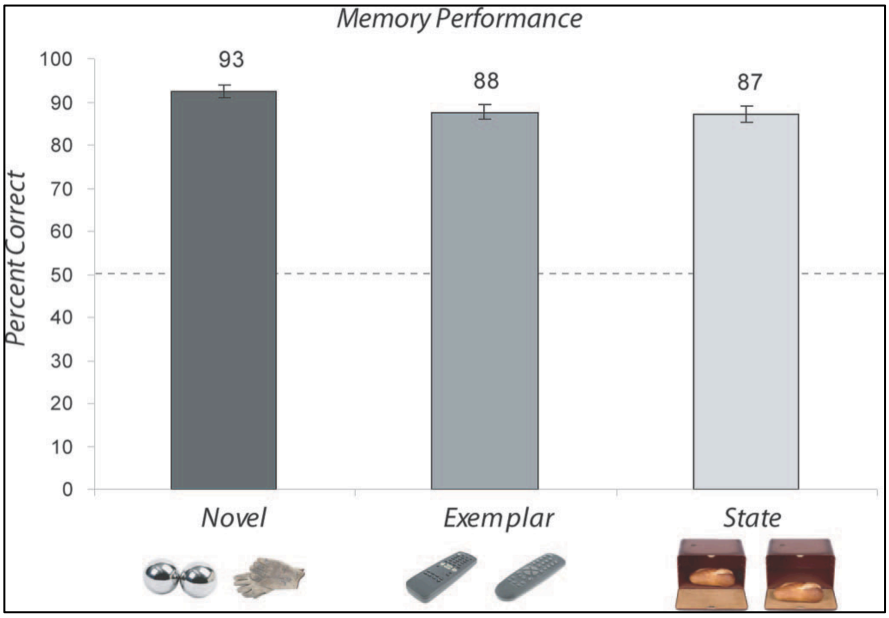


***Figure 6.1*. Within-object memory (Brady et al., 2012).**

More recent research has also continued the examination of recognition memory for faces, for example Bainbridge, Isola, Blank and Oliva (2012) explored what they defined as memory for human face photographs. Bainbridge et al., (2012) argued that it is unclear how object memory can be related to face memory and that memory for faces might be more distinct as faces are more memorable. To test memory for faces Bainbridge et al., compiled a large database of unfamiliar faces (see Figure 6.2 for examples). Conducted as an online experiment, participants were shown up to 3,600 images of different faces. During the exposure to the images randomly selected images would be repeated and it was participants task to indicate if they saw a repeated image by pressing the “r” key on their keyboard.

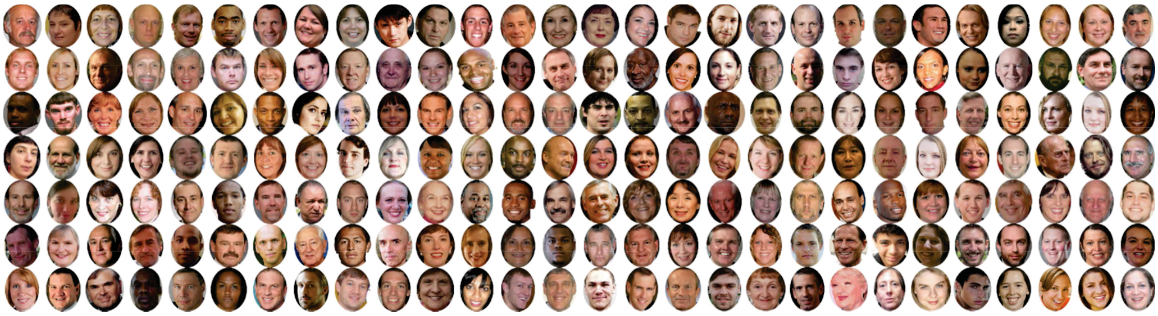


Figure 6.2. **Memory for face photographs (Bainbridge et al., 2012).**

Performance was surprisingly poor: participants' accuracy was just above chance at 53.6%. Therefore participants accurately indicated that they had seen a face before for only just over half of the repeated images. This suggests that it is difficult to remember photographs of unfamiliar faces. This finding supports previous research that has found identification and recognition memory to be poor for unfamiliar faces (Bruce 1982; Bruce et al., 2001; Burton et al., 1999).

The main conclusions that can be made from the studies discussed is that although memory for faces is better than memory for words and objects, memory for unfamiliar faces can be surprisingly poor. Memory for highly familiar faces on the other hand can be long lasting. Two important points can also be summarised: previous research has not explored image memory within-person, and research often confuses image memory with face memory.

Image Memory versus Face Memory

In previous chapters it has been discussed that one image is not an adequate representation of a face. In the context of memory research the use of one image produces a different limitation: the use of one image of a face at first exposure and test phases of a memory experiment does not equate to face memory, instead it is a test of image memory. The difference between face memory and image memory is an important distinction to make as the two concepts are not interchangeable. To investigate face memory different images of a face must be shown at the first exposure and memory phase of the experiment so that participants' task becomes looking past the image characteristics to identify whether it is the same person they previously saw.

A series of experiments from Bruce in 1982 is a clear example of face memory; participants were presented with images of faces at the initial exposure phase and then presented with either the same image or a different image at the test phase. The images presented during the test phase were either unchanged or manipulated for expression, pose, or expression and pose. At the test phase participants were asked, “Have you seen this person before?” thus participants were required to look past the specific image and decide on the basis of identity.

In a typical memory design, participants were shown 24 images of unfamiliar faces at the exposure phase. In a later phase they were shown the same 24 faces (either the same or different image of the faces), intermixed with images of 24 new faces. The images were shown individually and participants were made aware of the memory task prior to starting. Participants were 89.6% accurate when presented with the identical image after a delay of just 15 minutes. Accuracy dropped to 76% when the image at test phase was varied by one dimension (either pose or expression), and decreased further to only 60.5% accuracy when the image varied by two dimensions (pose and expression). Therefore face memory for unfamiliar faces is poor and is worse than image memory, as demonstrated through the use of unchanged images.

In a companion experiment the same procedure was repeated, this time using participants who were either familiar or unfamiliar with the faces in the stimuli. When participants were familiar with the faces accuracy was high, participants were 95.8% accurate when the same image was presented at test and 94.5% accurate when the image was varied. Performance by participants unfamiliar with the faces was similar to the first experiment, with 88.8% accuracy for the same image and only 54.8% accuracy when the image varied. This illustrates the difference between familiar and unfamiliar face performance, when a face is familiar participants were good at face memory, finding it easy to recognise a face they had seen before across differing images. Therefore when it comes to face memory being familiar with the face is beneficial.

Image-Specific Memory

Considering the previous studies is it possible to visualise the areas memory research has explored, and the questions that remain unaddressed within face and image memory (see Figure 6.3). Memory studies such as Standing (1973), Bahrick et al., (1975) and Bainbridge et al., (2012) addressed the issue of image memory by asking participants to remember if they had seen an image previously, and presenting participants with either the same image of a face or a different image of a new face. Meanwhile studies such as Bruce (1982) have addressed the issue of face memory by asking participants to remember whether they had seen a person previously and presenting participants with either the same image of a face or a different image of the same face. However previous memory studies have not addressed image memory using multiple images of the same face.

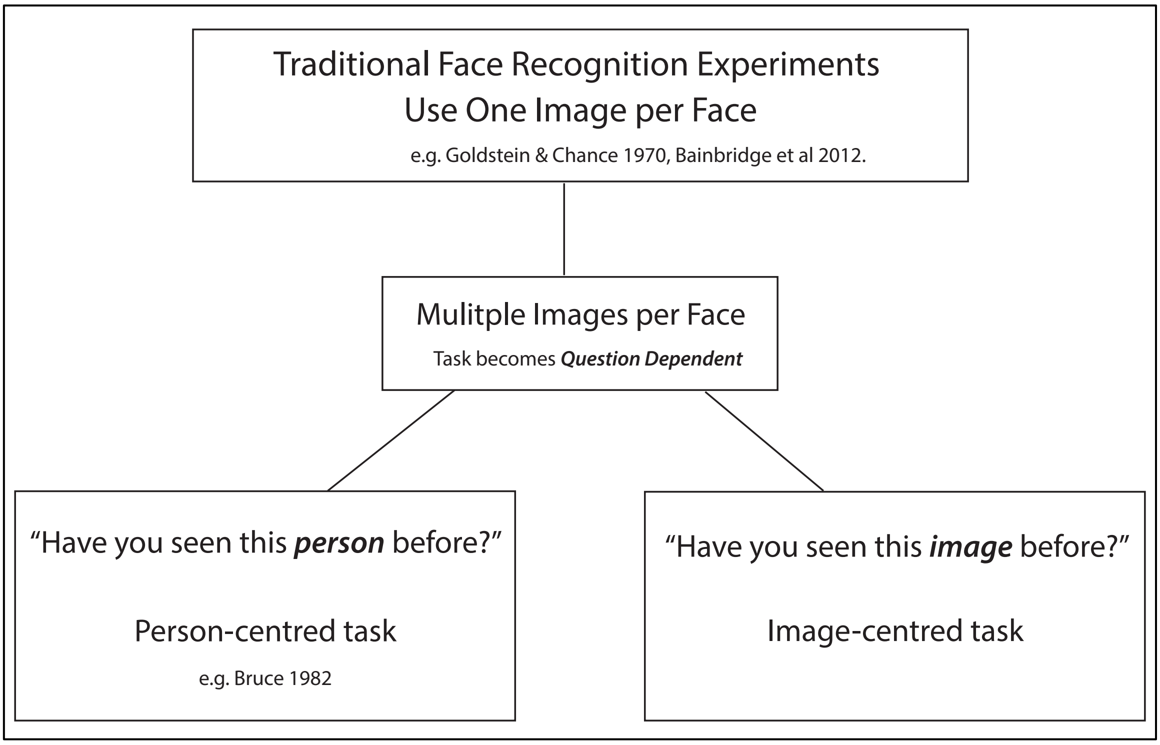


Figure 6.3. **Visual summary of memory recognition research.**
Visual description of memory research to date and what has not been explored.

It is important to explore image memory in the context of familiarity as familiar and unfamiliar face processing has been theorised to be separate processes. Unfamiliar face processing has been described as being strongly image bound (Hancock, Bruce & Burton 2000; Megreya & Burton 2006; Burton & Jenkins 2011). For example Hancock et al., (2000) described unfamiliar face processing as being reliant on low-level image descriptions. It has also been claimed that unfamiliar faces are not processed as faces but rather as pictures (Megreya & Burton 2006; Burton & Jenkins 2011).

In contrast familiar face processing has been described as involving face representations that go beyond the specifics of any particular image (Bruce & Young 1986; Burton, Jenkins, Hancock & White 2005). Bruce and Young's model of functional face processing (1986) describes a stage in familiar face processing involving face recognition units (FRUs), stored representations of any familiar face. When we see a familiar face we can use these FRUs to recognise and identify the person. This concept of having representations of familiar faces illustrates why participants are able to remember familiar faces in experimental conditions such as in Bruce's study (1982); we can use our FRUs of familiar faces to build a more generalised representation of a person's face that goes beyond image dependent descriptions.

If unfamiliar face processing is largely image-based and familiar face processing is less image-based, this could suggest a counterintuitive prediction: memory for specific images might actually be better for unfamiliar faces than for familiar faces. Therefore if a memory task is more image-based, participants might be better at remembering different images of unfamiliar faces than familiar faces. This prediction may appear counterintuitive as in face processing tasks performance is routinely better for familiar faces (Burton, Wilson, Cowan & Bruce 1999; Bruce, Henderson, Newman & Burton 2001). This finding has typically been shown across two different images of a person, such as a matching task with high qualities images and lower quality images or video stills. Similarly, familiarity has been shown to positively affect memory (Bahrick et al., 1975; Bruce 1982).

Given the extent of research findings showing an improvement in face processing when a face is familiar it may be unusual to consider the opposite. However it is possible that being familiar with a face is disadvantageous in a task where attending to a specific image is required. For example, as the current Prime Minister we are frequently presented with images of David Cameron. As we are familiar with David Cameron when we see an image of him we quickly recognise him and no longer require the image visually: instead we process it as a more abstract representation of identity. It could be suggested that in the context of an experiment when shown an image of David Cameron we think, "that's David Cameron I've seen him before" and so we do not retain the image. Given an image of an unfamiliar person in a similar context, this can't happen because the image is all we have to process.

With this in mind, when asked to recall specific images of familiar faces it may be more difficult than if asked to recall specific images of unfamiliar faces. This theory has been illustrated in memory research in a different context: processing x-ray images in radiology. In a study from Myles-Worsley, Johnston and Simons (1988) novices and radiologists, with varying expertise, were tested on their memory for clinically normal and clinically abnormal x-rays. As hypothesised, memory for clinically abnormal x-rays increased as the level of radiological expertise increased in participants. Interestingly however there was a decrease in radiological experts' memory performance for normal x-rays to the extent that novice participants had better memory performance for normal x-rays than the experts (see Figure 6.4). This suggests that comparing experts' memory for familiar stimuli (normal x-rays) to novices' memory for unfamiliar stimuli (normal x-rays), results in better performance when the images are unfamiliar. Applying these findings to faces could indicate that image memory may actually be better for unfamiliar faces than familiar faces.

