

Imagining Fictional Faces

Lily Lai Hang Chan

Department of Psychology

University of York

Submitted for the Degree of PhD

September 2019

Abstract

Fictional characters loom large in cultural traditions throughout recorded history, and are commonly portrayed in literature and visual arts. The persistence of these traditions demonstrates that information concerning the appearance of fictional characters - including facial appearance - can be preserved and shared among individuals. The current thesis is an attempt to understand the cognitive processes underlying mental imagery for fictional faces. It was already established that mental representations of real faces undergo qualitative change as visual exposure leads to familiarity. Fictional faces are never seen directly, though they may be represented in various ways. If fictional faces can acquire the psychological hallmarks of familiar faces, this would suggest alternative routes to face learning, besides natural visual exposure. To date however, this possibility has been largely ignored. The experiments in this thesis addressed learning of fictional faces by examining familiarity effects for fictional faces, and by assessing the consequences of reading descriptions on mental imagery for fictional characters. The main findings indicate that face representations and face learning may be more adaptable than previously assumed, accommodating photographic images, different types of drawings, and written descriptions. All of these representations can be quantified and compared using the common currency of social inference ratings. Written descriptions of physical and character attributes both contribute to mental imagery for faces, and these complementary types of information can converge on specific facial identities. As well as enriching our psychological understanding of face processing, the current thesis forms a bridge between the scientific study of faces, and portrayals of faces in the arts.

Table of Contents

Abstract	2
Acknowledgements	6
Declaration	7
Ethics Statement	8
Chapter 1 <i>General Introduction</i>	9
.....	9
1.1 Familiarity and Mental Representations of Faces	11
1.1.1. Unfamiliar face processing	11
1.1.2. Familiar face processing	15
1.2 Variability and Face Learning	21
1.3 Face Shape and Face Texture	22
1.4 Visual Representations of Faces	27
1.4.1. Photos.....	27
1.4.2. Drawings	28
1.5 Non-visual Representations of Faces	30
1.5.1. Verbal descriptions	30
1.5.2. Social inferences.....	30
1.6 Mental Imagery and Face Perception	34
1.7 General Methodological Approach and Overview	36
Chapter 2 <i>Familiarity and Identification for Fictional Faces</i>	39
.....	39

Introduction.....	40
Experiment 1	42
General Discussion.....	61
Chapter 3.....	<i>Social Inference Ratings for Fictional Faces</i>
.....	64
Introduction.....	65
Experiment 2	67
Experiment 3	73
Experiment 4	76
General Discussion.....	83
Chapter 4.....	<i>Identification from Sparse Drawings</i>
.....	85
Introduction.....	86
Experiment 5	89
Experiment 6	95
Experiment 7	98
General Discussion.....	102
Chapter 5.....	<i>Matching Photos to Written Descriptions</i>
.....	105
Introduction.....	106
Experiment 8	107

Experiment 9	114
Experiment 10	125
General Discussion.....	128
Chapter 6.....	<i>Mental Imagery from Written Descriptions</i>
.....	131
Introduction.....	132
Experiment 11	133
Experiment 12	143
Experiment 13	145
Experiment 14	150
General Discussion.....	153
Chapter 7.....	<i>General Discussion</i>
.....	156
References.....	171

Acknowledgements

I would like to thank Dr Rob Jenkins for his support and supervision throughout this thesis. I would also like to thank Professor Mike Burton and Dr Catherine Preston for their advisory.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Ethics Statement

The experiments in this thesis were approved by the Psychology Ethics Committee at the University of York. All experimental participants gave their written informed consent.

Chapter 1

General Introduction

Starting in the 1980s, face perception has become an established topic in cognitive psychology. The standard framing of this topic has been as a sub-branch of visual perception. As such, mainstream face perception research has inherited many of the questions, methods, and controversies that pervade visual perception as a whole. The overarching aim of the perceptual research agenda is to understand how the visual system makes sense of external reality. The current thesis steps outside the mainstream by considering faces that are not a part of external reality—specifically, fictional faces, how they are imagined, and how they are communicated.

Fictional characters loom large in cultural traditions throughout recorded history, and are commonly portrayed in literature and visual arts. The persistence of these traditions demonstrates that information concerning the appearance of fictional characters - including facial appearance - can be preserved and shared among individuals. From a cognitive science perspective, these observations raise interesting questions about the mental representations that could support such processes. The current thesis is an attempt to articulate some of these questions and to answer them experimentally.

I begin by reviewing some of the major areas of face perception research, and how they might relate to imagining fictional faces. The review sets out the important distinction between unfamiliar faces and familiar faces, and what this distinction tells us about face learning. I then examine how different types of visual and non-visual information contribute to mental representations of faces. Finally, I consider the intersection between face perception and mental imagery. The review informs the specific research questions outlined at the end of this chapter, and informs the choice of methods in the experimental chapters that follow.

1.1 Familiarity and Mental Representations of Faces

Mental representations of faces appear to be very different for familiar faces (e.g. friends, family members, celebrities) and unfamiliar faces (i.e. people that we have never seen before). Although familiarity is a continuous variable rather than a binary variable, much of the literature on familiarity focuses on the extremes. For convenience, this section considers unfamiliar faces and familiar faces separately before examining the transition between them during face learning. I end by summarising what it might mean to ‘know’ a fictional face.

1.1.1. *Unfamiliar face processing*

Identifying unfamiliar faces is difficult. Processing of unfamiliar faces seems to be highly image-bound. In an early demonstration of this, Bruce (1982) presented participants with some unfamiliar faces during a learning phase, but with identical images of learned faces or different images of the same person during the test phase. Recognition accuracy was about 90% when images in two phases were identical, but was only about 60% when the images were different. Unfamiliar faces were recognized more slowly and less accurately with either changes of viewpoint or expressions between study and test phase.

Although this reduction in task performance was initially thought to reflect memory failures, subsequent experiments have revealed similar difficulty in perceptual matching tasks. Bruce et al. (1999) designed an experimental method using a ‘line-up’ of face images to study how did changes in viewpoint and expression between the target and distractor faces affect the accuracy of matching unfamiliar faces. A single still target image captured from a video was presented along with an array of 10 high-quality photographs of men with similar physical appearance to the target (see Figure 1.1 for an example). Observers were required to decide if the target was present in the array, and if so, to indicate which one was the target. They were told that half of all arrays would be target absent. In target present trials, the

average accuracy (i.e., correct hit or correct rejection) for the task was only 70% even in the matched viewpoint and expression condition for both target-present and target-absent arrays. About 20% of trials, observers incorrectly decided that the target was not present (misses) and a wrong person was picked (misidentifications) on about 10% of occasions. In target absent trials, observers chose a foil from the line-up in about 30% of trials (false positives). All images in the task were taken in good lighting from similar full-face angle, on the same day. The only deliberate difference between these images was that the target images were taken by a high-quality video camera, whereas the line-up images were taken by a high-quality studio camera. Bruce et al. (1999) suggested that the difference in capture devices causes some superficial difference in quality and the appearance of faces, therefore making the comparison difficult.

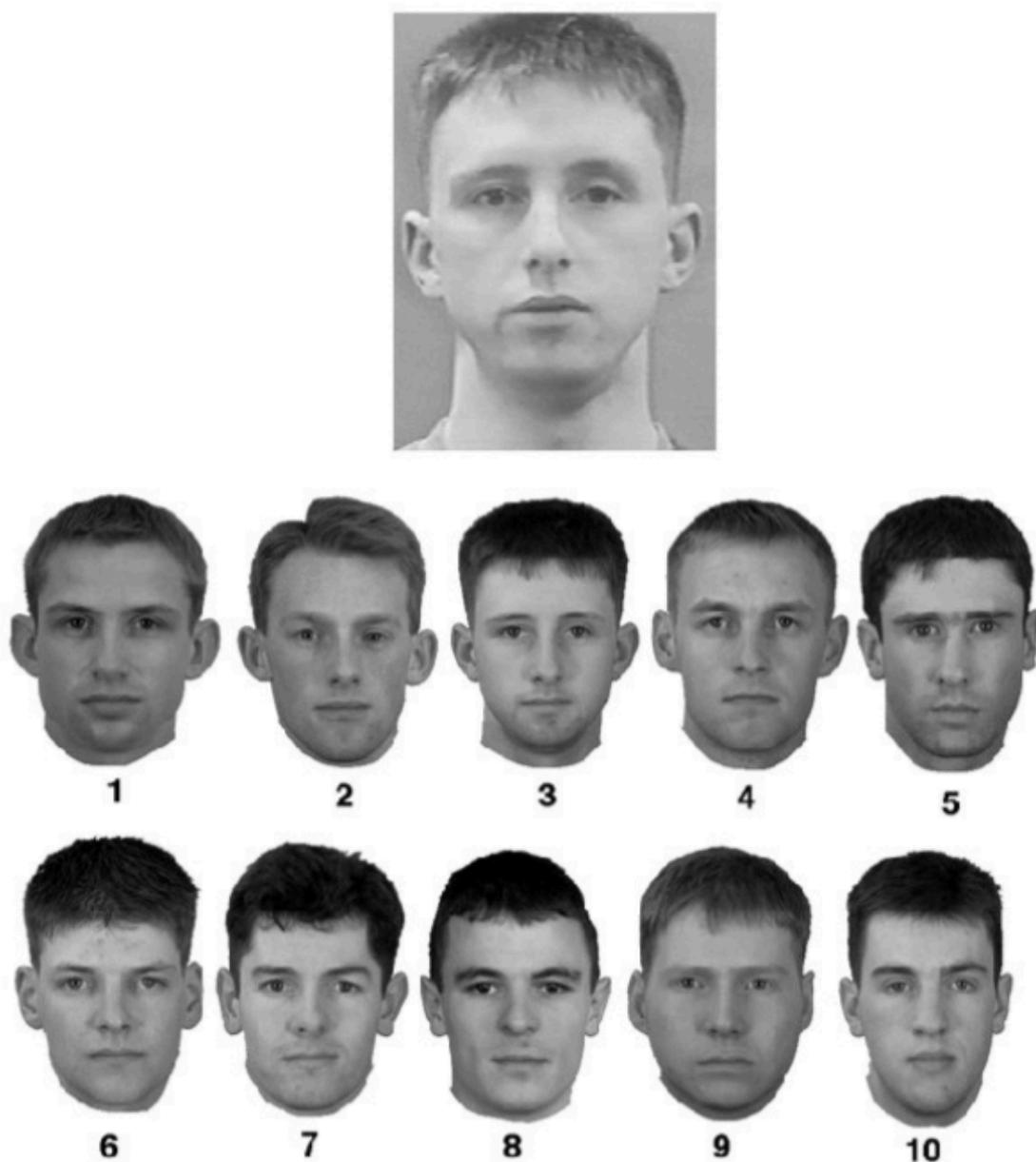


Figure 1.1. An example of the array task reproduced from Bruce et al., 1999. The target showed at the top may or may not appear in the ten images underneath. Participants' task was to decide if the target was present, and if yes, which one it was.

Later, Bruce, Henderson, Newman and Burton (2001) reduced the Bruce et al. (1999) array task to only target-present arrays that viewers were forced to choose one answer among the ten photos in the array. However, accuracy increased by only 9% compared with the finding

of Bruce et al. (1999), which was still surprisingly low. In a subsequent study, Henderson, Bruce and Burton (2001) further simplified into a single-item verification task that observers have to decide whether the two photos showed the same person. Accuracy was still low that only 45% participants correctly identified matching pairs, and about 28% incorrectly identified the target and the distractor as the same person.

Similar results were found in study matching passport photo to a live person (Kemp, Towell & Pike, 1997). In experiments conducted in a real supermarket with real staff, cashiers were asked to verify identity of customers and to decide whether to accept or reject their photo-credit cards. Each customer had four different photo-credit cards: (i) unchanged appearance card, that the photo showing the appearance of customer as on the day of shopping; (ii) changed appearance card, that the photo showing the customer with a minor paraphernalia; (iii) changed appearance card, that the photo showing a previously judged lookalike of the customer; and (iv) unmatched foil card, which with images of someone who was previously judged dissimilar to the customer. The cashiers accepted about 64% of matched foil cards and 34% of unmatched foil cards. In an applied context as in the lab, identification of unfamiliar faces is difficult and generates lots of errors.

Megreya and Burton (2006) compared Bruce et al.'s (1999) one-to-ten array task on matching unfamiliar faces and its covariation with other non-face objects (i.e., visual short-term memory, visual differentiation, and perceptual speed). Across a range of measures, participants who performed well in unfamiliar face matching tasks were also good at object matching. They further found that (i) performance on matching upright faces and matching inverted faces were highly correlated, (ii) accuracy for matching upright familiarized faces was significantly better than matching upright unfamiliar faces, and (iii) accuracy for matching upright unfamiliar faces was better than matching inverted familiarized faces. The

authors concluded that unfamiliar faces are not engaging the same processes that are engaged by familiar faces. Experiments that examine the role of different facial features in identification tasks reach similar conclusions. For familiar faces, internal features (i.e., eyes, nose, and mouth) seem to be more important in identity matching, whereas unfamiliar faces seem to rely more on external features (e.g., hairstyle and face shape) (Bonner & Burton, 2004; Osborne & Stevenage, 2008; Young, Hay, McWeeny, Flude, & Ellis, 1985).

1.1.2. Familiar face processing

Familiar face processing seems to rely on more abstract visual representations, compared with unfamiliar face processing. Although it is often hard to see that different photos of an unfamiliar face show the same person, we can recognise familiar faces over a wide range of images. It was this observation that led to the notion of a face recognition unit (FRU) in early models of face perception (e.g. Bruce & Young, 1986; Burton, Bruce, & Johnston, 1990; see Figure 1.2). In these models, the FRU functions as an image invariant representation of a particular face that is activated by any recognisable view of that face.

The IAC model of face recognition (IAC) model is a simple form of connectionist architecture including pools of simple processing units, and excitatory links connecting these units (Burton et al., 1990; Burton, Young, Bruce, Johnston, & Ellis, 1991; Burton & Bruce, 1993; Bruce, Burton, & Craw, 1992a). These links are bi-directional and initially of equal strength. *Face Recognition Units* (FRUs) are a pool of units corresponding to classification of face, with each single unit referring to a known face. These units are view-independent, such that any recognizable view of the face will cause activation in the appropriate FRU. A classification of the identity then occurs at the *Person Identity Nodes* (PINs). *Semantic Information Units* (SIUs) code information about known individuals. The model includes a recognition route for domains other than faces, such as names, and occupations. There is a

recognition route for domains other than faces. *Word Recognition Units (WRUs)* code names and other related semantic information of a person. Names (both forenames and surnames) are directly linked to *Name Recognition Units (NRUs)*, whereas other items of related semantic information (such as occupation and nationality) are connected to *Semantic Information Units (SIUs)*. *NRUs* and *SIUs* are both connected to *PINs*, where faces and identities can be connected together. Lexical output, a pool of units, are intended in processes involved in speech and other output modalities. This model included recognition route of faces and domains other than faces, and allowed comparison between different domains of face recognition (Burton, Bruce & Hancock, 1999).

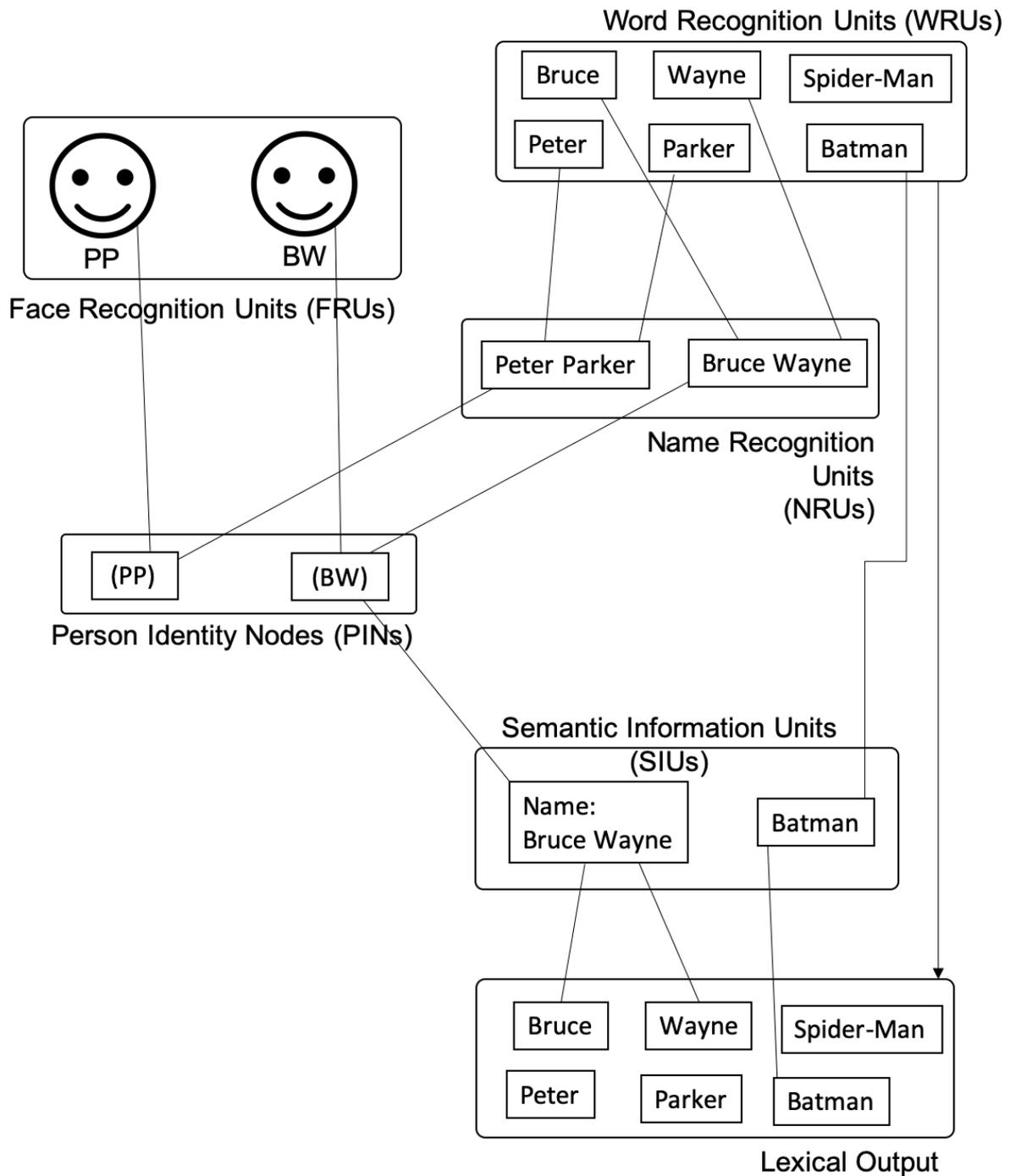


Figure 1.2. The Interactive Activation and Competition (IAC) Model of Face Recognition (Burton et al., 1990; Burton et al., 1991; Burton & Bruce, 1993; Bruce et al., 1992a)

As well as exhibiting image invariance, familiar face recognition is also robust against poor image quality. In the studies described above, high-quality full-face images were commonly used for recognition, with the goal of maximising task performance (e.g. Bruce et al., 1999).

However, there are many situations in which high-quality full-face images is unavailable, such that some security surveillance systems can only produce poor-quality videos. Burton, Wilson, Cowan, and Bruce (1999) examined people's face recognition ability in poor-quality video images, as well as how familiarity affected people's performance (see Figure 1.3 for sample images). Participants were presented with 10 poor-quality video clips and told they would be asked to identify the people in video. Afterwards, 20 high-quality face images were presented together. Participants were told that half of the 20 images were seen in video and asked to rate the level of certainty that the person appeared in the videos. Participants who were personally familiar with the targets assigned high scores to seen targets and low scores to unseen targets, whereas participants who were unfamiliar with the targets struggled to tell who was who. Although the video clips contained various cues to identity, such as body shape, clothing, and gait, the discrepancy between participants who were familiar versus unfamiliar with the target seemed to be driven by facial cues. Burton and his colleagues replicated the study, this time obscuring head, body or gait of targets in the original videos. Obscuring targets' heads significantly reduced the accuracy in recognizing the target, whereas obscuring the body or gait had negligible impact. This pattern of findings shows that familiarity significantly affects people's face recognition ability, and that viewers can recognize faces that they know even in poor quality images.



Figure 1.3. Samples of images used in face recognition task in Bruce et al. (1999): a still form a poor-quality video clip (left) and a high-quality face photograph taken in good lighting (right). Reproduced from Bruce et al. (1999).

Building on Bruce's (1982) same-image and different-image face recognition, Jenkins, White, Van Montfort and Burton (2011) drew attention to the range of variability in photos of the same face. This within-person variability is easily accommodated by viewers who are familiar with the faces concerned, but not by viewers for whom the faces are unfamiliar. In a card sorting task for face photographs, participants were given 40 different face photos (Figure 1.4), and asked to sort the photos by identity, so that photos of the same person were grouped together. What participants did not know is that the cards showed just two identities (i.e., 20 different photos of each face). Familiar viewers performed almost perfectly, correctly sorting the photos into two groups. However, unfamiliar viewers created 7 or 8 different groups on average. Apparently, unfamiliar viewers often perceived different photos of the same person as different individuals. Interestingly, they rarely mixed up photos of different

people. The difficulty seems to be cohering dissimilar images rather than separating similar faces.



Figure 1.4. The 40 face photos from Jenkins et al.'s (2011) card sorting task. Participants were required to sort the photos by identity, so that cards of the same person were grouped together. Solution: (Row 1) ABAAABABAB, (Row 2) AAAAABBBAB, (Row 3) BBBAAABBAA, (Row 4) BABAABBBB. Reproduced from Jenkins et al. (2011).