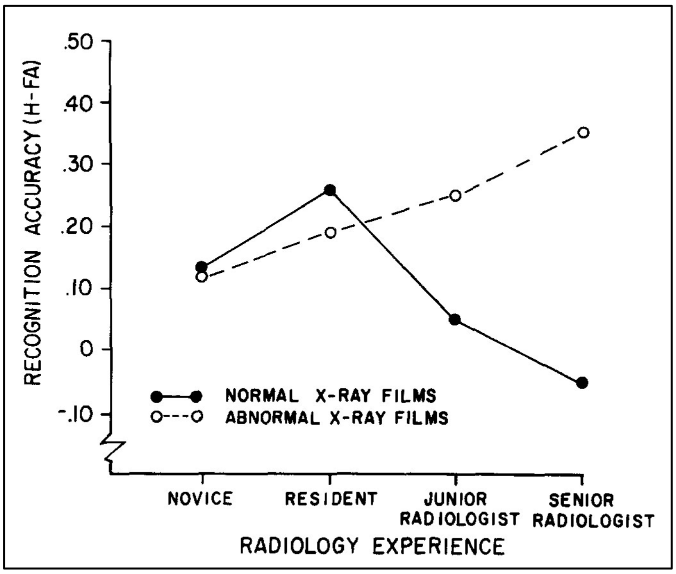


Figure 6.4. **Expert versus novice memory (Myles-Worsley et al., (1988).**
Summary of results from Myles-Worsley et al., (1988) illustrating better memory for unfamiliar stimuli.

Multiple images of the same face have been used in a few face recognition studies, such as Bruce (1982). However the images are usually presented across the exposure and test phase of the experimental task and only two images are typically used. For example one image will be presented at exposure and a different image of the same face will be presented at test. To explore the full extent of image memory this study will present multiple images of a face at first exposure and test participants' memory for each specific image. Therefore the question asked of participants is also important; it must be clearly underlined that participants are required to remember whether they have seen the specific image before as opposed to the face shown in the image. This task is therefore image-centred; the face shown in the image is somewhat irrelevant, the specific image must be remembered (see Figure 6.3).

As outlined above there is a gap in the investigation of image memory, therefore this chapter outlines two experiments that explore the extent of image memory. Based on the findings of previous experiments (e.g. Standing 1973), it is hypothesised that overall image memory for multiple images of the same face will be good. Familiarity is tested by utilising images of famous faces from two different locations: the UK and Australia, and testing participants from each of the locations. It is hypothesised that memory for images of unfamiliar faces will be better than memory for images of familiar faces.

6.2 Experiment 11: Image memory following a long delay

To examine image memory over a long delay the opportunity arose to approach participants who had taken part in an earlier experiment; the 20 participants from Experiment 8 were invited back to participate in Experiment 11 and 16 of the original 20 were able to take part. The length of delay varied due to participants' availability, therefore the length of the delay was between 42 days and 161 days (1.5 months and 5.7 months). Previous research has reported good memory following a delay of up to 47 years for familiar faces (Bahrick et al., 1975) therefore it could be expected that image memory for familiar faces will be good following a long delay. However as described above, it is predicted that memory for multiple images of unfamiliar faces would be better than familiar faces due to the distinct methods of processing familiar and unfamiliar faces.

Method

Participants

Sixteen undergraduate students participated in Experiment 11 in return for a small payment. All participants had previously participated in Experiment 8 (with the same stimuli) approximately two months prior to Experiment 11. This allowed for their previous data to be considered as their initial exposure to the stimuli and their memory for the images to be considered following a long period of delay.

Stimuli and Design

The stimuli comprised of 12 different images of 80 famous faces: 40 UK celebrities and 40 Australian celebrities (960 images in total). Participants had previously viewed six of the 12 images of each famous face (either images 1 - 6 or images 7 – 12 see Figure 6.5) and were now shown all 12 images to test image memory.

As the images had previously been used in Experiment 8, no further image preparation or selection was needed. In a design similar to Experiment 8, the images were inserted into a computer-based task using experiment-building software, LiveCode. The experiment began with an instruction screen with a “Start Experiment” button. Following this, all images were presented at random and individually on the screen. Below the image a “Old” and “New” button was positioned. To move onto the next image, a “Old/New” response was required from participants.

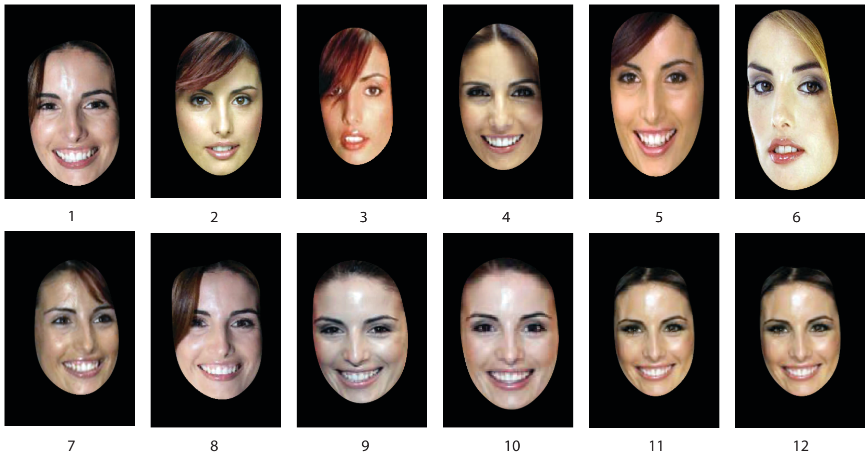


***Figure 6.5*. Images used in Experiment 11.**
Example of 12 different images of one face (Face 21) used to test participants image memory.

Procedure

Participants completed an "Old/New' recognition test: participants were presented with images of different faces and asked whether they had seen each image before. The images were presented on a computer screen, one at a time and in a random order. Participants were asked to indicate if the image was "Old" or "New" by clicking on the relevant button below each image. Following each response, the next image would appear. It is important to stress that participants were asked to indicate whether they had seen each *image* before as opposed to each face or person.

Participants were made aware that there was no time limit and to move onto the next image they were required to make a rating. At three equal intervals throughout the task participants were presented with 'break' screens allowing for them to take an optional break before continuing with the task.

As participants had previously completed a familiarity task at the end of Experiment 8, it was not necessary to establish familiarity with the faces in this experiment.

Results

Participants' mean correct responses (%) were plotted to compare memory performance for old and new images, across familiarity. From Figure 6.6, memory performance is shown to be near chance performance (50%) for both old and new images. In addition there appears to be no difference between memory performance for familiar and unfamiliar faces.

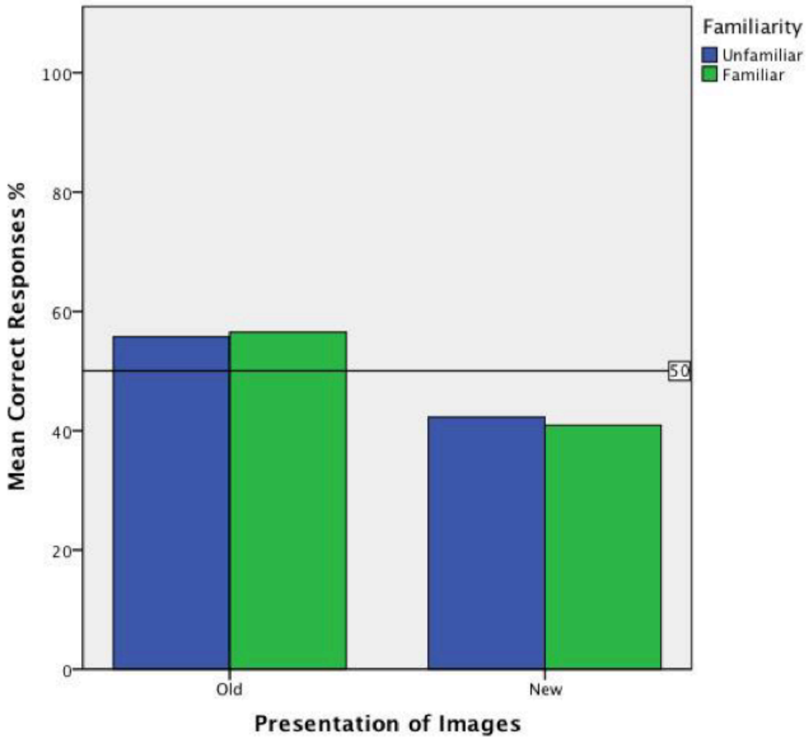


Figure 6.6. **Experiment 11: Image memory performance.**
Histogram of performance illustrating participants' memory accuracy for old and new images across familiarity.

Overall participants mean % accuracy for familiar images was 48.7% and for unfamiliar images was 49%. Considered separately by presentation, there is a clear bias towards responding that an image is old. Participants' mean accuracy for familiar old images was 56.5% and for familiar new images was 40.9%. Participants mean accuracy for unfamiliar old images was 55.7% and for unfamiliar new images was 42.2%. A 2 x 2 within-subjects ANOVA was conducted, with the factors familiarity (familiar, unfamiliar) and presentation (old, new).

No main effect of presentation was found [F (1, 15) = 1.44, p > 0.05, η_p^2 = 0.08], indicating that there was no significant difference in memory performance for old images or new images. There was also no main effect of familiarity [F (1, 15) = 0.08, p > 0.05, η_p^2 = 0.00], indicating that there was no significant difference in memory performance for images of familiar faces or unfamiliar faces. In addition, no significant interaction was found between familiarity and presentation of images [F (1, 15) = 0.10, p > 0.05, η_p^2 = 0.00].

Due to participants availability, the length of the delay between initial exposure to the images and memory test varied by several weeks, with the minimum delay being 42 days and the maximum delay being 161 days. Given this wide range of exposure to test intervals, memory performance and length of delay were examined to analyse whether the duration of the interval influenced test performance. Overall image memory performance as a function of delay is shown in Figure 6.7.

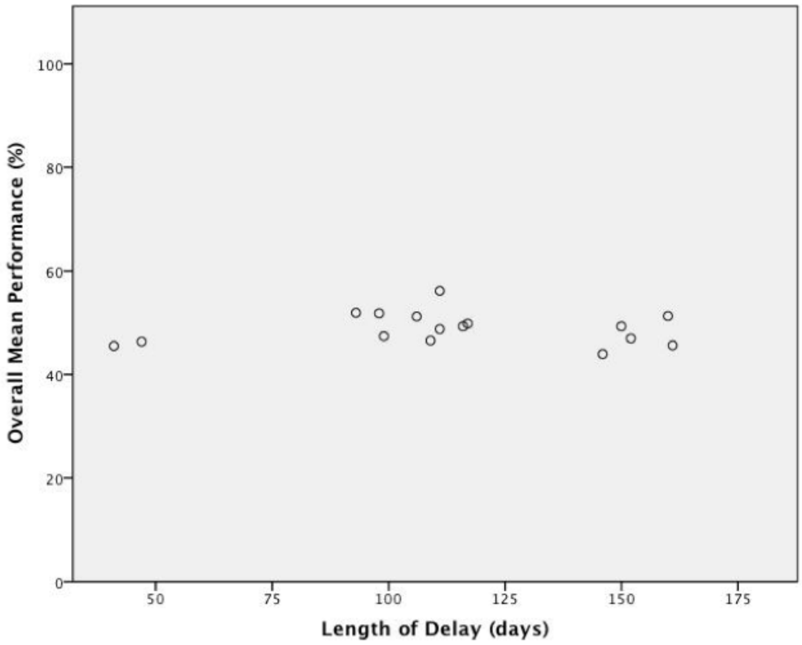


Figure 6.7. **Experiment 11: Effect of length of delay on performance.**
Scatterplot of overall memory performance by length of delay (in days) for each participant.

No significant positive correlation was found between length of delay (in days) and percentage correct performance [r = 0.00, p > 0.05]. This remains consistent when familiar data is considered independently [r = 0.00, p > 0.05] and unfamiliar data is considered independently [r = 0.00, p > 0.05]. Therefore the length of delay from 42 days to 161 days, did not affect memory performance in this task.

Discussion

In Experiment 11 participants were surprisingly bad at remembering images following a long delay. Performance was close to or below chance (50%) for both old and new images, and familiar and unfamiliar faces. Previous memory research has claimed that we are very good at remembering familiar faces over long periods of time, for example, Bahrick et al., (1975) tested participants' memory for familiar faces following a delay of 3 months – to – 47 years and reported accurate memory performance of between 70 – 90 %. The results from Experiment 11 suggest that image memory across multiple images of familiar faces is not good, despite previous research demonstrating that face memory for familiar faces is good.

It could be argued that as this study explores *image* memory, the poor results are not representative of *face* memory ability. This is an important point: image memory for multiple images of the same face has not previously been considered leaving a wide gap in memory research. Exploring image memory on this scale, will allow for a fuller understanding of face memory to be developed.

For example, the findings of Experiment 11 show no difference in image memory for familiar and unfamiliar faces. Reflecting on studies such as Bahrick et al., (1975), which report good long-lasting memory for highly familiar faces, and recognition studies such as Bruce et al., (2001) which highlight the superiority of familiar face recognition, this finding is particularly surprising. Perhaps there is a difference in the process of long-term multiple image memory that makes the advantage of familiarity irrelevant? This could occur as multiple images of familiar faces are processed into one representation of the person, and not as individual images. Therefore as previous research has suggested that unfamiliar faces are processed largely through pictorial codes (Megreya & Burton 2006; Bruce & Young 1986; Burton, Jenkins, Hancock & White 2005), an advantage for unfamiliar faces could have been expected but was not observed in this experiment.

6.3 Experiment 12: Image memory following a short delay

In Experiment 11 participants' memory for multiple images of the same face was surprisingly poor even with images of familiar faces. In an attempt to discover whether we have any memory for multiple images of the same face, the same experiment was conducted with one change; the length of delay between first exposure to the images and the memory test was reduced. It was predicted that overall performance would improve, and again image memory for unfamiliar faces would be better than image memory for familiar faces.

Method

Participants

Twenty undergraduate students recruited from the School of Psychology subject pool participated in Experiment 12 in return for a small payment (age range 18-23 years; 3 male: 17 female).