Another group of participants was asked to rate the likeness of 480 celebrity face photographs (40 celebrities, with 12 photos per celebrity). Two patterns were revealed from the ratings: (i) some photos of a person were rated for better in likeness than others, and (ii) there are differences in likeness ratings for different celebrities. The former indicated the within-person variability that some photos were better representing a person's appearance than others. The latter revealed that between-person variability is possibly due to different degree of familiarity. A strong correlation between familiarity and likeness was found. This proposed that familiarity is not simply familiar with a face but familiar over the range of

variability of the face (that the person has experienced). Increased familiarity to a face enhances the viewer's level of tolerance to image variability, so that more photos are considered as an acceptable representation of the face. Understanding between-person variability is important in telling different people apart, whereas understanding within-person variability enables people to integrate different images of the same person. Both between-person variability and within-person variability are essential in making correct judgements.

In these experiments, the familiar faces were celebrities that viewers had already learned through day-to-day exposure in the media (e.g. films and television). Films and television normally present high-quality images in which viewers can see the faces clearly. If the faces had been presented in poor quality images, it is unknown whether exposure to variability would compensate for loss of information. The next section addresses the types of information that are needed for face learning, and how much information would need to be preserved for effective face recognition.

1.2 Variability and Face Learning

Within-person variability facilitates learning of facial identities (Murphy, Ipser, Gaigg, & Cook, 2015; Ritchie & Burton, 2017). Murphy et al. (2015) presented participants with 16 trials in one of two different face learning procedures. In each trial, 48 coloured facial images, including 8 to-be-learned identities and comprising 6 images per identity (unknown to the viewers), were presented altogether in a 6 x 8 array. The training images presented consisted a wide range of within-person variability (e.g., poses, expressions, hairstyles, lighting conditions, and camera parameters). After each learning trial, participants had to judge the number of identities appeared in the array, as accurately as possible. Participants in the unique exemplar condition learned a novel set of 48 images on each training trial. Thus 96 different images of each identity could be learned after 16 trials. On the other hand,

participants in repeated exemplar condition observed the same set of images on each trial, therefore only 8 different images of each identity could be learned. Participants were then required to complete 160 test trials with 5 novel images of the 8 learned identities and 5 images of 8 novel identities. In each trial, a single greyscale test image was presented—either an image of a new identity or a novel image of a learned identity. The face images were cropped to exclude external features, thereby eliminating important cues used to recognize unfamiliar faces. Participants were required to judge if the face was present or absent during the learning trials. It was found that observers in both learning condition overestimated the number of identities presented, which is consistent with the finding of Jenkins et al. (2011). As familiarity increased, the identity estimates of all participants became gradually more accurate. In the testing phase, participants in the unique exemplar condition performed significantly better than participants in the repeated exemplar condition in hits (i.e., correctly identify a novel image of a learned identity as an image of a learned identity), but no significant difference was found in correct rejections (i.e., correctly identify an image of a new identity is not an image of a learned identity). This implies that face learning is not only determined by the time spent viewing a face, but also the degree of different face variation experienced by the learner. However, it is unclear that what specific image information supports accurate face identification. The next section considers two types of image information that are often separated in face perception research.

1.3 Face Shape and Face Texture

Analyses of face images often distinguish between face *shape* and face *texture*. In this context, shape refers to the outline of the head including external features such as hair, the configuration or layout of facial features, and perhaps specific distances between them. Texture refers to complexion, skin tone, and the pattern of light and dark across the image. These two types of information are not completely independent in visual processing. Changes

in lighting, shading, and even viewing distance can affect the perceived shape of the face. On the other hand, the shape of a face can affect the lighting and shading on that face, as when protruding features occlude regions of the face or cast them into shadow. Previous studies have shown that shape information is relatively less important than texture information to the recognition of familiar faces. For example, familiar faces remain recognizable even with the shape is distorted by stretching to several times its original height (Hole, George, Eaves, & Rasek, 2002) (Figure 1.5.A), morphing to a standard ‘anonymous’ face template (Burton, Jenkins, Hancock, & White, 2005) (Figure 1.5.B), or even morphing to the shape of another familiar face (Andrews, Baseler, Jenkins, Burton, & Young, 2016) (Figure 1.5.C). Evidently, texture can carry identity, even when divorced from shape. The converse is not true. For example, line drawings of familiar faces are poorly recognized, even though they preserve shape information (Davies, Ellis, & Shepherd, 1978; Rhodes, Brennan, & Carey, 1987). Interestingly, Bruce, Hanna, Bench, Healey, and Burton (1992) found that adding ‘mass’ (i.e. very basic *texture*) to sketches of familiar faces composed of ‘line’ (*shape*) information increased recognition performance almost to the level seen for photographic quality reconstructions (Figure 1.6.A). This kind of mass-adding is similar to the Ben Day dots or halftone dots in comic book drawing which added texture onto the simple line sketches (see Figure 1.6.B for an example). However, the ‘mass’ demonstration concerns recognition of faces that were already familiar to the viewer. It does not tell us what type of information is required to support learning of new faces.

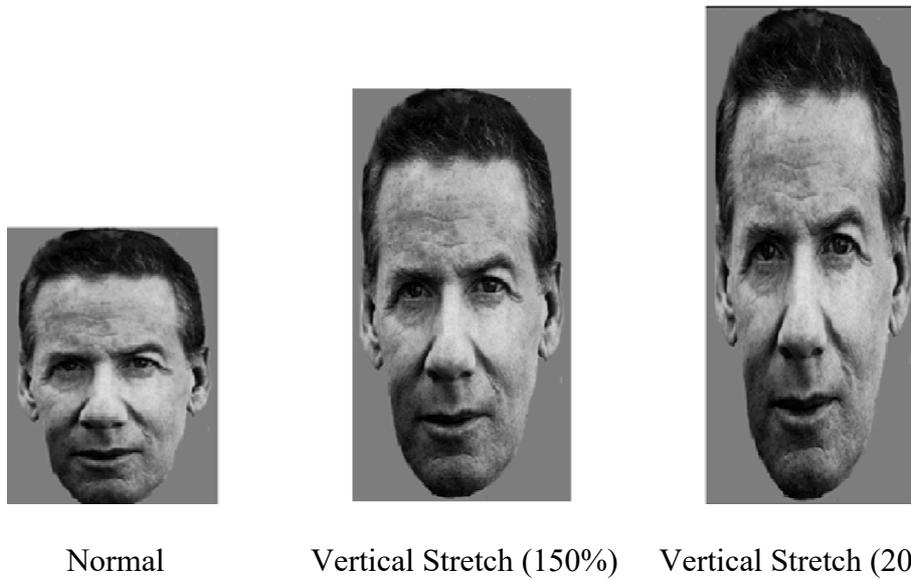


Figure 1.5.A. Example of vertically stretched faces. Familiar faces remain recognizable even when the shape is distorted by stretching to several times its original height. Reproduced from Hole et al., 2002.



Figure 1.5.B. Examples of faces morphing to a standard common shape. Familiar faces are still be recognizable in shape-free conditions. From left to right: John Travolta, Susan Sarandon, Sylvester Stallone, Leo di Caprio. Reproduced from Burton et al., 2005.

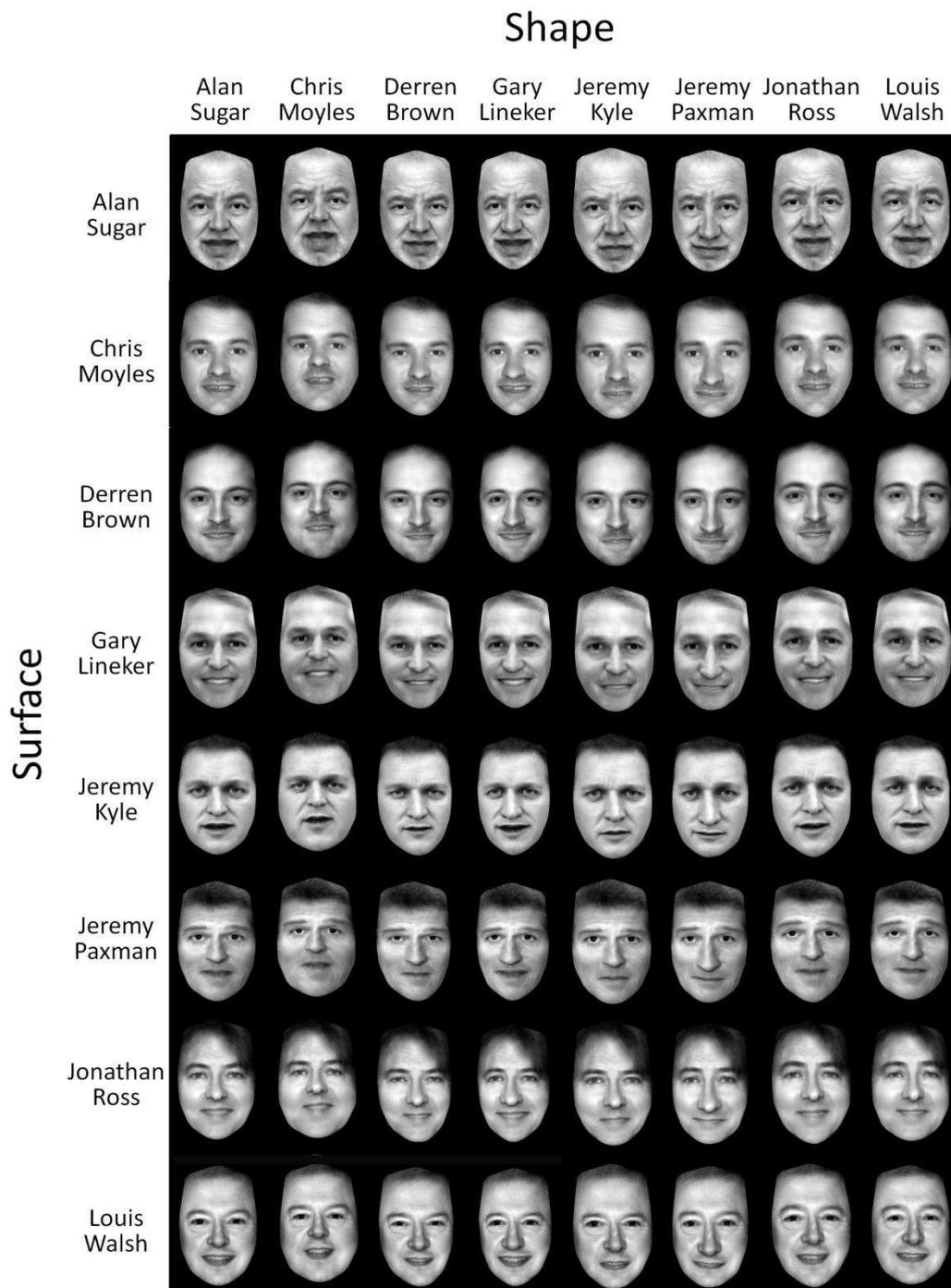


Figure 1.5.C. Examples of faces morphing to different shapes. Each individual (shown along the diagonal from the top-left to bottom-right of the Figure) were morphed into different peoples' face shapes (off-diagonal images in the Figure). Images in each row share the same surface (texture) information, and images in each column share the same shape. Images in the same row were usually regarded as the same person than images in the same column.

Reproduced from Andrews et al., 2016.



Figure 1.6.A. The mass (middle) that is added into shape (left) resulting an image with recognizability almost to the level seen for photographic quality reconstructions (right).
Reproduced from Bruce et al., 1992.



Figure 1.6.B. A comic style drawing, in which Ben Day dots (texture) were added onto simple line drawings. Illustration by Stafford (2016).

Kramer, Jenkins, Young and Burton (2017b) emphasised the importance of natural variation for learning new faces with an everyday setting – watching TV. Participants were asked to watch TV shows they had never seen before, and then their ability to recognize the actors were tested. TV shows were presented in three condition that were equivalent in terms of image variability: (i) conventional manner (natural variation), (ii) upside down (non-natural variation, texture preserved), and (iii) contrast-reversed (non-natural variation, shape preserved). As expected, participants in the natural condition learned the faces easily. However, participants were unable to learn the faces upside down or contrast-reversed, even when tested in the same format as learning. Kramer et al. (2017b) concluded that in order to support face learning, image variability must fall in the critical range that corresponds to natural, everyday variation. Moreover, gross disruption of either face shape (through inversion) or face texture (through contrast reversal) appears to rule out normal face learning. In the next section, I consider several common types of face representation, and the extent to which they preserve natural shape and texture cues.

1.4 Visual Representations of Faces

1.4.1. Photos

Although social face recognition usually involves faces that are physically present, that may not be the case in some applied situations. Often photographs are used as substitutes for real faces. Real faces are the most precise and accurate real-time representation of the face. It is possible for observers to learn as many different variations of the face as possible. However, it is not possible to learn how a person looked in the past from a real face. Alternatively, face photographs can capture the image of a person at a moment and reproduce the person's appearance at that moment even when the person no longer exists. Photographic images of faces have full (2D) shape and texture information, which is one of the best available method to capture visual information from faces and produce a representational image of the person

being captured. However, a single face photograph can only capture a single instance of the face. It does not capture all the possible ways that face can look. In addition, external factors such as lighting information and capturing device would affect how the face look from the photo.

1.4.2. Drawings

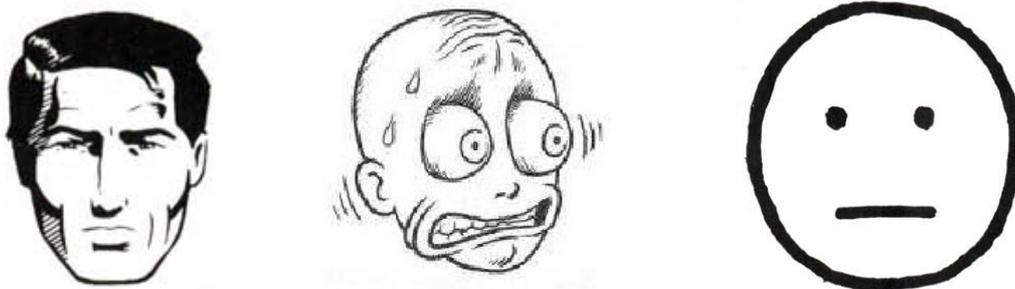
Face drawing are usually less representational than photos. However, style of drawing can vary from representational fine art portrait paintings (e.g., *Mona Lisa* by Leonardo da Vinci) to abstract line drawings of faces (e.g., *Self portrait, 1914* by Egon Schiele). Fine art portraits often retain many precise details and preserve a high degree of likeness. In some cases, they presumably maintain just a little bit less or even similar amount of shape and textural information as a photograph. As with photographs, drawn or painted photos capture a single instance of the depicted face. Thus, although we can be highly familiar with the *Mona Lisa* by da Vinci, we may not be able to recognize the model's face in another depictions, and instead identify her as another person. Unlike photos however, variation in texture information due to the specifics of environmental lighting, or characteristics of the capture device, need not constrain facial appearance in drawings or paintings.

Face drawings can also be more abstract and presented with mainly lines such as comic book or cartoon drawings. In comics, defined as “sequential art” by Scott McCloud (McCloud, 1993), different depictions of the same face are usually presented in a series (see Figure 1.7 for an example). In contrast to fine art portraits, a wider range of representations of the same face is available. This raises the prospect of learning within-person variability, albeit in a different form to the natural variability studied in previous experiments (e.g. Jenkins et al., 2011; Murphy et al., 2015; Kramer et al., 2017b). Although comics cover a wide range of styles, texture information is often simplified in comic drawings (perhaps more akin to

‘mass’; Bruce et al., 1992). Shape information is sometimes well preserved, sometimes exaggerated, and sometimes not based on real face shape at all (see Figure 1.8 for examples). Drawings are an important means of visualizing fictional characters, as there is no prospect of photographing fictional characters—or encountering them in social situations. Given that drawings are generated by the artist, they may provide insight into the artist’s mental representation of the fictional face.



Figure 1.7. Different depiction of the same face in sequential scenes from *Spider-Man: Death of the Stacys* (Lee, Kane & Romita, 2007)



(a) Well Preserved Shape (b) Exaggerated Shape (c) Not Based on Real Face Shape

Figure 1.8. Examples of face in comics with (a) well preserved shape (McCloud, 1993), (b) exaggerated shape (McCloud, 2011), and (c) not based on real face shape (McCloud, 1993).

1.5 Non-visual Representations of Faces

1.5.1. Verbal descriptions

Faces are not presented only through visual images. Verbal descriptions of faces do not have face shape or face texture, although they can convey both types of information (e.g., a big nose; rosy cheeks). Face descriptions are routinely used to develop fictional characters in novels (e.g., Sherlock Holmes, James Bond, and Harry Potter). They are also used in forensic settings (e.g., eyewitness testimony; Laub & Bornstein, 2008; Wells & Olson, 2003). In both cases, the idea is that by reading face descriptions, a person could learn about and visualize a face he or she has never seen before. The transition from words to images is completed when descriptions provided by eyewitnesses are essential to creating a face composite or sketch as well as the selection of foils in photo or live line-ups (Brandl, 2014). Casting for characters of a novel-based movie or television series is another situation in which descriptions of facial appearance are taken into account. In casting for television series, the writer generally introduces small parts with a minimum of physical descriptions and abstracts the character descriptions (Turow, 1978). Visualizing a fictional role is not based solely on physical appearance illustrations in the original work, but also on descriptions of character traits. As a result, discrepancies in the physical appearance of actors and actresses being selected can arise. For example, James Bond, a fictional character created by Ian Fleming, has been portrayed by seven different actors in James Bond movie series. Needless to say, those seven actors differ somewhat in their physical appearance, although they generally conform to the original written descriptions in the novels.

1.5.2. Social inferences

Many studies have found that people automatically make important social and personality inferences based on physical appearances (e.g., Oosterhof & Todorov, 2008; Todorov, 2008; Todorov, Mandisodza, Goren & Hall, 2005; Todorov, Said, Engell & Oosterhof, 2008; Willis

& Todorov, 2006; Zebrowitz & Montepare, 2008a, 2008b; Zebrowitz, Hall, Murphy & Rhodes, 2002). When viewing unfamiliar faces, adequate information for social judgements can be gathered in as little as tenth of a second (Willis & Todorov, 2006).

Oosterhof and Todorov (2008) generated a two-dimensional (trustworthiness and dominance) model of first impressions from faces by applying principal component analysis (PCA) to viewers' ratings of face images. PCA is a dimension reduction technique that delivers a small number of axes, ordered by the amount of variance explained. The two cardinal dimensions appear to code facial features signalling physical strength (the dominance dimension) or expressions signalling approach or avoidance (the trustworthiness dimension; see Figure 1.9). Using this two-dimensional model, an unlimited number of artificial faces could be generated, and face variations along a given dimension could be mapped.

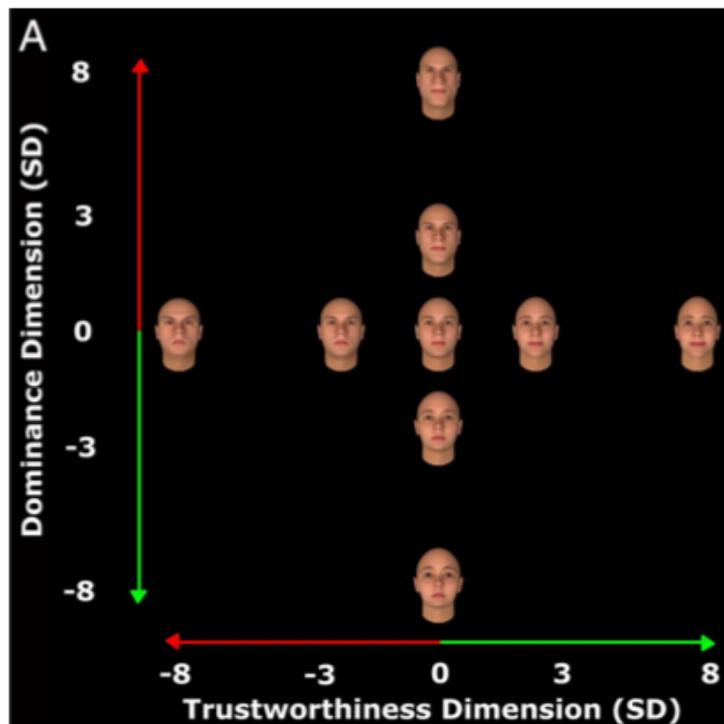


Figure 1.9. A 2D model of face evaluation, along the dimensions of dominance and trustworthiness. Reproduced from Oosterhof & Todorov, 2008.

Sutherland et al. (2013) extended this approach to examine first impressions for natural face photographs (as opposed to computer generated faces). Their analysis was based on 1000 ambient face images (500 male and 500 female Caucasian adults). These faces images were intentionally chosen to display a wide range of variability, including but not limited to age, expression, pose, facial hair, and accessories. The photographs were taken with different cameras, from different angles, with different backgrounds and lighting. The intention was to generate a model that applied to natural face photographs that span the range of variability encountered in real life. All facial photographs were rated on trustworthiness, approachability, degree of smiling, attractiveness, intelligence, dominance, sexual dimorphism, skin tone, confidence, aggressiveness, age and baby-facedness using 7-point Likert scales. Traits were rated in separate blocks to prevent carryover effects. After that, 20 face photographs that rated highest and lowest in each of seven selected traits (i.e., age, sexual dimorphism, attractiveness, intelligence, trustworthiness, dominance, and confidence) were averaged and morphed respectively. Each pair of average images (high and low) was morphed in steps of 10%, creating a morphed continuum with 11 different images for each of the seven traits (77 images in total). The resulting images were then rated on the same trait dimensions that were used for the source photographs. The ratings were consistent with the manipulated level of their corresponding traits, indicating that the traits on morphed photograph in the continuums were successfully manipulated. The results showed that consistent facial cues subserve a range of social inferences, even for non-standardized photographs with high variations.