Stimuli and Design

As in Experiment 11, the stimuli comprised of 12 different images of 80 famous faces: 40 UK celebrities and 40 Australian celebrities, 960 images in total (see Figure 6.1 for an example). Participants completed an initial exposure task viewing half of the images, and returned the following day to complete a memory task with all images.

As the images had previously been used in both Experiment 8 and 11, no further image preparation or selection was required. Participants completed the same computer-based task used in Experiment 8, in order to act as initial exposure to the images. A follow-up task, as used in Experiment 8, established participants' familiarity with the faces shown.

For the memory task, the same computer-based task in Experiment 11 was used. The experiment began with an instruction screen with a "Start Experiment" button. Following this, all 960 images (half viewed the previous day, half new to participants) were presented at random and individually on the screen. Below the image a “Old” and “New” button was positioned. To move onto the next image, a “Old/New” response was required from participants.

Procedure

Participants were asked to complete Experiment 12 over two consecutive days. On Day 1 participants viewed 480 images and were asked to rate each individual image, as opposed to each identity, for the personality traits trustworthiness and dominance on a 7-point scale (see Chapter 4). The images were presented on a computer screen, individually and in a randomised order. Following each response, the next image would appear on the computer screen. This stage acted as an initial exposure phase to the images.

On Day 2 participants completed an “Old/New’ recognition test: participants were presented with all images and asked whether they had seen each image before. The images were presented on a computer screen, one at a time and in a random order. Participants were asked to indicate if the image was “Old” or “New” by clicking on the relevant button below each image. Following each response, the next image would appear. It is important to stress that participants were asked to indicate whether they had seen each *image* before as opposed to each face or person.

Participants were made aware that there was no time limit and to move onto the next image they were required to make a rating. At three equal intervals throughout the task participants were presented with ‘break’ screens allowing for them to take an optional break before continuing with the task.

Familiarity is a key element of this experiment. To establish participants’ familiarity with the faces, following the initial exposure task an additional familiarity task was conducted. Participants were shown an image of each face again along with the name of the famous person, on the computer screen, and were asked to indicate with a “Yes/No” response whether they were familiar with each person, by clicking on either a “Yes” or “No” button below the image. Participants’ responses could then be analysed and if any of the appointed familiar faces were “unfamiliar” or any of the appointed unfamiliar faces were “familiar”, data could then be excluded. All participants were familiar with the appointed familiar faces and unfamiliar with the appointed unfamiliar faces and so all data was analysed.

Results

Participants' mean correct responses (%) were plotted to compare memory performance for old and new images, across familiarity. In Figure 6.8, memory performance is shown to be near chance performance for old and new images. In additionally there appears to be more of a difference between memory performance for images of familiar and unfamiliar faces.

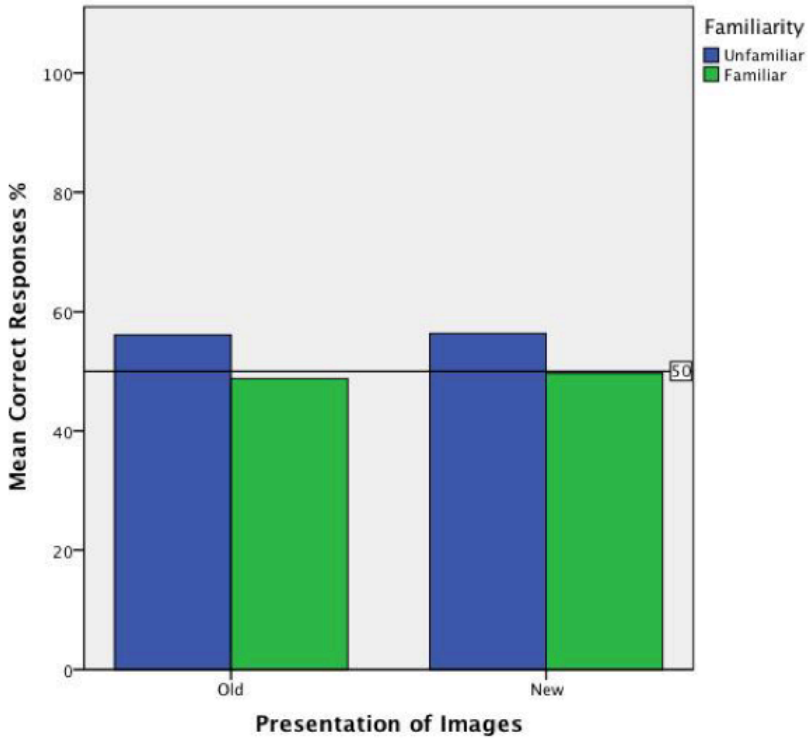


Figure 6.8. **Experiment 12: Image memory performance.**
Histogram of performance illustrating participants' memory accuracy for old and new images across familiarity.

Overall participants mean % accuracy for familiar images was 49.1% and for unfamiliar images was 56.2%. Considered separately by presentation, participants mean accuracy for familiar old images was 48.7% and for familiar new images was 49.6%. Participants mean accuracy for unfamiliar old images was 56.1% and for unfamiliar new images was 56.3%. A 2 x 2 within-subjects ANOVA was conducted, with the factors familiarity (familiar, unfamiliar) and presentation (old, new).

No significant main effect of presentation was found [F (1, 19) = 0.00, $p > 0.05$, $\eta_p^2 = 0.00$] indicating that there was no significant difference between memory performance for old and new images. A significant main effect of familiarity was found [F (1, 19) = 14.29, $p < 0.01$, $\eta_p^2 = 0.11$] indicating that there was a significant difference between image memory for familiar and unfamiliar faces; memory for images of unfamiliar faces is significantly higher than memory for images of familiar faces. No significant interaction was found between familiarity and presentation [F (1, 19) = 0.00, $p > 0.05$, $\eta_p^2 = 0.00$].

Due to participants' availability across the two days, the length of the delay between initial exposure and memory test varied between 21 – 29 hours. As in Experiment 11, memory performance and length of delay were examined to analyse whether the duration of the interval influenced test performance. Overall image memory performance as a function of delay is shown in Figure 6.9.

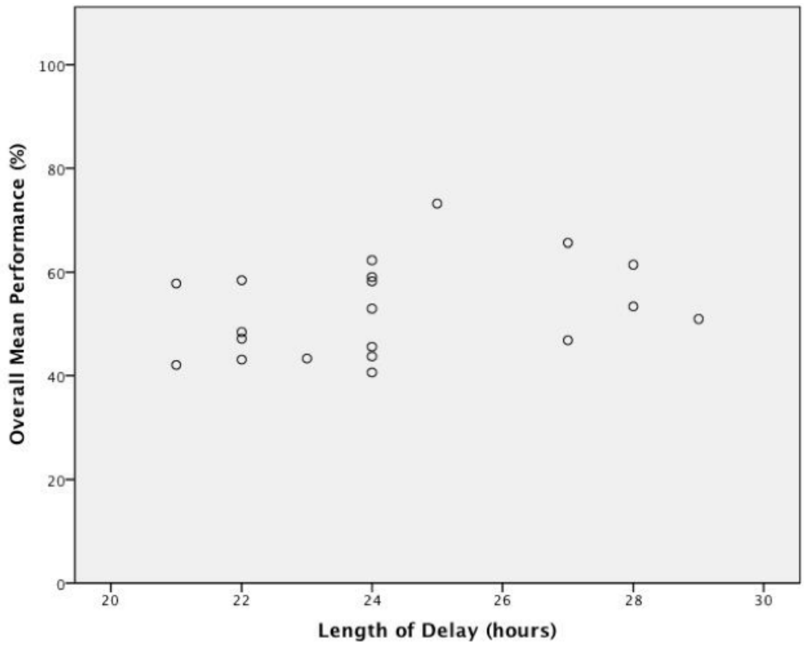


***Figure 6.9*. Experiment 12: Effect of length of delay on performance.**
Scatterplot of overall memory performance by length of delay (in hours) for each participant.

No significant correlation was found between length of delay (in hours) and percentage correct performance [$r = 0.30$, $p > 0.05$]. This remains consistent when familiar data is considered independently [$r = 0.14$, $p > 0.05$] and unfamiliar data is considered independently [$r = 0.08$, $p > 0.05$]. Therefore the difference in hours of the length of delay did not affect memory performance.

Discussion

In Experiment 12 participants are surprisingly bad at remember images following a short delay of just 21 – 29 hours. Similar to the findings in Experiment 11, image memory was close to chance for new images and only just above chance for old images. These poor results are particularly surprising given the short delay of an average of 24 hours between participants first viewing the images and being asked to remember the images. Previous studies have shown a small decay in image memory following short delays of 2 hours or 3 days (Shepard 1967; Goldstein & Chance 1970), but not on a similar level as the poor accuracy in Experiment 12. Considering the differences between this study and previous memory research, i.e. exploring image memory with *multiple* images of the same face, it could be suggested that perhaps we are simply bad at remembering images in this way and on such a large scale.

More importantly however, Experiment 12 also produced interesting results when comparing familiarity conditions: participants were significantly better at recognising images of unfamiliar faces compared with familiar faces. Although memory performance is still low (see Figure 6.8), this suggests that participants were able to make a clearer distinction between images of unfamiliar faces that they had seen before than images of familiar faces they had seen before. This is an unusual result as being familiar with a face helps face recognition in many other tasks. (Burton et al., 1999; Bruce et al., 2001; Bruce 1982). The important difference here is that the task required image-specific memory.

This result is strongly reminiscent of the findings of Myles-Worsley et al., (1988). Myles-Worsley et al., showed that expert radiologists memory for normal x-ray images was worse than novices' memory for normal x-ray images. Therefore unfamiliar (novice) memory was actually better than familiar (expert) memory. The results of Experiment 12 can be directly compared to Myles-Worsley et al.,'s findings (1988).

This familiarity finding can also be explained by the concept of having mental representations, or face recognition units (FRUs) for familiar faces (Bruce & Young 1986) that are quickly accessed when presented with an image of a familiar face, resulting in the specific image shown not being attended to in great detail. As we don't have stored mental representations of unfamiliar faces, it could be suggested that we process each image of an unfamiliar face as a predominantly visual code and therefore may be able to remember each individual image better. This idea is reflected in the results of Experiment 12. This is an important finding as it provides a rare counter example to the usual finding that familiarity improves performance in face tasks. While familiarity undoubtedly benefits *face* recognition when the image is task-irrelevant, familiarity can apparently harm *image* recognition when the identity of the face is task-irrelevant.

6.4 General Discussion

In Experiment 11 following a long delay, image-specific memory i.e. image memory for multiple images of one face, was found to be very poor. Participants' memory accuracy was at 50% chance level for both previously seen "old" images and "new" images. Surprisingly participants performed no more accurately for familiar faces in comparison to unfamiliar faces. In Experiment 12 the task remained the same but the delay was reduced to a short delay of 24 hours. Participants' memory accuracy was still surprisingly poor. However interestingly in Experiment 12 participants' memory performance for unfamiliar faces was significantly better than their memory performance for familiar faces. Overall the findings of Experiment 11 and 12 demonstrate how difficult it is to remember images within-person, and the findings of Experiment 12 suggest that it is easier to remember different images of unfamiliar faces than familiar faces.

Experiment 11

In Experiment 11 it was shown that participants found it difficult to remember multiple images of different faces, performing just below chance level (across old and new presentations) at 48.7% accuracy for images of familiar faces and 49% accuracy for images of unfamiliar faces. As previous research has not explored image memory within-person, a direct comparison cannot be made with this study. However object and face memory research has examined image memory over long delays. For example, Shepard (1967) showed that participants' image memory for 612 pictures was just above chance at 57% following a delay of 120 days from first exposure. The findings here support

Shepard's results and together the results suggest that remembering large numbers of images is very difficult over long periods of delay.

The findings here are comparable to Shepard's findings, however it was expected that image memory for familiar and unfamiliar faces would differ. In comparison, Bahrick et al., (1975) demonstrated that participants' memory for familiar faces (old classmates) following the same approximate length of delay (3 months) was very good at 90% accuracy. However in Experiment 11 there was no difference between memory for images of familiar or unfamiliar faces. This suggests that image memory within-person is perhaps a different process and more difficult to do, for both familiar and unfamiliar faces, in comparison to traditional image memory tasks.

Experiment 12

In Experiment 12 the delay experienced by participants was reduced substantially to just 24 hours so that exposure to the stimuli and participants' memory test were conducted on two consecutive days. Despite this performance did not changed and actually got worse for familiar faces in comparison to Experiment 11; participants' memory accuracy (across presentation) was 49.1% for images of familiar faces and 56.2% for images of unfamiliar faces. Previous research has not explored image memory within-person, however object and face memory research often examines memory following a short delay of one or two days. For example, Goldstein and Chance (1970) tested participants' image memory for faces following a delay of two days and found 71.5% accuracy. Comparing this result with the findings of Experiment 12, a likely conclusion is that the image memory required in Experiment 11 and 12 is more difficult than image memory as examined by previous research such as Goldstein and Chance (1970).

The results from Experiment 12 demonstrated that even following a short delay image memory within-person is very poor. More importantly the results from Experiment 12 have shown that participants' memory for images of unfamiliar faces was better than memory for images of familiar faces. Participants accurately remembered 56.3% of images of unfamiliar faces that they had seen before, and accurately remembered 48.7% of images of familiar faces they had seen before. This difference was found to be significantly difference, indicating that there was an advantage for unfamiliar faces.