The ambient face images were then used to test the Oosterhof and Todorov (2008) two-dimensional (trustworthiness and dominance) model. The two-dimensional model was replicated, and approachability, youthful-attractiveness, and dominance were merged into the model using principal axis factor analysis, see Figure 1.10 for an example of the three-

dimensional map of faces. The new three-dimensional model was cross-validated using face averages directly constructed from the factor scores.

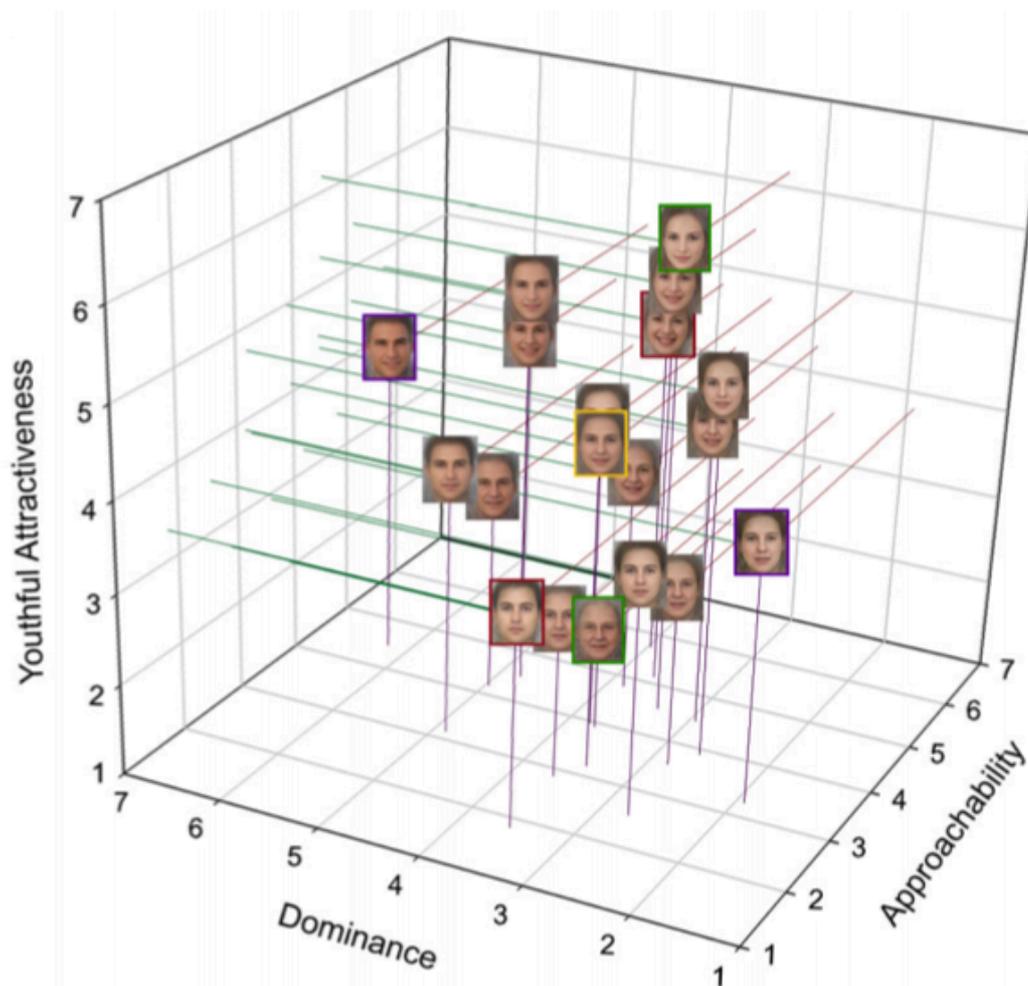


Figure 1.10. A 3D map of faces, along the dimensions of approachability, dominance, youth attractiveness. Reproduced from Sutherland et al., 2016.

This study showed that trait ratings could be a systematic mean to retrieve social inferences from faces. A relatively small set of ratings (two or three numbers) allows us to turn social inferences into quantitative data, so that statistical comparisons can be made. This is an important result for the central aim of this thesis. Because social inference ratings are abstracted away from the visual properties of faces, they can potentially apply to different types of face representation, including drawings, written descriptions, and mental images.

This suggests a route into quantifying representations of fictional faces (for which photographs do not exist), and may provide a ‘common language’ by which different types of face representation can be compared.

1.6 Mental Imagery and Face Perception

Mental imagery, a visual representation in the absence of environmental input, engages many neural and cognitive mechanisms that are involved in visual perception. For example, research with brain-damaged patients has shown that visual imagery and visual perception share specific and specialized mechanisms (Kosslyn, Ganis, & Thompson, 2001). Among the cortical areas activated by imagery and by perception, about two thirds are activated in common by both functions (Kosslyn, Thompson, & Alpert, 1997). Cognitive control processes function comparably in both imagery and perception, however, more overlap in activation was found in the frontal and parietal lobes than occipital and temporal lobes, that was interpreted that perception relies partly on bottom-up processes that are not used in imagery (Ganis, Thompson, & Kosslyn, 2004).

In addition to evidence from neuroscience, imagery and perception are closely tied at the behavioural level. Kosslyn, Sukel, and Bly (1999) examined difference between imaginary and perceptual comparisons on arrays of stripes (see Figure 1.11). Participants were asked to compare four quadrants, each containing a set of stripes, according to their length, spacing, orientation, or width. Participants made similar errors whether performing the task via imagery or perception. In addition, more time was needed to evaluate high-resolution patterns in imagery than perception, suggesting that additional effort is required in imagery to represent high resolution information. Although imagery and perception may activate common brain regions, it may be more difficult to process visual detail in imagery, relative to normal perception.

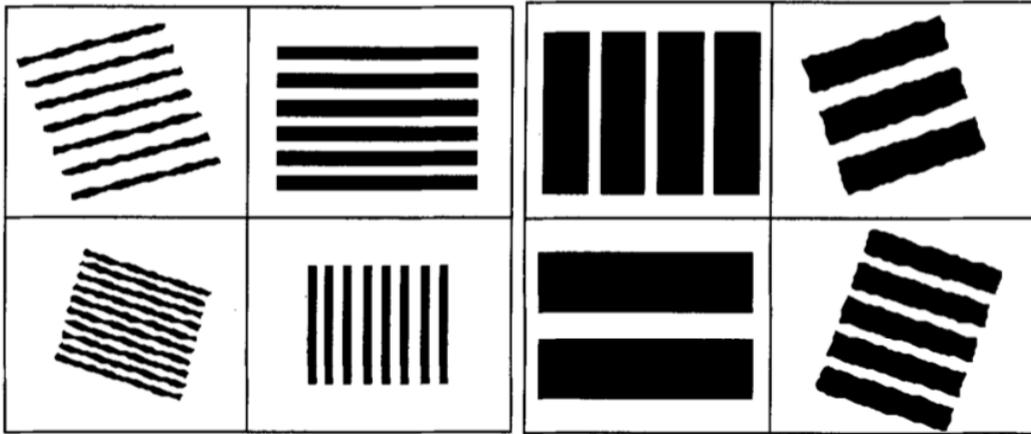


Figure 1.11. Two sets of gratings with arrays of stripes. Participants' task was to compare the for quadrants according to their length, spacing, orientation, or width by viewing (perception) or visualizing (mental imagery) them. Reproduced from Kosslyn et al., 1999.

Faces are often claimed to be a 'special' category of visual object, in part because of their distinct neural signature. O'Craven and Kanwisher (2000) demonstrated that brain regions specialized for face perception were also activated during mental imagery of faces.

Activation during face imagery was maximal in the lateral fusiform face-selective region, which also responds maximally during face perception (Ishai, Ungerleider, & Haxby, 2000). Greater response magnitudes for was found in perception of faces than in imagery of faces. (O'Craven & Kanwisher, 2000). However, this may not be surprising if imagery requires greater effort than normal perception, or involves less visual detail (Kosslyn, et al., 1993).

All of these studies relied on neuroimaging techniques to compare patterns of brain activation during face imagery and face perception. Participants in these experiments were instructed to generated mental images of familiar faces (Ishai et al., 2000) or famous faces (O'Craven & Kanwisher, 2000) from long-term memory, or to imagine faces that they had learned during the experiment (O'Craven & Kanwisher, 2000). In this thesis, I would like to deal with faces that participants have never seen in the flesh—faces of fictional characters. In the case of

fictional faces, the mental representation cannot be built up in the normal way, that is, through visual exposure to the live face, or to photographic or film footage (bottom-up perception). Instead it must be built ‘from the inside’, based on semantic or emotional descriptions (top-down perception), or from visual approximations of the imagined face (e.g. drawings or dramatic depictions).

Although this topic draws on many of the areas reviewed in this chapter, to my knowledge, it has not been directly addressed by any previous studies. Consequently, there are many open questions. One of the questions I tackle in this thesis is whether it is possible to learn a fictional face, so that the fictional face comes to show the behavioural hallmarks of familiar face recognition. A related question is whether radically different visual representations of a fictional face can cohere into a single mental representation. In pursuing these questions, I will test representations in which the visual component is increasingly sparse. Ultimately, I will examine people’s ability to transform written descriptions of fictional faces into mental images. In so doing, I will ask whether this transformation could be mediated by social inferences, how much convergence there is among different readers, and the extent to which descriptions of physical traits and character traits contribute to overall impressions of fictional faces. These are novel research questions, and answering them requires novel research methods. In developing the experiments in this thesis, I borrow several techniques from mainstream face perception research, and modify them to address the questions outlined here.

1.7 General Methodological Approach and Overview

The processing of fictional faces was examined here by adapting standard behavioural methods that were developed to address normal face perception, such as card-sorting (Chapter 2), line-up identification (Chapter 4), and face matching (Chapter 4) tasks. Chapter

2 examined familiarity effects in identifying fictional faces using Jenkins et al.'s (2011) card-sorting task. Photographic images of real faces were replaced with varied portrayals of each fictional character's face—comic drawings by different artists and movie stills featuring different actors. Participants who were familiar or unfamiliar with the characters were required to sort the face cards according to *the fictional identities* that the cards represented.

Chapter 3 attempted to distinguish the two fictional characters used in Chapter 2 based on social inference ratings (*trustworthiness, dominance, attractiveness, and age*). Linear Discriminant Analysis (LDA) used to separate the images by identity, and compared to human perceptual performance.

The experiments in Chapter 4 examined face identification across drawings and photos with simple cartoon drawings that were based on a single photograph. The line-up task of Bruce et al. (1999) and the paired matching task of Henderson et al. (2001) were adapted for use with drawings of faces. Two experiments assessed the effect of a change in image format on identification accuracy.