This finding is a clear illustration of the differences between familiar and unfamiliar face processing. Previous research has demonstrated the difference in performance with familiar and unfamiliar faces in face experiments such as identification and recognition memory tasks (Bruce 1982; Burton et al., 1999; Bruce et al., 1999; 2001; Clutterbuck & Johnston 2004) and different theories have suggested differences in familiar and unfamiliar face processing (Hancock, Bruce & Burton 2000; Megreya & Burton 2006; Burton & Jenkins 2011). For example Hancock et al., (2000) propose that unfamiliar face processing is strongly image-bound while Megreya and Burton (2006) suggest that unfamiliar faces are not processed as faces but as pictures. Meanwhile familiar face processing has been suggested to be reliant on face representations that abstract away from specific image features and instead comprise of a more general representation of a face (Bruce & Young 1986; Burton et al., 2005).

The findings of Experiment 12 support the concept that unfamiliar face processing is dependent on the image as when participants were required to remember specific images of faces, they were better at remembering images of unfamiliar faces than images of familiar faces. If familiar faces are not processed in the same way but instead involves 'matching' images of familiar faces to their stored mental representation then it follows that participants would struggle to differentiate between multiple images of the same face. Therefore Experiment 12 not only supports face processing theories, but also shows that this difference in processing can be beneficial for unfamiliar faces in the context of within-person image memory.

Future Research

Memory for multiple images of the same face has not previously been explored in memory research and the results from Experiments 11 and 12 indicate that participants find it surprisingly difficult across familiarity. It would be interesting to develop this research finding to examine within-person image memory further, such as using a more sensitive measure of image memory through repetition priming.

Repetition priming research has been demonstrated with images of faces and has shown repetition priming can occur with different images of the same faces (Bruce & Valentine 1985). Research has also demonstrated the long lasting effects of repetition priming with pictures and object drawings (Cave 1997; Mitchell 2006). It would be interesting to explore within-person memory accuracy through repetition priming and to explore whether there is a long lasting effect. Experiment 11 found poor image memory following a long delay, however perhaps the volume of faces used was overwhelming for participants. Future research could reduce the number of faces but retain the number of images of each face therefore developing the concept of within-person image memory.

Examining image memory for multiple images of the same face is a new concept in memory research. From the results of Experiment 11 and 12 showing participants' poor overall performance and especially the better performance with unfamiliar faces in comparison to familiar faces in Experiment 12, it can be concluded that image memory within-person can give an insight into how images of faces are processed. It is therefore important that this concept is developed in future memory research. From reviewing previous memory research it is also important that future research considers not just image memory within-person, but also the distinction between image memory and face memory. This will allow for more accurate conclusions to be made from memory performance.

Chapter 7

General Discussion

7.1 Summary of Results and Key Themes

Chapter 2: Within-person variability in face identification

Chapter 2 focuses on the effect of within-person variability on face identification. A stimuli set was built consisting of 20 different images for each of two different faces. Participants were required to sort these images by identity. In Experiment 1, participants were unaware of how many identities existed within the images and reached a mean number of 9.7 different identities. This finding clearly illustrates the difficulties of mapping naturally varying images onto identities. Following the task it was established that some of the participants were familiar with Face 2; their performance was better for Face 2 than for Face 1. Consistent with previous research, this suggests that being familiar with a face helps improve identification across within-person variability. However, being confronted with the range of within-person variability that can exist for one face (Face 2) did not help participants to cope with variability in another face (Face 1).

Analysis of identification errors revealed that participants made more identity-split errors than identity-merge errors. This suggests that identification difficulties came from being unable to group images of the same face together, as opposed to being unable to split images of two different faces apart. In Experiments 2 and 3 the experimental design was modified to establish whether participants could sort the images reliably when identity-split errors were precluded.

In Experiment 2 participants were given two constraints: firstly, they were informed that two identities were present in the image set and secondly, they were given a reference image of each identity (a current passport photograph) to indicate their appearance. Unsurprisingly all participants sorted the images into two identities. However the number of identity-merge errors increased relative to Experiment 1. In Experiment 3, participants were given just one constraint: they were informed that two identities were present amongst the images, but this time no reference images were provided. Again all participants sorted the images into two identities as expected. Participants continued to make identity-merge errors. Interestingly the frequency of identity-merge errors was actually lower when the reference image was absent (Experiment 3) than when it was present (Experiment 2). This suggests that providing passport photographs as visual guidance actually made the task harder.

Chapter 3: Between-person similarity in face identification

The experiments in Chapter 2 were all based on the same set of stimuli. In Chapter 3, new faces were introduced to examine the generality of identity-split and identity-merge errors. Combining items from different stimulus sets also allowed for within-person variability as a function of between-person variability to be examined.

In Experiment 4 participants were given forty images of two identical twins (Face 3 and Face 4) and were asked to sort the images by identity. Half of participants were unfamiliar with these faces and half were familiar with the faces. It might be expected that participants who were unfamiliar with the identical twins would perceive the images to be of the same person. As it turned out unfamiliar participants perceived 6 different identities from the images on average. Familiar participants correctly determined that only two faces were present in the set but made just as many identity-merge errors as unfamiliar participants. This suggests that participants who were familiar with the twins may have used semantic knowledge to resist making identity-split errors but had trouble assigning images to the right identities, which was solely a visual task.

In Experiment 5 participants were given forty images of two very dissimilar faces, one younger male face from Experiment 1 and one older male face from Experiment 4, and were asked to sort the images by identity. By manipulating the between-person differences so that the two faces were very dissimilar, it was possible to observe whether the within-person variability would result in fewer identity-split errors. Overall participants made a mean number of 4.7 different identities suggesting that sorting images into identities is still difficult even with highly dissimilar faces. In comparison to previous participants' free sorting performance with the same faces (in Experiments 1 and 4), the number of identities made for Face 2, 3 and 4 in Experiment 5 were significantly less than in the previous experiments. This suggests that identification improved to a certain extent with the manipulation of between-person variability. As improvement was found for only three of the four faces used as stimuli in Experiment 5, this adds further support to the theory that within-person variability may be learnt on a face-by-face basis.

In Experiment 6 the issue of between-person similarity for the two faces was avoided by reducing the number of identities in the stimuli to just one face. Participants were given twenty images of one face and were asked to sort the images by identity. Participants made a mean number of 5 identities: directly comparable to the number of perceived identities made (per identity) in Experiment 1. This confirms that the effect of within-person variability can occur when applied to just one person and is not an artefact of between-person similarity.

In the final image sorting experiment (Experiment 7), participants were again given twenty different images of just one face. This time, the images were randomly selected from a face database (Multi-PIE), created to capture within-person variability across pose, illumination and expression (Gross et al., 2010). It was hypothesised that if the database successfully captured within-person variability, the results of an image-sorting task with images from the database would be similar to results in Experiment 6. Participants perceived 3.3 identities per actual identity on average, significantly less than the number of identities perceived by participants in Experiment 6. This suggests that identification was somewhat easier for the database images than for the ambient images used in the preceding experiments. The implication of this is that samples of ambient images encompass a wider range of within-person variability than was captured in the Multi-PIE database and this is important for identity tasks.

Together Chapters 2 and 3 illustrate the difficulty of identifying faces across within-person variability. The effect of within-person variability on identification performance clearly demonstrates the importance of incorporating within-person variability analysis into future research and suggests that image sampling should be considered very carefully when designing experimental stimuli. Applied implications of these findings are discussed in a later section.

Chapter 4: Within-person variability in social dimensions

In Chapter 4 within-person variability is explored in the perception of the social dimensions trustworthiness and dominance. In Experiment 8 participants were asked to rate multiple images of different faces for the two personality traits. The results showed that within-person variability in appearance can strongly affect perceptions of personality. Indeed within-person variability often exceeded between-person variability in its range. Participants rated both familiar and unfamiliar faces; it was expected that less variability would occur within images of familiar faces due to impressions of the faces having already been formed through previous exposure to the person. In fact no difference in variability was found between ratings for familiar and unfamiliar faces for the trait trustworthiness, and only a small number of faces were rated as more variable for the trait dominance when they were unfamiliar to participants. Therefore within-person variability in appearance affected personality perceptions even when participants were familiar with the faces they were rating.

These findings suggest that within-person variability will have to be incorporated into social perception research to deliver a complete theory of personality perception from facial appearance. It also draws attention to the traditional use of just one image to represent a face in experimental stimulus, as simply presenting a different image of a face can radically alter impressions of that person. This emphasises a recurrent theme in this thesis that a single image is not an adequate representation of a face.

Chapter 5: Within-person variability in facial attractiveness

Chapter 5 examines within-person variability in the context of facial attractiveness. In Experiment 9 participants were given multiple images of different faces to rate for attractiveness. As in Experiment 8, perceptions were strongly influenced by within-person variability. Simply presenting different images of the faces could reverse participants' preferences for the faces.

Experiment 10 examined first impressions and anchoring effects in perceptions of facial attractiveness. Participants were shown images of the same face in either ascending or descending order of attractiveness. They were then asked to give an attractiveness rating for that person, based on a reference image of the face, which remained constant in both conditions. The findings showed that the order of the sequence had a significant effect on the perceived attractiveness of the face. Specifically attractiveness ratings were higher when the images were presented in descending order of attractiveness (high to low) than when the images were presented in ascending order of attractiveness (low to high). This finding suggests an anchoring effect whereby the first exposure to a person disproportionately affects impression formation. Surprisingly this effect was found even when participants were familiar with the faces they were rating. The findings of Experiment 9 and 10 clearly imply that perceived attractiveness is not a stable attribution.

Chapter 6: Image memory

In Chapter 6 a novel face memory task was introduced based on memory for specific images of familiar and unfamiliar faces. In Experiment 11 participants who had taken part in Experiment 8 were invited back to complete an incidental memory task for the images they had rated in Experiment 8. Participants were shown all of the images they had seen (at least 6 weeks earlier) in Experiment 8, intermixed with the same number of new images of the faces. Participants were asked to indicate with a yes/no response whether or not they had seen each image before. The results indicate that image specific memory was extremely poor over a long delay, for both familiar and unfamiliar faces.

To investigate whether the poor image memory seen in Experiment 11 could be attributed to the long exposure-test interval, Experiment 12 repeated the procedure with a shorter delay. Participants were asked to complete the experiment over two consecutive days resulting in a mean delay time of 24 hours. On the first day participants were presented with the same experimental task as Experiment 8, rating multiple images for different personality traits (trustworthiness and dominance). This ensured that participants were attending to the images that were presented to them. On the second day, participants viewed these images again intermixed with new images of the same faces, and were asked to indicate whether they had seen each image before. Overall performance was still poor, but importantly, image memory was significantly better for unfamiliar faces than for familiar faces. This pattern of results is consistent with the idea that familiar faces are coded in representations that abstract away from specific visual images, for example a face recognition unit (FRU) or person identity node (PIN) as described by Bruce and Young (1986). In contrast, images of unfamiliar faces remain bound to visual codes.

7.2 Main Themes

Within-person variability

The twelve experiments in this thesis examined within-person variability in the context of face identification, face perceptions of personality, perceptions of facial attractiveness and image memory. Based on the findings it can be concluded that within-person variability should be a major concern across all the topics discussed. Previous research has almost entirely ignored within-person variability. Instead research has looked at between-person variability, for example in face perception research comparing a single image of one face to a single image of another face (e.g. Todorov et al., 2005). As discussed in a review from Burton (in press), and as concluded from the findings here, to progress in our understanding of face recognition, within-person variability must be acknowledged and integrated into both our theorising and research methods.

Although some previous studies have asked how well participants can recognise a face across different images (Bruce 1982), the images have typically been controlled to isolate changes in set expressions or viewpoints. Again, as discussed in a review from Burton (in press), within-person variability is grossly underestimated when restricted to such controlled manipulations. As the comparison between Experiment 6 and Experiment 7 emphasises, within-person variability consists of much more than just systematic changes in pose, illumination and expression. As discussed in Chapter 1, within-person variability can be classified into three distinct categories: *1)*. Variations in face characteristics, *2)*. Environmental variations, and *3)*. Camera variations. Ambient images of faces, as used in this thesis, can capture variability that falls into each of these categories. Therefore to capture real within-person variability, less limited and less controlled methods of image sampling will need to be developed.

To ignore within-person variability would be to overlook an important factor of face processing, akin to ignoring between-person variability. In this thesis within-person variability has been shown to have a huge effect on different areas of face processing, including face processing with familiar and unfamiliar faces. Therefore to gain a more complete understanding of face processing, within-person variability must be taken seriously.

Familiarity

Another theme running through this thesis is the relationship between within-person variability and familiarity. Familiarity is explored in all chapters in relation to face identification, perception and memory. Previous research has underlined many behavioural distinctions between familiar and unfamiliar face processing (Burton, Jenkins & Schweinberger 2011; Carbon 2008; Megreya & Burton 2006; Bruce 1982). Among the most striking is the ability of observers to recognise a familiar face across a wide range of images. The image-sorting experiments in Chapter 2 and 3 illustrate this ability in a novel task. Intriguingly, the only situation in which performance was poor for familiar faces was in Experiment 12, which specifically required image-specific processing. This finding complements the existing distinction between familiar and unfamiliar face processing. Not only does familiarity improve performance in tasks that involve abstracting over different images, it can actually make performances worse in tasks that depend on image-specific processing.

Interestingly familiarity had a less striking impact on tasks that did not involve identity. Ratings of trustworthiness and dominance (Chapter 4) and facial attractiveness (Chapter 5) exhibited wide within-person variability for familiar and unfamiliar faces alike. In summary, familiarity with a particular face reduces the influences of within-person variability on judgments of identity but not for judgments of trustworthiness, dominance or attractiveness. This separation mirrors the structural separation between changeable and non-changeable aspects of the face in cognitive and neuroscientific models of face processing (e.g. Bruce & Young 1986; Haxby, Hoffman & Gobbini 2000).