Chapters 5 and 6 examined transformations between written descriptions of fictional faces and mental pictures in the reader's imagination. Todorov et al. (2008) developed social inference ratings for measuring first impressions from faces, and connecting these impressions to aspects of facial appearance. The experiments in Chapter 5 used similar ratings to project (i) written descriptions of fictional faces, and (ii) photographs of real faces into a common metric space. Participants' mental imagery could be visualised by asking them to select the best match photograph for a given description. In Chapter 6, the software package InterFace (Kramer, Jenkins, & Burton, 2017a) was used to refine these visualisations by creating weighted averages of the chosen photographs. The resulting images were

presented in a series of perceptual tasks to assess the influence of different types of descriptive information on mental imagery for fictional faces.

Chapter 2

Familiarity and Identification for Fictional Faces

Introduction

A key insight from the Jenkins et al. (2011) card-sorting task is that sorting photos by identity is much more efficient for familiar face than for unfamiliar faces. In the original version of the task, familiar viewers correctly sorted the photos into two identities, whereas unfamiliar viewers perceived many different identities in the set. For all participants, mixing the two identities was rare. Andrews, Jenkins, Cursiter and Burton (2015) replicated this card-sorting task with a constraint that participants were told the number of identities presented. Unlike the original card-sorting task, in which participants were free to create as many or as few identities as they liked, participants were highly accurate in this modified card sorting. Participants had little difficulty in distinguishing between faces when they were told there were only two identities, and rarely mistook one person for the other. These results show that unfamiliar faces can be cohered on the basis of visual information, but only when viewers are specifically encouraged to do so (e.g. given the additional information - number of existing identities). Moreover, Andrews et al. (2015) found that exposure to within-person variability in this two-sort task improved performance in a subsequent face-matching task. The authors concluded that exposure to this variability led viewers to form stable mental representations of the faces concerned, and that these stable representations allowed them to recognise previously unseen images of those faces.

In a separate experiment (Jenkins et al., 2011), participants were recruited to rate the likeness of 480 celebrity face photographs (40 celebrities, with 12 photos per celebrity). Two patterns emerged from these ratings. The first was that some photos were seen as better likenesses than others. The second was that there were differences in likeness ratings for different celebrities. The latter finding may seem puzzling at first, but can be explained by differing degrees of familiarity. As it turned out, there was a strong correlation between familiarity and

likeness. Apparently, increased familiarity to a face enhances the viewer's level of tolerance to image variability, so that more photos are considered an acceptable representation of that face. Between-person variability is important in telling people apart, and within-person variability is important for 'telling people together'. To make accurate identity judgements, it is necessary to understand both forms of variability.

What is it that varies? Some previous studies have broken face images into shape information and texture information. Faces that present shape information only are generally hard to recognise, but can sometimes be made more recognizable through caricature, which emphasises idiosyncratic characteristics of face shape. Rhodes et al. (1987) generated three types of drawing representations (i.e., veridical line drawings, caricatures, and anti-caricatures) algorithmically. Veridical line drawings represented a face by 37 lines, based on a fixed set of 169 points. Caricatures were created by exaggerating all metric differences between a face and a norm, whereas anti-caricatures were produced by reducing all differences between a face and a norm. Participants were required to identify the person in the drawings. There were 18 faces altogether. For each face, one-third of the participants saw the veridical line drawing, one-third saw the caricature, and the remaining third saw the anti-caricature. Caricatures of familiar faces were identified more quickly than veridical line drawings or anti-caricatures. Both veridical line drawings and caricatures preserve less texture information than photos. However, caricatures capture (or even exaggerates) the essential characteristics or shape of a face, and, therefore, is 'super-fidelity' (Rhodes et al., 1987). However, using Principal Component Analysis (PCA), Hancock, Burton and Bruce (1996) found that texture information is more dominant. Models based solely on shape information provided extremely poor identification rates (Hancock et al., 1996, 1998). Thus, while shape information may provide good support for face recognition, it does not appear to be the dominant cue. In current experiment, I specifically focus on texture information. For

stimuli in which the texture information is impoverished (e.g. face drawings), is exposure to within-person variability enough to support face learning, such that standard familiarity effects emerge?

Before choosing stimuli presented in the study, two main constraints of the design were considered. First, I would like the same faces to be presented both as photos (full texture information) and as drawings (impoverished texture information). Second, I would like these faces more familiar to some participants and less familiar to others, so that effects of familiarity can be examined. To this end, I chose two well-known fictional characters that have been portrayed in movies by different actors, and depicted in comic books by different artists: Bruce Wayne (Batman) and Peter Parker (Spider-Man).

Experiment 1

The purpose of this experiment was to test whether facial identity can be learned from variability in drawings of the same face, and if so, whether standard familiarity effects emerge. To this end, I developed two separate versions of the card-sorting task (Jenkins et al., 2011) based on movie captures and comic book drawings of the same fictional characters (Bruce Wayne and Peter Parker). Each of these fictional characters was portrayed by different actors and by different artists. The main question is whether viewers can learn faces from exposure to variation, even when the information in the images is impoverished (e.g. reduced texture information in the drawings).

Methods

Participants

A total of 60 students [38 female, 22 male; mean age (*SD*): 22.85 (6.06); age range: 18 – 45], were recruited from the University of York. All participants gave written informed consent

prior to the experiment and took part in the study in exchange for a small payment or course credit.

Stimuli and Design

Participants were presented with two sets of stimuli containing 40 photographs and 40 drawings respectively, as listed in *Figure 2.1*. All these images were printed in their original colour on laminated cards measuring 20 × 29 mm each. For each set of 40 cards, 20 “Bruce Wayne”, they were instructed to sort 40 face cards (either photographs or drawings) by fictional identity. The 40 cards contained 2 fictional characters presented by 4 different actors (photographs) or artists (drawings). They were asked to cleave the cards by the fictional character each card represents, so that cards of the same fictional character were grouped together, and cards of different fictional characters were placed in separate piles. They were free to sort the cards into as many or as few groups as they wished. In each group, they were free to have as many or as few cards as they like. There was no time limitation on this task. After the first task, participants were asked to repeat the procedure with the other set of stimuli (either drawings or photographs). Finally, participants completed the questionnaire.

Results and Discussion

Sorting cards by format

As expected, all participants performed perfectly in the practice task when sorting the cards by image format. All cards were correctly sorted into two piles of 40 drawings and 40 photos.

Sorting the photos by identity

Card-sorting performance was initially evaluated by two measures: number of perceived identities and number of misidentification errors. Number of perceived identities is the number of image piles that participants sorted the 40 photographs into, each pile representing one different perceived identity. Number of misidentification errors is the number of times participants' image piles featured more than one fictional identity (i.e., both Peter Parker and Bruce Wayne). For instance, in a pile with 2 'Peter Parker' and 3 'Bruce Wayne', the minority (2 'Peter Parker') scored 2 in misidentification errors. These scores were summed to create the misidentification errors score of each participant (maximum score = 20).

Descriptive statistics on card-sorting performance measures are presented in Table 2.1.

Table 2.1. Descriptive statistics on photo card-sorting performance measures.

	Mean (<i>SD</i>)	Median	Mode	Range
Number of perceived identities ^a (<i>max</i> = 40)	4.78 (2.75)	4	2	2-15
Number of misidentification errors ^b (<i>max</i> = 20)	.55 (1.67)	0	0	0-8

^a Number of perceived identities is the number of image piles (each representing a different perceived identity) that participants sorted the 40 photographs into.

^b Number of misidentification errors is the number of times participants' image piles featured more than one fictional identity (i.e., both Peter Parker and Bruce Wayne).

Simple analysis Consistent with Jenkins et al. (2011), images of the same identity were often perceived as different identities, meaning that participants divided Peter Parker and Bruce Wayne into several piles (median = 4, mode = 2, range = 2-15). However, misidentification errors were rare (median = 0, mode = 0, range = 0-8), meaning that images of the two different identities were seldom confused.

These two performance measures (number of perceived identities and misidentification errors) were originally used to compare responses to familiar and unfamiliar faces (Jenkins et al., 2011). Because standard familiarity effects are so large in identification tasks, they are easy to detect even with crude measures. However, these measures ignore much of the information that is available in participants' solutions. For example, simply counting the piles of cards does not differentiate between a solution comprising three piles of equal size (a moderate solution), and a solution comprising two large piles plus one singleton (an almost perfect solution). The current study addresses research questions that are much more subtle than the original familiarity contrast, as the identities involved were fictional characters, and each character was represented by different actors.

Matrix analysis A few recent studies have developed more sophisticated analyses of card-sorting tasks that use all of the available data (e.g. Neil, Cappagli, Karaminis, Jenkins & Pellicano, 2016; Yan, Andrews, Jenkins & Young, 2016). Following Neil et al. (2016), a confusion matrix was created to visualise and integrate task performance across participants (see Figure 2.2).

In this matrix, all 40 images (20 of Peter Parker and 20 of Bruce Wayne) were placed along both x- and y- axes. Each cell in the matrix represents the total number of times two different images were placed into the same pile by the 60 participants, in other words, the number of participants who placed the two images together in the same pile (minimum score = 0; maximum score = 60). Perfect performance would result in maximum scores (60) for every cell in the match quadrants (top-left and bottom-right), and minimum scores (0) for every cell in the mismatch quadrants (top-right and bottom-left). Non-zero values in the latter two quadrants indicate misidentification errors, that is, photos of different identities being grouped together.