The difference between a face and an image of a face

A key theme that has emerged throughout the experiments is that an image is not always an accurate or adequate representation of a person or their face. Firstly it is important to make a clear distinction between what is a face and what is an image. It may sound like an obvious difference however it can often be overlooked and confused. For example, in the context of face memory if a participant is shown an image of a face and later shown the same image and asked to remember if they have seen that face before, this is an example of image memory not face memory (e.g. Goldstein and Chance 1970). A test of face memory would be to differ the images showed at exposure and test, testing participants' memory for the face across different images (e.g. Bruce 1982). It is important that this distinction is clarified and is consistent in research.

Returning to the idea that an image is not a good representation of a face, the results from experiments 8, 9 and 10 illustrate the importance of an image in social perceptions. Perceptions can be formed on facial images alone but can also be manipulated dependent on the image. In Experiment 2 using passport photographs as reference images was found to increase identity-merge errors i.e. identifying an image with the wrong passport photograph. Therefore the importance of a good representation for identification purposes is also emphasised. Overall this theme raises two important points: a clear distinction between face and image in experimental tasks is needed, and images are not always an adequate representation of a face. Both of these points are important for interpreting previous research.

7.3 Extension of Previous Findings

Chapters 2: Within-person variability in face identification, and Chapter 3: Between-person similarity in face identification

Experiment 1 continues on from a study from Jenkins et al., (2011) in which participants struggle to identify faces when presented with images that illustrate the extent of within-person variability. Experiment 1 replicates the findings of Jenkins et al., (2011) and also extends the finding by showing that the effect of within-person variability occurs with different faces. With this base level of knowledge about identification ability, the six subsequent experiments built on this finding to explore what effect different conditions and images have on performance.

Experiment 4 found that participants made similar numbers of merge errors for identical twins regardless of familiarity. From this it could be concluded that participants familiar with the faces had the semantic information to determine that two identities were present in the images but did not have the perceptual knowledge to accurately identify each image. This could suggest that when we are presented with images of a familiar face we spend less time building a perceptual representation of the face. Therefore individual images of familiar faces might not be processed as thoroughly as images of unfamiliar faces. This finding supports similar theories put forward by face processing research, such as Bruce (1982), that non-visual codes may be involved in recognising familiar faces. This also accords with studies of expertise in other disciplines, such as x-ray image recognition in novice and expert radiologists (Myles-Worsley et al., 1988).

In Experiment 7 images of faces from the Multi-PIE database were used (Gross et al., 2010) to examine whether the variability in pose, illumination and expression in the database adequately captures within-person variability. Participants consistently perceived multiple identities in the random selection of twenty images of a single face. However in comparison to identification performance with the ambient images used in Experiment 6, participants in Experiment 7 perceived significantly fewer distinct identities in the Multi-PIE images. This suggests that although the Multi-PIE database does capture variability across pose, illumination and expression, realistic samples of images cover a broader range of variability than this. It is important to note that the Multi-PIE database was chosen precisely because of its objective to capture a wide range of variability for each face. As it turns out, even this creditable project greatly underestimates the within-person variability encountered in daily life.

Prior to this study previous identification research has reported that we are good at identifying familiar faces and bad at identifying unfamiliar faces (e.g. Bruce et al., 2001). When exploring within-person variability, Jenkins et al., (2011) reported that identification for familiar faces was good and identification for unfamiliar faces was poor. The main finding from Chapter 2 and 3 supports this previous literature but extends the findings to illustrate just how surprisingly poor unfamiliar identification is with within-person variability, and that familiar identification can also be negatively affected by within-person variability: for example when images of a familiar face fall outwith the range of variation that has been encountered by the observer (for example, images that depict the person in childhood), or when prior perceptual exposure has not been sufficient to allow highly similar faces to be reliably distinguished (for example, the identical twins in Experiment 4).

Furthermore Chapters 2 and 3 have shown that one of the major difficulties with images illustrating within-person variability is being able to group images of the same person together. This is an important insight as it brings into focus the fact that previous research has been concerned almost exclusively with telling faces apart. Indeed when each face is represented by a single image, telling them apart is the only problem. It is only when multiple images of each face are used that the problem of telling faces together can arise. The experiments in this thesis show that this problem is just as important, both for our theoretical understanding of face recognition and for applied implications.

Chapter 4: Within-person variability in social dimensions

Perceptions of personality across different images of the same face have not previously been explored in face perception research. Therefore the finding of Experiment 8, that perceptions of trustworthiness and dominance of a person can vary dependent on an image of the person, is a new finding that has interesting implications for future research. The use of the dimensions trustworthiness and dominance lends support to previous research that has found trustworthiness and dominance to be important variables in facial appearance and also traits that are easily perceived in faces (e.g. Oosterhof & Todorov 2008). Participants were able to perceive the two traits in faces across different images and across familiarity.

The findings of Experiment 8 extend the knowledge of the perception of personality traits in faces and show that perceptions can actually be image based rather than based solely on the face. In Experiment 8 trustworthiness ratings are shown to vary widely across within-person variability such as expression and eye gaze. This supports the theory proposed by Oosterhof and Todorov (2008) that the trait valence (or trustworthiness) is dependent largely on facial expressions. However, in Experiment 8 dominance ratings are also shown to vary widely across within-person variability. This finding does not coincide with Oosterhof and Todorov's theory that the trait dominance is more dependent on anatomical features (Oosterhof & Todorov 2008). This suggests that there is more to the perception of dominance than the structural features of a face.

Research into personality perceptions often create computer generated faces and manipulate the faces along different dimensions to explore the different effects changes can have on participants' perceptions (Oosterhof & Todorov 2008). This method of exploring face perceptions is the equivalent of exploring perceptions between-persons, as each newly generated face is a new face as opposed to a new image of the same face. The findings of Experiment 8 suggest that simply by changing an image of a face, perceptions of the face can change significantly. It could be proposed that the same end result is reached in each experimental scenario i.e. variability occurs and induces different perceptions. It could therefore be of interest for images of human faces to be used in perception research, specifically multiple images of within-person variability. The use of such images may lead to a better applied understanding of what causes changes in perceptions, since what can lead to a higher perception of trustworthiness in one face may not be the same in another.

The important finding from Experiment 8 is that perceptions of social dimensions from facial appearance are not stable but can vary widely depending on the image shown of a face. Furthermore, perceptions are not based solely on anatomical features but can vary on more superficial image features. Future research should therefore examine how changes in facial appearance, the environment and the camera used to capture an image, interact to determine the perception of personality from an image. Overall these findings highlight the difference between face and image perception, and the use of ambient images versus artificial faces, suggesting that research should develop to examine the effect of more realistic within-person variability on personality perceptions.

Chapter 5: Within-person variability in facial attractiveness

As with experiments 1 – 7, Experiment 9 continues from an experiment in Jenkins et al., (2011) in which perceptions of attractiveness were shown to vary significantly within-persons. Experiment 9 replicates this basic finding and extends it in two important ways. First it shows that the effect of within-person variability on attractiveness generalises to new faces. Second, it shows that the effects can occur across images of familiar and unfamiliar faces. In some ways this is a surprising result, as it could be expected that perceptions of attractiveness for familiar faces would already have been settled from previous exposure to the faces. However participants' perceptions varied across images even when they were familiar with the face, further underlining the strong effect of within-person variability.

Experiment 10 illustrates that anchoring effects can occur in the perception of facial attractiveness. Participants were presented with images of a face ordered in ascending or descending attractiveness, followed by a final image and were asked to rate how attractive they thought the person was. Anchoring occurred for the first image that participants saw, so that if an attractive image was shown first (i.e. if the images were shown in descending order) participants attractiveness perceptions were higher than if they were shown an unattractive image first (i.e. if the images were shown in ascending order).

This result mirrors the classic anchoring phenomenon whereby presentation of information prior to a judgment can act as an anchor point that influences participants' decision-making (Tversky & Kahnemann 1974). It closely follows the findings of Asch (1946): merely manipulating the order in which the preceding information is presented (rather than the content of the information) is enough to alter the impression that is formed. However, this is the first time that such effects have been shown in the context of face perception. Indeed it is only through within-person variability that this aspect of impression formation becomes visible.

The anchoring effect occurred with both familiar and unfamiliar faces. This was again surprising, as it might be expected that perceptions of attractiveness would already have been formed from previous exposure to the familiar faces.

Instead the data shows that even perceptions of familiar faces can be manipulated by simply changing the images viewed prior to making a social judgment. This further emphasises that facial perceptions are not always stable and raises the question of how the theory of lasting first impressions fits into the results seen here. For example, Willis and Todorov (2006) argued that first impressions are made quickly upon seeing a face and that these impressions endure. However the present results with familiar faces suggest that first impressions are not entirely fixed and indeed may be manipulated rather easily, at least over a certain range.

Research into personality perceptions typically focuses on biological variations between faces such as sexual dimorphism and facial symmetry (e.g. Perrett et al., 1998). To do this, research often utilises computer generated faces to explore how different physiognomic changes can affect perceived attractiveness (Little et al., 2012). The present results show that attractiveness is dependent on more than physiognomic changes such as face symmetry. Future research should therefore examine how changes in facial appearance, the environment and the camera used to capture an image, interact to determine the attractiveness of an image.

Chapter 6: Image memory

Reviewing previous memory research (Bainbridge et al., 2012; Bruce 1982; Standing 1973; Goldstein & Chance 1970), it is possible to visualise the areas that remain unaddressed within face and image memory: specifically, memory for multiple images of the same person has not been examined. It is particularly important to examine image memory in the context of familiarity, since familiar and unfamiliar face processing have been theorised to be separate processes. In Experiments 11 and 12 it was hypothesised that memory for multiple images of unfamiliar faces might actually be better than memory for multiple images of familiar faces, due to unfamiliar faces being processed based on more pictorial codes than familiar faces.

The results of Experiment 11 showed that memory for multiple images of familiar and unfamiliar faces was extremely poor, with participants performing at chance (50%) when exposure and test phases were separated by several weeks. As image memory on this scale has not been researched before the results cannot be directly compared to other research. However previous image memory research has reported high levels of memory accuracy for large sets of images, and across different lengths of delay, especially for highly familiar faces (Standing 1973; Bahrick et al., 1975). Based on these findings, better performance could therefore have been expected. Previous research has also shown recognition memory to be better for familiar faces (Bruce 1982). Based on these findings, better performance could have been expected for familiar faces in the present study. In fact image-specific memory was extremely poor for familiar and unfamiliar alike. It implies that good memory for familiar faces is not entirely driven by image memory but may involve more abstracted representations (e.g. Burton et al., 2005; Jenkins & Burton 2011).

Experiment 12 shortened the time delay between participants' first exposure to images and their memory test. Following the results of Experiment 11 it was expected that a shorter delay would improve performance. Not only did performance improve but image memory for unfamiliar faces was better than familiar faces. This supports an idea discussed by Bruce (1982) that non-visual coding occurs with images of familiar faces. This could suggest that when participants see an image of a familiar face, they do not process the image in the same way that they would process it if it was an image of an unfamiliar face. Instead the face is recognised and a non-visual code is assigned to the image.

This follows the functional model of face processing set out by Bruce and Young (1986), which introduced the idea of face recognition units (FRUs) for familiar face processing. Face recognition units are mental representations of familiar faces. In Bruce and Young's model (1986) when a familiar face is processed it is matched to the face recognition unit for that person. If a face is familiar then the image of the face may become redundant as it can be quickly linked to the right identity. The findings of Experiment 12 support this: images of unfamiliar faces can't be assigned a face recognition unit therefore the images might be processed more thoroughly and make a more distinct impression on memory.

From Chapter 6 it was discovered that memory for multiple images of the same face can be surprisingly poor. Furthermore it was discovered that memory for multiple images of different unfamiliar faces is better than for familiar faces, supporting the idea that images of familiar and unfamiliar faces are processed differently. This finding mirrors an observation made by Myles-Worsley et al., (1988) in an x-ray image memory task; participants with no radiology experience were better at remembering images of normal x-rays than participants who were senior radiologists. This result suggests that unfamiliar images were better remembered by some participants, (the normal x-rays viewed by non-radiologists), than when the same images were familiar to other participants (the normal x-rays viewed by senior radiologists). This supports the theory that unfamiliar image memory may be better than familiar image memory and demonstrates that this effect may occur in memory for other images, not just faces.

7.4 Strengths and limitations of the current methodology

Chapters 2: Within-person variability in face identification, and Chapter 3: Between-person similarity in face identification

The main strength of Experiments 1 – 7 is the novelty of the image sorting tasks. The task allows identification behaviour with multiple within-person images to be quickly and simply determined. The results shed light on the major influence that within-person variability has on accurate face identification. The experiments in Chapters 2 and 3 introduce a previously neglected side of face variability to show that within-person variability occurs to a great extent, across different faces and can affect our ability to accurately identify familiar and unfamiliar faces.

Experiments 1 – 7 expand the findings of Jenkins et al., (2011) to demonstrate that within-person variability can occur to similar extents with images of different faces. The experiments also illustrate that identification problems across within-person variability can happen in different conditions and with altered instructions. Importantly the stimuli used in Experiments 1 – 7 are ambient images as opposed to controlled images taken specifically for experimental purposes. This is an important strength of the experiments since in Experiments 6 and 7 it was shown that ambient images contain more within-person variability than images taken for the Multi-PIE database (Gross et al., 2010). This suggests that there is more to within-person variability than 'PIE' i.e. pose, illumination and expression.

A possible limitation of Experiments 1 – 7, is that the process of participants' identification behaviour was not observed or recorded. Therefore participants' decision-making behaviour and how they reached the final number of identities was not observed. It would have been interesting to see how participants approached the task, especially in view of the large range in time taken to complete the task.

Another limitation of the image-sorting experiments is that images of only 1 or 2 faces were given to participants to identify. By introducing images of more faces it could be observed whether this improves identification or results in similarly poor performance. Performance might be expected to improve: as participants are exposed to more within-person variability they become more aware of it. However, as shown in Experiment 5, awareness of within-person variability seems to occur on a face-by-face basis and cannot be generalised. Alternatively, performance might remain the same, as experiments 1–7 have shown that identification errors can occur across different faces; increasing the number of faces shown could simply result in similar within-person variability confusion.