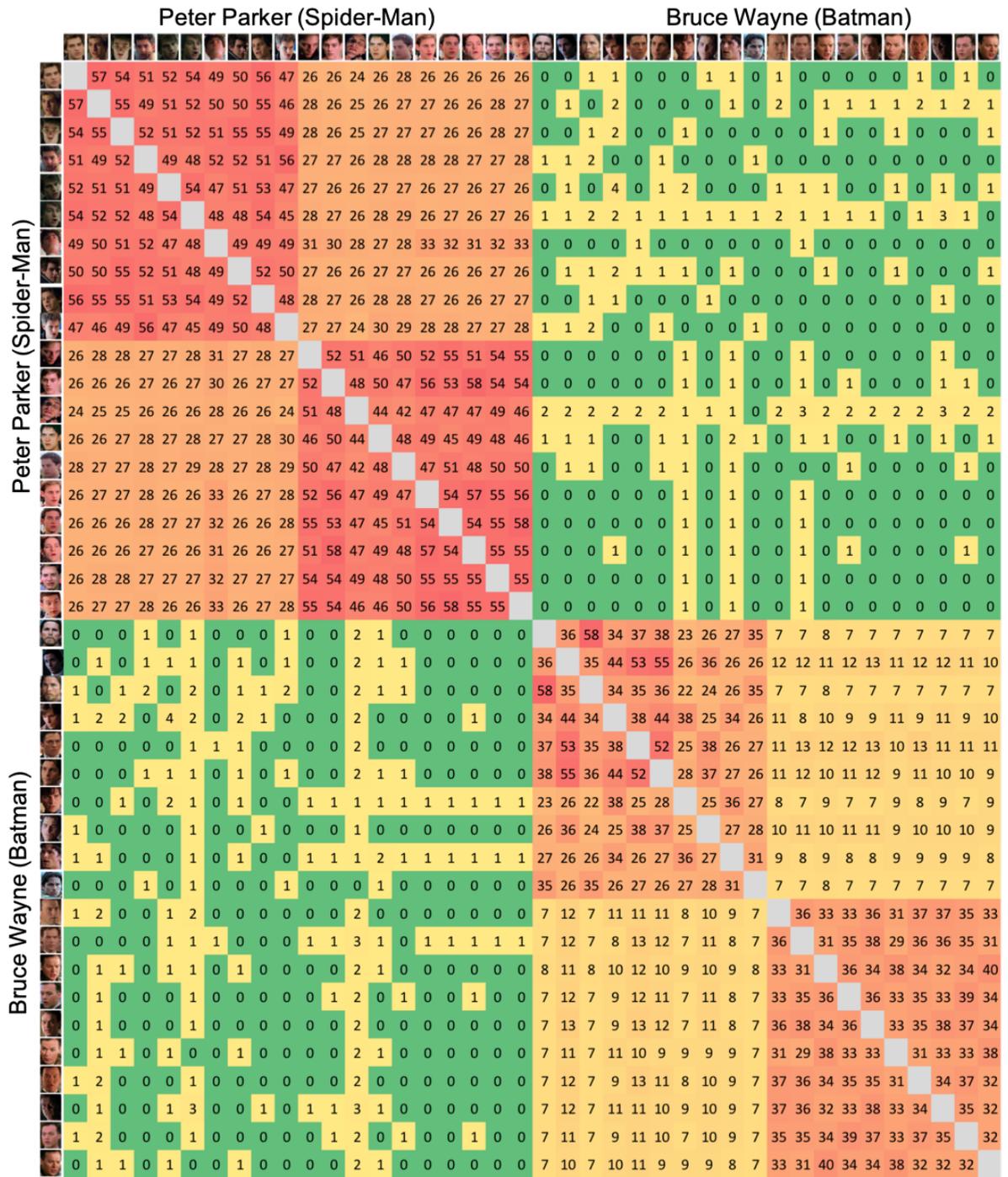


Figure 2.2. Response matrix for 60 participants in the photo card-sorting task. The x- and y-axes indicate the 20 cards of each of 2 identities (Peter Parker/Bruce Wayne). Each cell in the matrix represents the number of times two different photos were sorted into the same pile by the 60 participants. Therefore, values reflect the number of participants placing the two images together in the same pile. Perfect performance would result in scores of 60 in the match quadrants (top-left and bottom-right), and scores of 0 in the mismatch quadrants (top-

right and bottom-left). Non-zero values in the latter two quadrants specify misidentification errors, that is, photos of different identities being grouped together. Cell values are highlighted in a green (low scores) to yellow and then red (high scores) colour gradient. Because an image cannot be matched with itself, blank cells run in a diagonal line from the top-left to bottom-right corners of the matrices.

Figure 2.2 shows a clear distinction among the four major quadrants. The match quadrants (top-left and bottom-right) contain moderate-to-high values, whereas the mismatch quadrants (top-right and bottom-left) contain very low values. Some structure is also visible within each quadrant. For example, the match quadrants show some separation between different actors portraying the same character.

To assess the effect of identity in this sorting task, a paired sample *t*-test was conducted to compare the average score for match cells (the number of times each card appeared in a pile with the same perceived identity, $M = 32.60$, $SD = 6.12$) and the average score for mismatch cells (the number of times each card appeared in a pile with different perceived identities, $M = .51$, $SD = .42$). Scores were significantly higher in match cells than in mismatch cells [$t(39) = 32.84$, $p < .001$], confirming that participants could reliably group these fictional characters, across changes in actor. As in the standard task with real identities (Jenkins et al., 2011), merge errors were rare.

Having summarised overall performance, I next set out to examine effects of familiarity in this task.

Familiarity effect in photos of fictional characters

Participants' level of familiarity towards the faces of the fictional characters was measured by the average scores (out of 10) of 'I am familiar with the face of Bruce Wayne', and 'I am

familiar with the face of Peter Parker'. Participants were ranked by this familiarity score to allow comparison of high- and low-familiarity subgroups. An independent samples *t*-test confirmed that participants in the upper quartile ($N = 15$, $M = 9.03$, $SD = .72$) were significantly more familiar with these fictional faces than participants in the lower quartile ($N = 15$, $M = 1.97$, $SD = .81$) [$t(28) = 25.24$, $p < .001$].

Simple analysis To test for an effect of familiarity, I first compared the number of perceived identities and number of misidentification errors for high-familiarity (upper quartile) participants and low-familiarity (lower quartile) participants (see descriptive statistics in Table 2.2).

Table 2.2. Descriptive statistics by familiarity on photo card-sorting performance measures.

	Low-Familiarity Participants	High-Familiarity Participants
Number of perceived identities ^a ($max = 40$)		
Mean (<i>SD</i>)	5.33 (2.29)	4.87 (3.58)
Median	5	4
Mode	5	2
Range	2-11	2-15
Number of misidentification errors ^b		
Mean (<i>SD</i>)	1.00 (2.48)	.53 (2.07)
Median	0	0
Mode	0	0
Range	0-8	0-8

^a Number of perceived identities is the number of image piles (each representing a different perceived identity) that participants sorted the 40 photographs into.

^b Misidentification errors is the number of times participants' image piles featured more than one fictional identity (i.e., both Peter Parker and Bruce Wayne).

Independent sample *t*-tests showed that although the high-familiarity participants created fewer piles ($M = 4.87$, $SD = 3.58$) than the low-familiarity participants ($M = 5.33$, $SD = 2.29$), this difference was not significant [$t(28) = .61$, $p = .549$]. There was also no significant difference in the number of misidentification errors between high-familiarity participants ($M = .53$, $SD = 2.07$) and low-familiarity participants ($M = 1.00$, $SD = 2.48$) [$t(28) = .56$, $p = .580$]. Given the relative insensitivity of these basic measures, I next undertook a matrix analysis to compare the performance of these subgroups more thoroughly.

Matrix analysis Figure 2.3 shows separate confusion matrices for the high- and low-familiarity subgroups. As with the overall data, both matrices show clear distinctions among the four major quadrants. Specifically, match quadrants (top-left and bottom-right) contain higher scores than the mismatch quadrants (top-right and lower-left) overall. Within match quadrants, there is some separation between different actors portraying the same character, but this is much less pronounced.

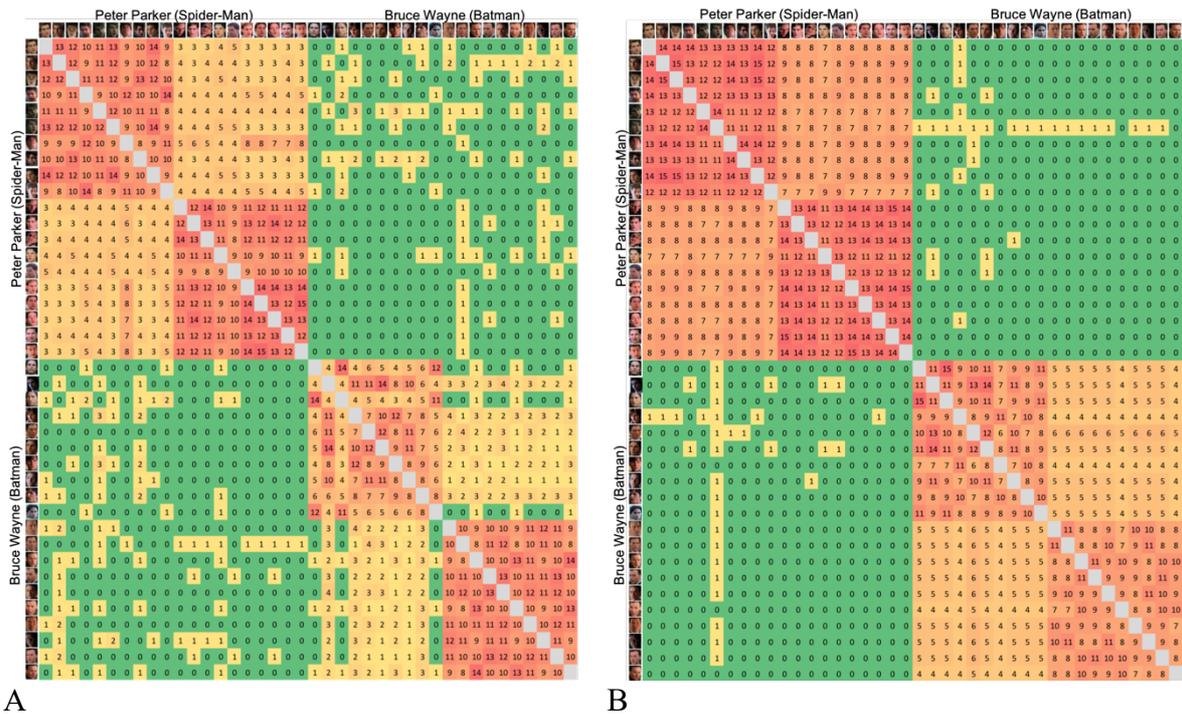


Figure 2.3. Response matrix for (A) low-familiarity and (B) high-familiarity participants in the photo card-sorting task. The x- and y- axes indicate the 20 cards of each of 2 identities (Peter Parker/Bruce Wayne). Each cell in the matrix represents the number of times two different photos were sorted into the same pile across 15 low-familiarity and high-familiarity participants. Therefore, values reflect the number of participants placing the two images together in the same pile. Perfect performance would result in scores of 15 in the match quadrants (top-left and bottom-right), and scores of 0 in the mismatch quadrants (top-right and bottom-left). Non-zero values in the latter two quadrants specify misidentification errors, that is, photos of different identities being grouped together. Cell values are highlighted in a green (low scores) to yellow and then red (high scores) colour gradient. Because an image cannot be matched with itself, blank cells run in a diagonal line from the top-left to bottom-right corners of the matrices.

To compare performance for the two subgroups, I first computed a ‘match score’ (averaged values from match cells) and a ‘mismatch score’ (averaged values from mismatch cells), separately for the low-familiarity subgroup and the high-familiarity subgroup (see Table 2.3 for descriptive statistics).

Table 2.3. Descriptive statistics of confusion matrix measures for low- and high-familiarity subgroups.

	Low-Familiarity Participants	High-Familiarity Participants
Match Score ^c		
Mean (<i>SD</i>)	6.21 (1.35)	8.57 (1.80)
Median	6.42	8.42
Mode	7.26	6.89, 10.26, 10.68
Range	3.00-7.89	5.84-10.74
Mismatch Score ^d		
Mean (<i>SD</i>)	.25 (.19)	.09 (.14)
Median	.20	.05
Mode	.15	.05
Range	.00-.80	.00-.85

^c Match score is the averaged values in the cells at the interceptions of a target card and the cards with the same identity as the target card.

^d Mismatch score is the averaged values in the cells at the interceptions of a target card and the cards with a different identity as the target card.

A two-way ANOVA was conducted to examine the effects of identity (match and mismatch) and familiarity (high- and low-familiarity participants) on sorting performance (the number of times two cards has been placed into the same pile).