Chapter 4: Within-person variability in social dimensions and Chapter 5: Within-person variability in facial attractiveness

A strength of Experiments 8, 9 and 10, is the new understanding of how perceptions can vary depending on an image of a face. This clearly demonstrates that perceptions are not stable impressions but can be manipulated by representing a face with a different image. Previous research has shown the importance of facial appearance on social judgments and decision-making, therefore understanding how faces can vary in attractiveness can lead to knowing how to manipulate facial appearance to a person's advantage.

Variability in perceptions was also found for images of familiar faces. Despite having been previously exposed to the celebrities used as familiar faces, participants were still influenced by within-person variability and so their ratings of different images of familiar faces were varied. This further underlines the instability of personality perceptions based on facial appearance and the strength of within-person variability. This allows for better understanding of familiar face processing: for example, these findings suggest that although we use non-visual codes to recognise and identify a familiar image we are still influenced by visual aspects when perceiving attractiveness.

A limitation of Experiments 8, 9 and 10, is the lack of image analysis to examine whether there is a specific aspect of within-person variability that causes the variability in perceptions. It would have been good to explore what aspects of within-person variability were causing these social perceptions to be swayed, particularly since Experiment 10 found that participants continued to be influenced by the order effect of images shown prior to attractiveness ratings being made for familiar faces. This is a surprising finding as participants had already been exposed to the faces either in image form or from seeing the person in their professional role, yet were influenced by a short sequence of images they were shown briefly prior to rating the person. Returning to the within-person variability categories outlined in Chapter 1, it would have been interesting if Chapter 4 and 5 had examined whether perceptions were influenced by variations in face characteristics, environment variation or camera variations.

Chapter 6: Image memory

A key strength of Experiments 11 and 12 is establishing participants' image memory ability with multiple images of large numbers of different faces. A strength of Experiment 12 in particular is the finding that memory for different images of unfamiliar faces is better than for familiar faces. This is important because it supports theories that familiar and unfamiliar face processing are different processes (Bruce & Young 1986; Bruce 1982). Another strength in Chapter 6 is the clarification of the difference between image and face memory. This is an important distinction to make in interpreting previous research and designing future experiments.

A possible limitation of Chapter 6 is the absence of further investigation into the effect of time delays on image memory. Experiment 11 demonstrated that image memory for multiple images of different faces was very poor following a long time delay of at least 6 weeks. In reducing the time delay to 24 hours, Experiment 12 demonstrated that overall image memory did not improve but that memory for images of unfamiliar faces was higher than for familiar faces. It would have been interesting to complete a further experiment exploring image memory with no time delay, and with a time delay somewhere between 24 hours and 6 weeks to gain a better idea of how memory performance increases or decreases. It would also be interesting to explore whether image memory is ever better for familiar faces than for unfamiliar faces. This would also help build a better picture of image memory overall and the decay rate of image memory.

7.5 Applied Implications

Forensic

The findings from Experiments 1–7 have important applied implications for identification in forensic settings. Experiments 1–7 concluded that participants' identification ability was poor when faced with images that contained within-person variability. This suggests that in an applied setting, identification may be negatively affected if images that show within-person variability are involved, for example if identification is needed from an image taken during a crime with an image taken of the individual following arrest. In an applied setting, any misidentification can have serious consequences. A greater understanding and awareness of within-person variability needs to be introduced to forensic procedures in order to improve identifications made based on photographs.

Previous research has shown identification to be difficult across two or more images of a face and have applied identification to imitate forensic settings. For example, Bruce et al., (1999) simulated a real life identification process by asking participants to identify if a person shown in a video still was present in an array of 10 high quality photographs. Participants' performance was poor and it was concluded that impressions of likeness or dissimilarity between different images resulted in inaccurate identifications. In Burton et al., (1999) police officers described as experienced with forensic identification participated in a recognition task. They viewed short video clips of target people and later were asked to indicate whether the target was present in a selection of high quality images. Correct performance was poor and police officers performed no better than non-experts in forensic identification.

The images used by Bruce et al., (1999) and Burton et al., (1999) were taken either on the same day or the following day to keep variability to a minimum. Despite this, poor performance was still observed suggesting that identification across multiple images is difficult even with high quality images taken within a short period of time. Experiments 1-7 expand this finding to show the extent of within-person variability and the difficulties experienced by participants in identification across within-person variability. As images will not be controlled in applied settings, the findings from Experiments 1-7 may better represent potential difficulties in forensic identification.

Photographic identification

The findings from Experiments 1-7 also have important applied implications for the use of photographic identification. If faces vary widely within-persons, using one image to represent a face may not be sufficient. As demonstrated by Kemp et al., (1997), matching a target person to their presented photographic identification can be highly error prone. Building on this, Experiments 1 – 7 illustrate how difficult it is to correctly group multiple images together as one identity. This finding highlights how difficult it is to match one image of a face with another. Specifically, Experiment 2 demonstrated the difficulties of identifying images showing within-person variability to current passport photographs of that identity: participants frequently made errors by matching an image to the wrong passport photograph. This is a clear indication of problems that could occur with photographic identification.

A possible solution to the limitation of displaying only one photograph on passports would be to utilise the security chip embedded in newly issued UK passports. The security chip has the capacity to store additional images, therefore additional images could be added to the chip at regular intervals or, for example, when passing through border control in airports. These additional images would then be available for border control officers to see upon scanning a passport, allowing officers to gain a better visual representation of the passport holder. A further development of passport photographs could also address the limitations of expressions allowed in passport photographs as the findings of Experiments 1 – 7 clearly demonstrate that expressions play a key role in within-person variability.

Photographic representation

The findings from Experiments 8, 9 and 10 have important applied implications for how we use images to represent ourselves. The results from Chapters 4 and 5 have shown that perceptions of trustworthiness, dominance and attractiveness are not stable across multiple images of a person's face. Photographs are used as representations of people across a range of different platforms from social networking sites, to company websites and more recently images are frequently required prior to job interviews, as well as official forms of identification including passports and driving licenses. With this in mind the findings from Experiments 8 and 9 have an important applied implication: if perceptions are quickly formed based on facial appearance, as shown in previous research (Willis & Todorov 2006), and these perceptions can vary, then the images we use to represent ourselves are important and influential. For example, if pulled over by the police while driving, presenting a license displaying an image that is perceived as trustworthy may be favourable over an image that may be perceived as untrustworthy.

If perceptions are image dependent then we can vary the images we use to best suit different contexts. This illustrates the control we can have over perceptions made by other people based on our appearance. Previous research has shown the importance facial appearance can have on social perceptions and decision-making (Watkins and Johnston 2000; Lenz & Lawson 2011). Therefore the implication that we can control these perceptions is important for how we choose to visually represent ourselves, and also interesting as it suggests we base social perceptions on visual image dependent information. This would be an interesting future research development.

7.6 Future Directions

Training

In Experiments 1-7 identification performance was improved when descriptive and visual constraints were introduced and when images of two very dissimilar faces were shown together. This improvement in identification suggests that under certain conditions participants were better able to cope with within-person variability. Based on this finding it would be interesting to manipulate conditions further to explore whether people can learn, or be trained, to better cope with within-person variability. Experiments 2 and 3 demonstrated improved identification performance when participants were made aware of how many identities existed amongst the images shown. If this was developed in a similar image-sorting design, it could be established whether being exposed to within-person variability for practise faces during a learning phase could help understanding of within-person variability with images of other faces.

In Experiment 1, being familiar with one of the faces, and so able to identify the person in multiple images, did not help improve identification for the remaining face. This finding suggests that to improve understanding of within-person variability, a learning phase would have to involve more extensive repetition and perhaps an overall review of what within-person variability can include - for example, across the three categories of variations in face characteristics, environmental variations and camera variations. If it is possible to train participants in within-person variability it would be interesting to develop this training in a more applied setting such as with agencies that use photographic identification, such as border control agencies and the police service.

Automatic face recognition

In Experiment 7 an image-sorting task was conducted with images selected from the Multi-PIE database created by Gross et al., (2010). The aim of the Multi-PIE database was to build a collection of varying images across the dimensions of pose, illumination and expression (PIE) to progress the advancement of face recognition algorithms for especially challenging and realistic image sets. The results in Experiment 7 suggested that the images selected from the database were not as variable as the ambient images used in Experiments 1 - 6. This could suggest that by focusing solely on pose, illumination and expression the database underestimates within-person variability in important ways. An interesting future direction would be to take what was learnt from Experiment 7 and build a database of more realistic within-person variability. The database could then be used either in human face recognition tasks or in testing automatic face recognition systems.

The effect of ageing

In Experiments 1-6 images taken of faces when the person was younger were frequently wrongly identified. This suggests that ageing can be an influential variable in within-person variability. However, some iconic images of famous faces are easily recognisable even in younger images, for example images of Paul McCartney in the Beatles. It would be interesting to explore whether this ease of recognising highly familiar faces is due to frequent exposure to the person, or whether ageing effects differ in their strength across individuals. To explore the effect of ageing within-person, a similar image-sorting design could be implemented with images deliberately selected to show variability across the age span of the face. In the same way that training or learning the extent of within-person variability could improve identification, gaining a better understanding of the effects of ageing on facial appearance could help improve identification performance.

The other race effect

Previous research on the other-race effect has shown that humans are better at identifying faces from their own race as opposed to other races (Malpass & Kravitz 1969; Lindsay, Jack & Christian 1991). It would be interesting to develop an image-sorting task to explore the other-race effect with images of within-person variability. As people often find faces from other races difficult to distinguish, it could be predicted that images of within-person variability would look more alike if the face was from a less familiar race. Therefore the images would be less frequently split into different identities resulting in a better overall performance. This would be a really interesting step to explore and could help with the understanding of how images of within-person variability are processed.

Do different faces vary differently?

Results from Experiments 1 – 12 have shown that within-person variability occurs across different faces. This shows that within-person variability is not limited to particular individuals, but instead seems to be a general phenomenon of facial appearance. An interesting development from this study would be to establish whether or not different faces vary in the same ways. For example in the context of identification, do all faces exhibit particular types of within-person variability that lead to identification errors? Similarly in the context of personality perception, do all faces exhibit particular types of within-person variability that lead to high or low ratings of trustworthiness, or do the causes of perceived trustworthiness differ between individuals?

To begin to explore such a hypothesis, a simple image analysis could be conducted to establish common variations in facial appearance such as hairstyle changes, changes in weight, or changes in physical signals of health. Analysis could be conducted on images of both male and female faces, ensuring that the image selection consists of ambient images as opposed to images that have been controlled. From here the analysis could develop to form a model of how a face varies, and this model could then be applied to different faces to see if variability is consistent or if some faces vary differently than others.

Image analysis could also explore variability with social and personality perceptions. Using social dimensions that previous research has shown can be perceived facially, such as trustworthiness and dominance (Oosterhof & Todorov 2008) or warmth and competence (Fiske, Cuddy & Glick 2007), a database of ratings for multiple images of different faces could be developed. Comparing the images and the ratings assigned to the images, it would then be possible to conduct an image analysis to establish whether certain variations are directly comparable to social dimensions. For example, does an image of a smiling expression consistently equate to a high trustworthy rating or is the relationship more face dependent?

Individual differences in observers

Previous research has begun to explore variability in the face recognition ability of observers; for example, exploring performance with people at either end of the prosopagnosia scale: “Super-recognisers” (Russell, Duchaine & Nakayama 2009) and “Developmental prosopagnosics” (Duchaine & Nakayama 2006). It would be interesting to develop this research to establish if there are some people who are able to cope better with the effect of within-person variability, for example conducting an image-sorting task with “Super-recognisers”. It would also be of interest to examine whether people who are poor at identifying across within-person variability will always struggle, or whether there is a possibility of learning more about within-person variability and being able to apply it to other faces. Previous research has also explored face recognition abilities in different professionals (Burton et al., 1999). It would be interesting to explore recognition abilities across within-person variability with professionals who are not necessarily experienced in forensic identification but who perhaps observe within-person variability in a more natural way, such as teachers who see the same pupils every day across changes in dimensions such as health and ageing.

This thesis introduces and explores the concept of within-person variability in the context of face identification, face perception and image memory. In summary, the results illustrate how within-person variability can cause problems in identification; how perceptions can vary within-person and are often image dependent; and that memory for multiple images of the same face is surprisingly poor across different time delays, but is better for unfamiliar faces over a short time delay. Overall, it can be concluded that within-person variability is an important aspect of face processing but one that has been overlooked, possibly due to experimental methods that have emphasised stimulus control. The findings of this thesis stress the importance of understanding within-person variability for future experimentation and theoretical development.

Appendix A

Chapter Four: Within-person variability in Social Dimensions

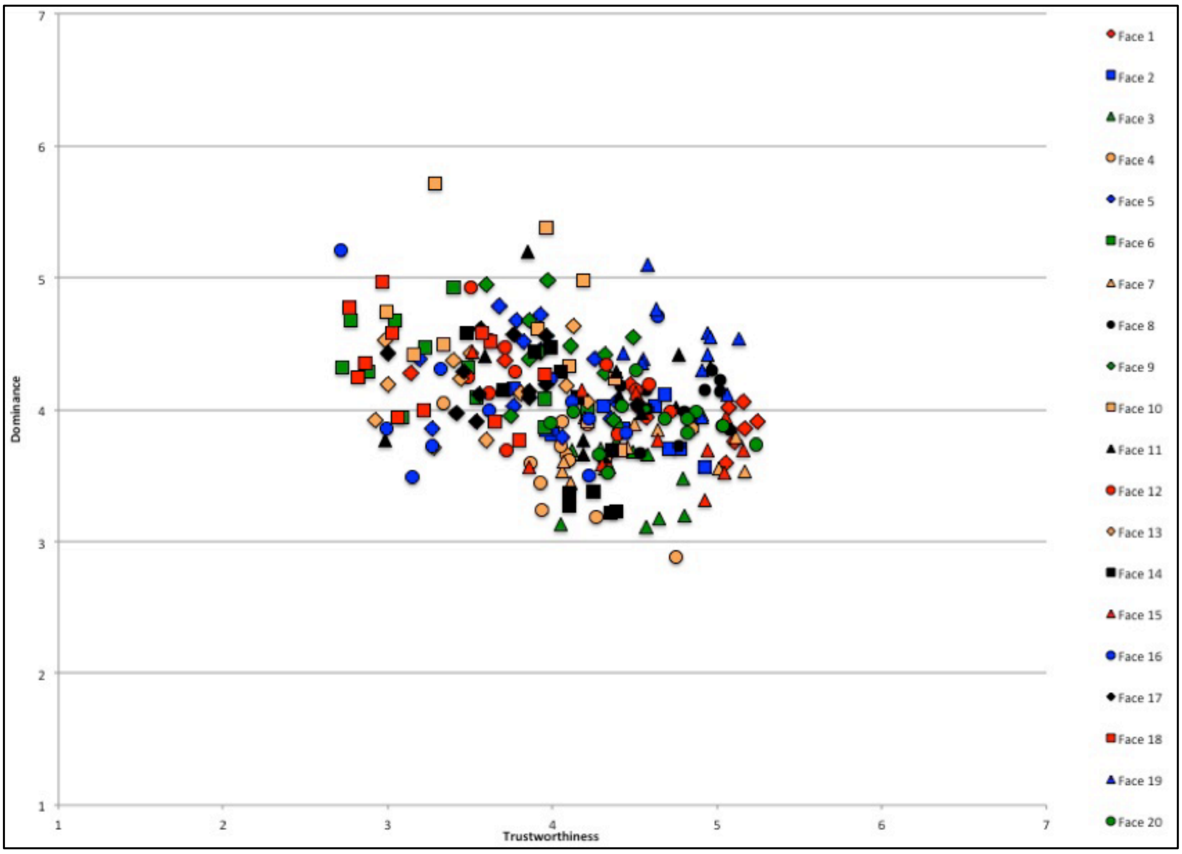


***Figure A.1*. Experiment 8: Scatterplot of ratings for Faces 1 – 20.**
Scatterplot illustrating the trustworthy ratings (x-axis) and dominance ratings (y-axis) of every image of Faces 1 – 20, made by participants in Experiment 8. Each face is represented by a different symbol as shown in the key.

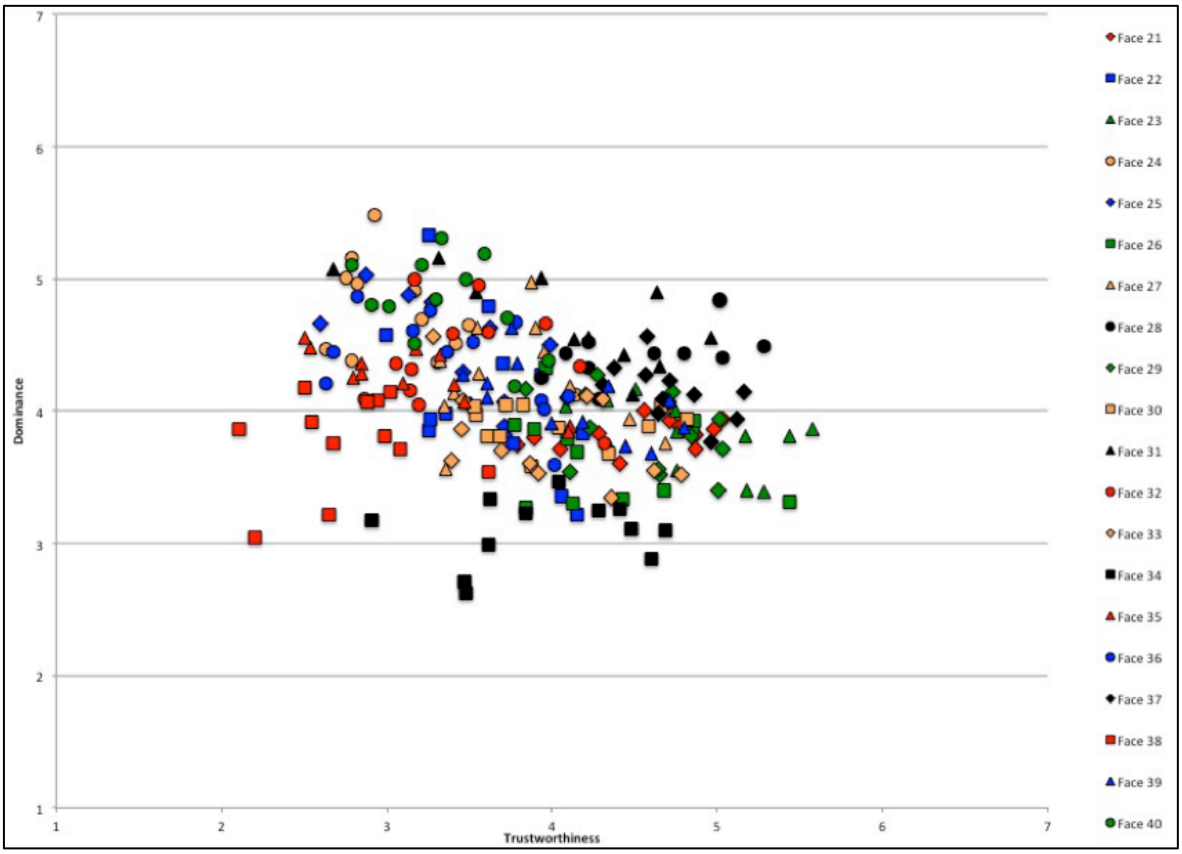


***Figure A.2*. Experiment 8: Scatterplot of ratings for Faces 21 – 40.**
Scatterplot illustrating the trustworthy ratings (x-axis) and dominance ratings (y-axis) of every image of Faces 21 – 40, made by participants in Experiment 8. Each face is represented by a different symbol as shown in the key.

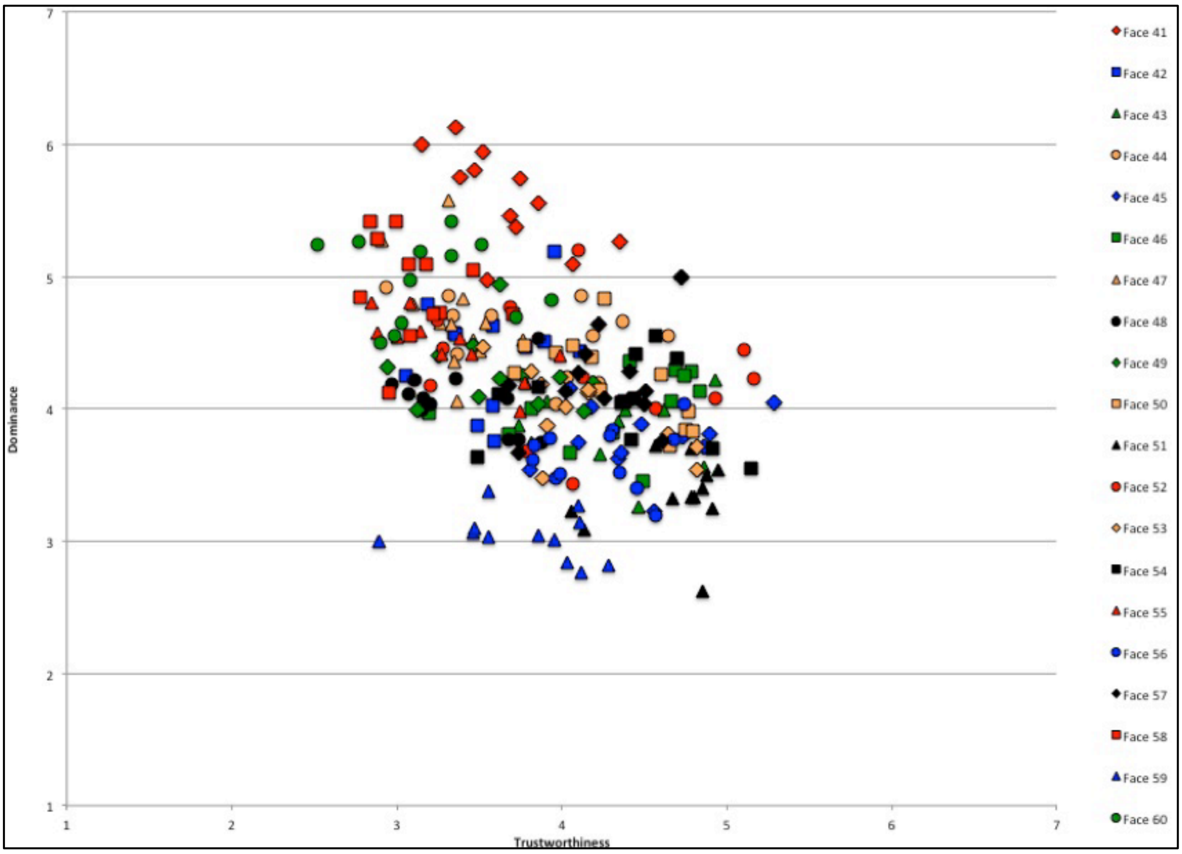


Figure A.3. **Experiment 8: Scatterplot of ratings for Faces 41 – 60.**
Scatterplot illustrating the trustworthy ratings (x-axis) and dominance ratings (y-axis) of every image of Faces 41 – 60, made by participants in Experiment 8. Each face is represented by a different symbol as shown in the key

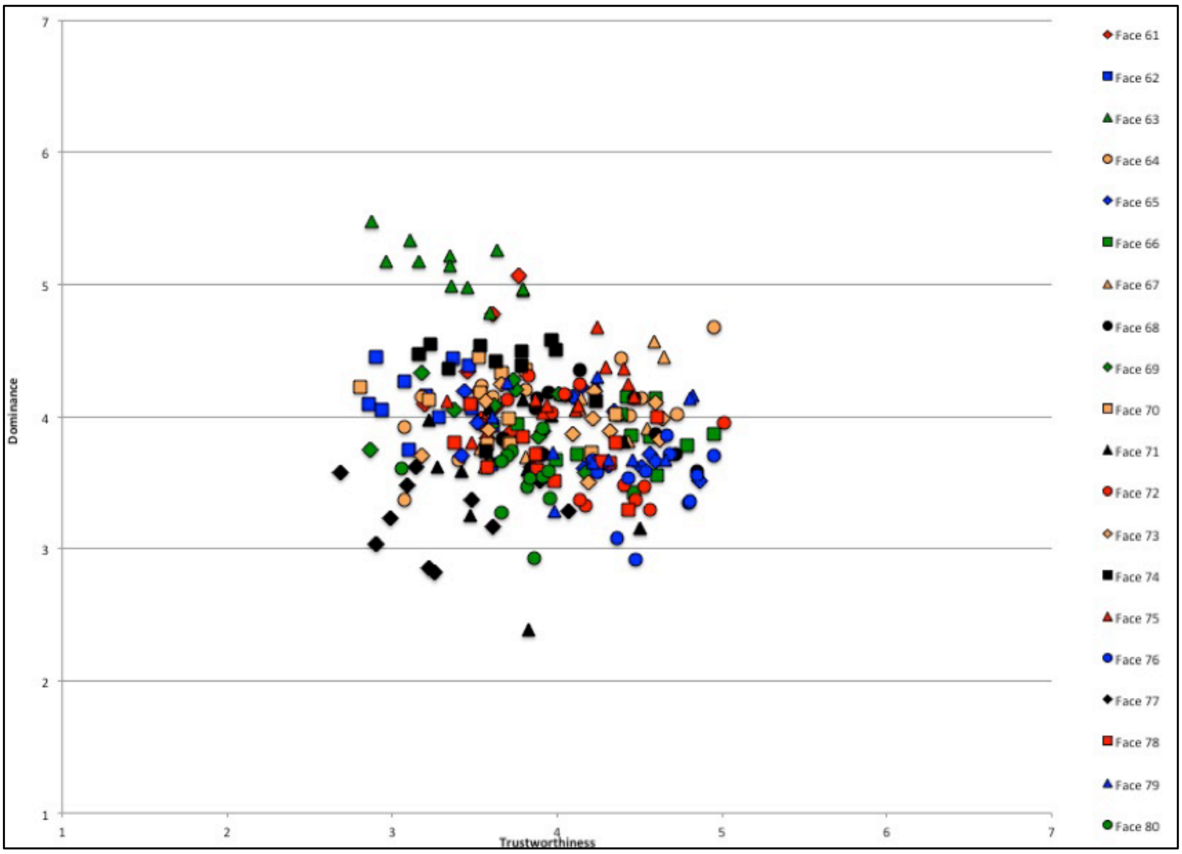


Figure A.4. **Experiment 8: Scatterplot of ratings for Faces 61 – 80.**
Scatterplot illustrating the trustworthy ratings (x-axis) and dominance ratings (y-axis) of every image of Faces 61 – 80, made by participants in Experiment 8. Each face is represented by a different symbol as shown in the key.

Personality Trait Definitions

Trustworthiness

To be trustworthy is to be reliable and dependable, in addition to being honest or truthful.

Dominance

To be dominant is to have a manner that alludes to a certain power or influence, often in addition to a tendency to be in control.

Figure A.5. **Definitions of the personality traits trustworthiness and dominance.** Definitions given to participants in Experiment 8, and also to participants in Experiments 11 and 12.

Table A.1. **Experiment 8: The identities of the 80 famous British and Australian faces used as stimuli.**

Face	Identity	Face	Identity
1	Ada Nicodemou	41	Alan Sugar
2	Anna Coren	42	Alex Salmond
3	Antonia Kidman	43	Alexa Chung
4	Bec Cartwright	44	Bruce Forsyth
5	Bert Newton	45	Carol Smilie
6	Brendan Nelson	46	Carol Vorderman
7	Dawn Fraser	47	Charles Kennedy
8	Don Burke	48	Chris Moyles
9	Eddie Mcguire	49	Chris Tarrant
10	George Gregan	50	Cilla Black
11	Georgie Parker	51	Connie Huq
12	Grant Hackett	52	Davina Mccall
13	Gretel Killeen	53	Denise van Outen
14	Guy Sebastian	54	Dermot O'Leary
15	Hamish Blake	55	Derren Brown
16	Ian Dickinson	56	Fearne Cotton
17	Ita Buttrose	57	Gary Liniker
18	James Packer	58	Jack Dee
19	Jamie Durie	59	Jennifer Ellison
20	Jessica Rowe	60	Jeremy Kyle
21	Johanna Griggs	61	Jeremy Paxman
22	Julia Gillard	62	Jodie Marsh
23	Kate Ritchie	63	Jon Snow
24	Kim Beazley	64	Jonathan Ross
25	Kyle Sandilands	65	Judy Finigan
26	Lisa Mccune	66	Kelly Brook
27	Mark Holden	67	Kirsty Gallacher
28	Megan Gale	68	Lorraine Kelly
29	Melissa Doyle	69	Louis Walsh
30	Merrick Watts	70	Michael Barrymore
31	Michael Klim	71	Miquita Oliver
32	Morris Iemma	72	Myleen Klass
33	Natasha Stott Despoja	73	Natasha Kaplinski
34	Nikki Webster	74	Noel Edmonds
35	Pauline Hanson	75	Phillip Schofield
36	Peter Costello	76	Rachel Stevens
37	Sandra Sully	77	Simon Amstell
38	Schapelle Corby	78	Tess Daly
39	Shannon Noll	79	Ulrika Johnson
40	Willie Mason	80	Vernon Kay

Appendix B

Chapter Five: Within-person variability in facial attractiveness

Table B.1. **Experiment 9 and 10: The identities of the 20 famous British and Austrian faces used as stimuli.**

Face	Identity
1	Armin Assinger
2	Moritz Bleibtreu
3	Dieter Bohlen
4	Thomas Gottschalk
5	Harald Krassnitzer
6	Anke Engelke
7	Verona Pooth
8	Franka Potente
9	Christina Stumeur
10	Ingrid Thurnher
11	Derren Brown
12	Gary Liniker
13	Jonathon Ross
14	Alex Salmond
15	Phillip Schofield
16	Alexa Chung
17	Tess Daly
18	Konnie Huq
19	Carol Smillie
20	Carol Vorderman

References

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science*, *323*(5918), 1183-1183.

Asch, S. E. (1946). Forming Impressions of Personality. *The Journal of Abnormal and Social Psychology*, 41(3), 258-290

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of experimental psychology: General*, *104*(1), 54-75.

Bainbridge, W. A., Isola, P., Blank, I., & Oliva, A. (2012). Establishing a Database for Studying Human Face Photograph Memory. Poster presented at the *34th Annual Conference of the Cognitive Science Society*. Sapporo, Japan.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*(46), 17948-17953.

Berry, D. S. & Finch Wero, J. L. (1993). Accuracy in face perception: A view from ecological psychology. *Journal of personality*, *61*(4), 497-520.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325-14329.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(1), 105-116.

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

Bruce, V., Henderson, Z., Newman, C. & Burton, A. M. (2001). Matching Identities of Familiar and Unfamiliar Faces Caught on CCTV Images. *Journal of Experimental Psychology: Applied, 7(3),* 207-218

Bruce, V., & Valentine, T. (1985). Identity priming in the recognition of familiar faces. *British Journal of Psychology*, *76*(3), 373-383.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305-327.

Burton, A. M. (in press). Why has research in face recognition progressed so slowly? The importance of variability. Quarterly Journal of Experimental Psychology.

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

Burton, A. M., & Jenkins, R. (2011). Unfamiliar face perception. *The Oxford handbook of face perception*, 287-306.

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

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943-958.

Burton, A. M., White, D. & McNeil, A. (2010). The Glasgow Face Matching Test. *Behaviour Research Methods, 42 (1),* 286-291

Burton, A. M., Wilson, S., Cowan, M. & Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence From Security Surveillance. *Psychological Science, 10(3)*

Carbon, C. C. (2008). Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception*, *37*(5), 801.

Carey, S., De Schonen, S., & Ellis, H. D. (1992). Becoming a Face Expert [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 95-103.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science*, *195*(4275), 312-314.

Cash, T. F., & Kilcullen, R. N. (1985). The Aye of the Beholder: Susceptibility to Sexism and Beautyism in the Evaluation of Managerial Applicants1. *Journal of Applied Social Psychology*, *15*(4), 591-605.

Cave, C. B. (1997). Very long-lasting priming in picture naming. *Psychological Science*, *8*(4), 322-325.

Chen, C. C., Chen, C. M., & Tyler, C. W. (2013). Depth Structure from Asymmetric Shading Supports Face Discrimination. *PloS one*, *8*(2), e55865.

Clifford, M. M., & Walster, E. (1973). The effect of physical attractiveness on teacher expectations. *Sociology of Education*, 248-258.

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *PERCEPTION-LONDON-*, *31*(8), 985-994.

Clutterbuck, R. & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition, 11(7),* 857-869

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.

Coetzee, V., Perrett, D. I., & Stephen, I. D. (2009). Facial adiposity: a cue to health?. *Perception, 38*(11), 1700.

Coetzee, V., Re, D., Perrett, D. I., Tiddeman, B. P., & Xiao, D. (2011). Judging the health and attractiveness of female faces: Is the most attractive level of facial adiposity also considered the healthiest?. *Body Image*, *8*(2), 190-193.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559-564.

Davies, J. P. & Valentine, T. (2009). CCTV on Trial: Matching Video Images with the Defendant in the Dock. *Applied Cognitive Psychology, 23,* 482-505

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of personality and social psychology*, *24*(3), 285-290.

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

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

Goldstein, A. G., & Chance, J. E. (1970). Visual recognition memory for complex configurations. *Perception & Psychophysics*, *9*(2), 237-241.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image and Vision Computing*, *28*(5), 807-813.

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

Harper, B., & Latto, R. (2001). Cyclopean vision, size estimation, and presence in orthostereoscopic images. *Presence: Teleoperators & virtual environments*,*10*(3), 312-330.

Harrison, A. A. (1969). Exposure and popularity. *Journal of Personality*, *37*(2), 359-377.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223-233.

Henderson, Z., Bruce, V. & Burton, A. M. (2001). Matching the Faces of Robbers Captured on Video. *Applied Cognitive Psychology, 15,* 445-464

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.

Hood, B. M., Macrae, C. N., Cole-Davies, V., & Dias, M. (2003). Eye remember you: The effects of gaze direction on face recognition in children and adults. *Developmental Science*, *6*(1), 67-71.

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition.*

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

Jenkins, R., White, D., van Montfort, X. & Burton, A. M. (2011). Variability in photos of the same face. *Cognition,* 121, 313-323

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

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of personality and social psychology*, *89*(6), 899.

Kalick, S. M., Zebrowitz, L. A., Langlois, J. H., & Johnson, R. M. (1998). Does human facial attractiveness honestly advertise health? Longitudinal data on an evolutionary question. *Psychological science*, *9*(1), 8-13.

Kanade, T., Cohn, J. & Tian, Y. L. (2000). Comprehensive database for facial expression analysis. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46–53.

Kemp, R., Towell, N. & Pike, G. (1997). When Seeing should not be Believing: Photographs, Credit Cards and Fraud. *Applied Cognitive Psychology, 11,* 211-222

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology*, *29*(3), 299.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M. & Smoot, M. (2000). *Psychological Bulletin,* 126(3), 390-423

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological science*, *1*(2), 115-121.

Langlois, J. H., Roggman, L. A., & Musselman, L. (1994). What is average and what is not average about attractive faces? *Psychological science*, *5*(4), 214-220.

Lenz, G. S., & Lawson, C. (2011). Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science*, *55*(3), 574-589.

Lindsay, D. S., Jack, P. C., & Christian, M. A. (1991). Other-race face perception. *Journal of Applied Psychology*, *76*(4), 587-589.

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior*, *28*(1), 18-27.

Little, A. C., Burt, D. M., Penton-Voak, I. S., & Perrett, D. I. (2001). Self-perceived attractiveness influences human female preferences for sexual dimorphism and symmetry in male faces. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1462), 39-44.

Little, A. C. & Hancock, P. J. B. (2002). The role of masculinity and distinctiveness in judgments of human male facial attractiveness. *British Journal of Psychology,* 93, 451-464

Little, A. C., Hockings, K. J., Apicella, C. L., & Sousa, C. (2012). Mixed-ethnicity face shape and attractiveness in humans. *Perception*, *41*, 1486-1496.

Livingston, R. W., & Pearce, N. A. (2009). The Teddy-Bear Effect Does Having a Baby Face Benefit Black Chief Executive Officers? *Psychological Science*, *20*(10), 1229-1236.

Logie, R. H., Baddeley, A. D., & Woodhead, M. M. (1987). Face recognition, pose and ecological validity. *Applied Cognitive Psychology*, *1*(1), 53-69.

Lorenzo, L. G., Biesanz, J. C. & Human, L. J. (2010). What Is Beautiful Is Good and More Accurately Understood: Physical Attractiveness and Accuracy in First Impressions of Personality. *Psychological Science* 21(12), 1777-1782

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of personality and social psychology*, *13*(4), 330.

Mazella, R. & Feingold, A. (1994). The Effects of Physical Attractiveness, Race, Socioeconomic Status and Gender of Defendants and Victims on Judgments of Mock Jurors: A meta-analysis. *Journal of Applied Social Psychology,* 24(15) 1315-1344

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

Mitchell, D. B. (2006). Non-conscious Priming After 17 Years Invulnerable Implicit Memory? *Psychological Science*, *17*(11), 925-929.

Morrison, E. R., Morris, P. H., & Bard, K. A. (2013). The stability of facial attractiveness: is it what you've got or what you do with it? *Journal of Nonverbal Behavior*, 1-9.

Mueller U. & Mazur A. (1996). Facial dominance of West Point cadets as a predictor of later military rank. *Social Forces,* 74(3), 823-850.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*(9), 1142-1150.

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

Olson, I. R., & Marshuetz, C. (2005). Facial attractiveness is appraised in a glance. *Emotion*, *5*(4), 498-502.

Oosterhof, N. N. & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the USA, 105(32),* 11087-92

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *JOSA A*, *10*(3), 405-411.

Perrett, D. I., Burt, D. M., Penton-Voak, I. S., Lee, K. J., Rowland, D. A., & Edwards, R. (1999). Symmetry and human facial attractiveness. *Evolution and human behavior*, *20*(5), 295-307.

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., Henzil, S. P., Castles, D. L. & Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, *394*(6696), 884-887.

Principe, C. P., & Langlois, J. H. (2013). Children and adults use attractiveness as a social cue in real people and avatars. *Journal of experimental child psychology*.

Robbins, R., & McKone, E. (2003). Can holistic processing be learned for inverted faces? *Cognition*, *88*(1), 79.

Rosar, U., Klein, M. & Beckers, T. (2008). The frog pond beauty contest: Physical attractiveness and electoral success of the constituency candidates at the North Rhine-Westphalia state election of 2005. *European Journal of Political Research*, 47, 64-79

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

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior*, *6*(1), 156-163.

Sigelman, C. K., Sigelman, L., Thomas, D. B. & Ribich, F. D. (1986). Gender, Physical Attractiveness and Electability: An Experimental Investigation of Voter Biases. *Journal of Applied Social Psychology,* 16(3), 229-248

Standing, L. (1973). Learning 10000 pictures. *The Quarterly journal of experimental psychology*, *25*(2), 207-222.

Tanaka, J. W. (2001). The entry point of face recognition: Evidence for face expertise. *Journal of experimental psychology General*, *130*(3), 534-543.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623-1626.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in cognitive sciences*, *12*(12), 455-460.

Tracy, J. L., & Beall, A. T. (2011). Happy guys finish last: The impact of emotion expressions on sexual attraction. *Emotion-APA*, *11*(6), 1379.

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

Walster, E., Aronson, V., Abrahams, D., & Rottman, L. (1966). Importance of physical attractiveness in dating behavior. *Journal of personality and social psychology*, *4*(5), 508-516.

Watkins, L. M. & Johnston, L. (2000). Screening Job Applicants: The Impact of Physical Attractiveness and Application Quality. *International Journal of Selection and Assessment,* 8(2)

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, *37*(3), 395.

Willis, M. L., Palermo, R., & Burke, D. (2011). Social Judgments are Influenced By Both Facial Expression and Direction of Eye Gaze. *Social Cognition*, *29*(4), 415-429.

Willis, J. & Todorov, A. (2006). First Impressions: Making up your mind after a 100-ms exposure. *Psychological Science,* 17(7)

Yang, J., Zhang, D., Frangi, A. F., & Yang, J. Y. (2004). Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *26*(1), 131-137.

Young, A. W., Hay, D. C., & Ellis, A. W. (1985). The faces that launched a thousand slips: Everyday difficulties and errors in recognizing people. *British Journal of Psychology*, *76*(4), 495-523.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. 1. *Pers. Soc. Psycho*, *9*(2 pt 2).