There was a significant main effect of identity, with higher values for match scores ($M = 7.39, SD = 1.98$) than for mismatch scores ($M = .17, SD = .19$) [$F(1, 39) = 899.29, p < .001, \eta_p^2 = .96$]. This analysis confirms that cards showing the same identity were more frequently placed in the same pile (perceived as the same identity) than cards sharing different identities.

There was also a significant main effect of familiarity, with higher overall scores for high-familiarity participants ($M = 4.33$, $SD = 4.46$) than for low-familiarity participants ($M = 3.23$, $SD = .3.14$) [$F(1, 39) = 173.32$, $p < .001$, $\eta_p^2 = .82$]. This means that, overall, high-familiarity participants grouped cards together more frequently than low-familiarity participants.

The interaction between these factors was also significant [$F(1, 39) = 219.75$, $p < .001$, $\eta_p^2 = .85$]. Simple main effects analyses were conducted to further investigate this interaction. At the levels of familiarity, simple main effects of identity were significant at both high familiarity [$F(1, 78) = 1103.78$, $p < .001$, $\eta_p^2 = .93$] and low familiarity [$F(1, 78) = 542.64$, $p < .001$, $\eta_p^2 = .84$], that match scores were consistently higher than mismatch scores. At the levels of identity, simple main effects of familiarity at mismatch scores were not significant [$F(1, 78) = 1.96$, $p > .05$, $\eta_p^2 = .02$], that there were no significant differences between mismatch scores for high or low familiarity. However, simple main effects of familiarity at match score was significant [$F(1, 78) = 392.31$, $p < .001$, $\eta_p^2 = .83$], that high-familiarity participants cohered the cards of same identity more frequently than low-familiarity participants.

Sorting the drawings by identities

As with the analysis of photo card sorting performance, performance on drawings card-sorting was evaluated by the simple analysis (with measures: number of perceived identities and number of misidentification errors). Descriptive statistics on card sorting performance measures are presented in Table 2.4.

Table 2.4. Descriptive statistics on drawing card-sorting performance measures.

	Mean (<i>SD</i>)	Median	Mode	Range
Number of perceived identities ^a (<i>max</i> = 40)	5.35 (2.57)	5	3	2-13
Number of misidentification errors ^b (<i>max</i> = 20)	1.27 (3.04)	0	0	0-15

^a Number of perceived identities is the number of image piles (each representing a different perceived identity) that participants sorted the 40 photographs into.

^b Misidentification errors is the number of times participants' image piles featured more than one fictional identity (i.e., both Peter Parker and Bruce Wayne).

Simple analysis Consistent with Jenkins et al. (2011) and our photo cards sorting task, drawings of the same identity often misidentified as different identities, meaning that participants divided Peter Parker and Bruce Wayne into several piles (median = 5, mode = 3, range = 2-13). However, drawings of the two different identities were seldom confused, that misidentification errors were rare (median = 0, mode = 0, range = 0-15). Generally, the number of perceived identities for drawings ($M = 5.35$, $SD = 2.57$) were slightly more than the photo card sorting task ($M = 4.78$, $SD = 2.57$), which is understood that task difficulty increases as drawings contain much less texture detail than photos. As with the analysis of photographic images, I next carried out a matrix analysis that makes use of all the available data.

Matrix analysis Following Neil et al., (2016), a confusion matrix was created to visualise and integrate task performance across participants (see Figure 2.4).

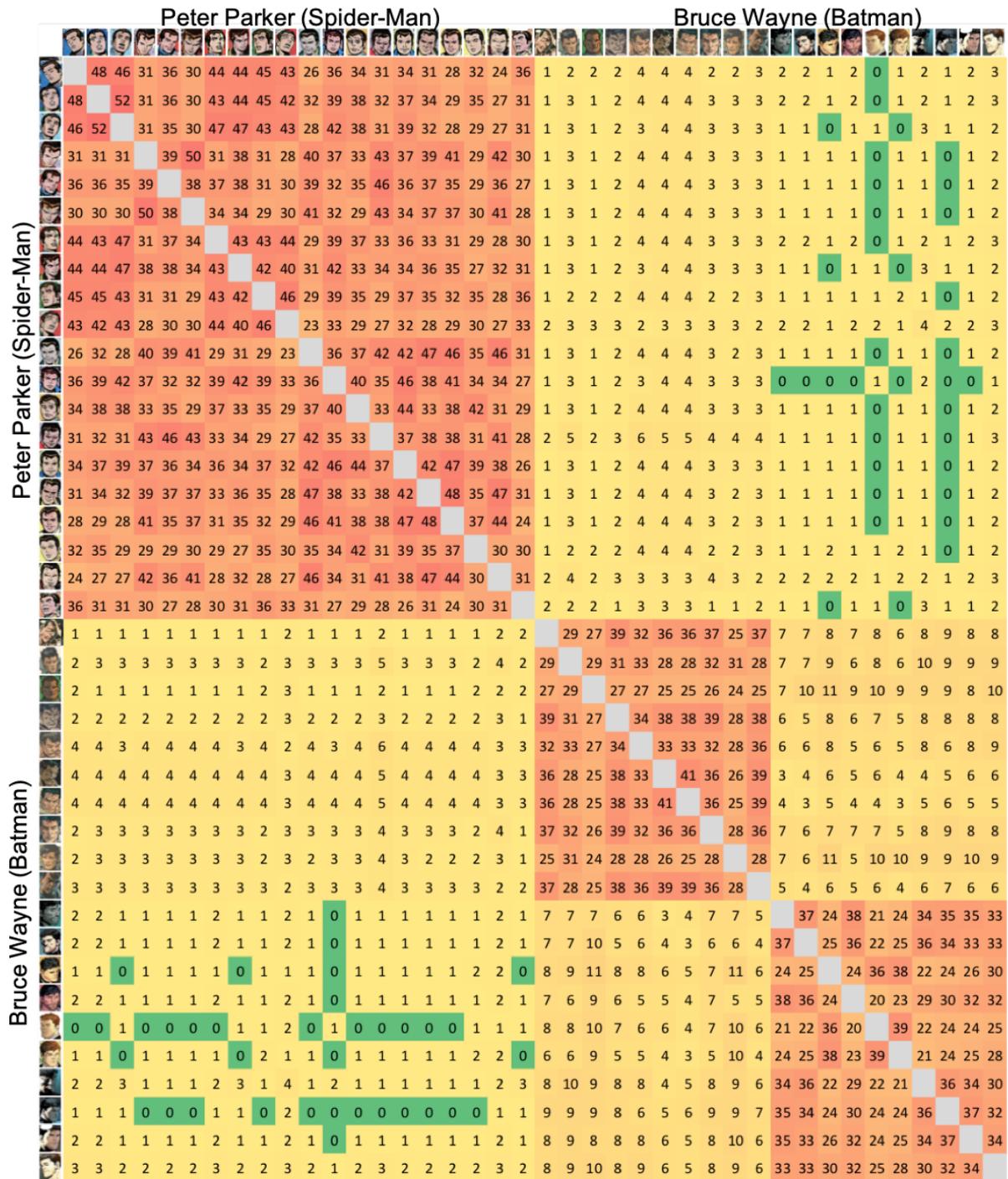


Figure 2.4. Response matrix for 60 participants in the drawing card-sorting task. The x - and y - axes indicate the 20 cards of each of 2 identities (Peter Parker/Bruce Wayne). Each cell in the matrix represents the number of times two different drawings were sorted into the same pile by the 60 participants. Therefore, values reflect the number of participants placing the two images together in the same pile. Perfect performance would result in scores of 60 in the match quadrants (top-left and bottom-right), and scores of 0 in the mismatch quadrants (top-

right and bottom-left). Non-zero values in the latter two quadrants specify misidentification errors, that is, drawings of different identities being grouped together. Cell values are highlighted in a green (low scores) to yellow and then red (high scores) colour gradient. Because an image cannot be matched with itself, blank cells run in a diagonal line from the top-left to bottom-right corners of the matrices.

This matrix is arranged as with the matrix of photographic images. It shows a distinction among the four major quadrants. The match quadrants (top-left and bottom-right) contain moderate-to-high values, whereas the mismatch quadrants (top-right and bottom-left) contain very low values. Some structure is also visible within each quadrant. For example, the match quadrants show some separation between different the same character.

To assess the effect of identity in this sorting task, a paired sample *t*-test was conducted to compare the average score for match cells (the number of times each card appeared in a pile with the same perceived identity, $M = 29.00$, $SD = 6.86$) and the average score for the mismatch cells (the number of times each card appeared in a pile with different perceived identities, $M = 2.47$, $SD = .92$). Scores were significantly higher in match cells than in mismatch cells [$t(39) = 24.33$, $p < .001$], confirming that participants could reliably group these fictional characters, across changes in artist. As in the standard task with real identities (Jenkins et al., 2011), merge errors were rare.

Having summarised overall performance, I next set out to examine effects of familiarity in this task.

Familiarity effect in drawings of fictional characters

As with the analysis of photographic images, comparison of high- and low- familiarity subgroups were made for the drawings.

Simple analysis To test for an effect of familiarity, the number of perceived identities and number of misidentification errors for high-familiarity participants and low-familiarity participants were compared (see descriptive statistics in Table 2.5).

Table 2.5. Descriptive statistics by familiarity on drawing card-sorting performance measures.

	Low-Familiarity Participants	High-Familiarity Participants
Number of perceived identities ^a (<i>max</i> = 40)		
Mean (<i>SD</i>)	6.27 (2.60)	5.20 (2.70)
Median	6	4
Mode	8	2
Range	3-13	2-9
Number of misidentification errors ^b		
Mean (<i>SD</i>)	.73 (1.58)	1.13 (2.61)
Median	0	0
Mode	0	0
Range	0-6	0-10

^a Number of perceived identities is the number of image piles (each representing a different perceived identity) that participants sorted the 40 photographs into.

^b Misidentification errors is the number of times participants' image piles featured more than one fictional identity (i.e., both Peter Parker and Bruce Wayne).

Independent sample *t*-tests showed that although the high-familiarity participants created fewer piles ($M = 5.20, SD = 2.70$) than the low-familiarity participants ($M = 6.27, SD = 2.60$), this difference was not significant [$t(28) = 1.10, p = .281$]. There was also no significant difference in the number of misidentification errors between high-familiarity participants ($M = 1.13, SD = 2.61$) and low-familiarity participants ($M = .73, SD = 1.58$) [$t(28) = -.51, p = .616$]. Given the relative insensitivity of these basic measures, I next undertook a matrix analysis to compare the performance of these subgroups more thoroughly.

Matrix analysis

Figure 2.5 shows separate confusion matrices for the high- and low-familiarity subgroups. As with the overall data, both matrices show clear distinctions among the four major quadrants. Specifically, match quadrants (top-left and bottom-right) contain higher scores than the mismatch quadrants (top-right and lower-left) overall. Within match quadrants, there is some separation between different actors portraying the same character, but this is much less pronounced.

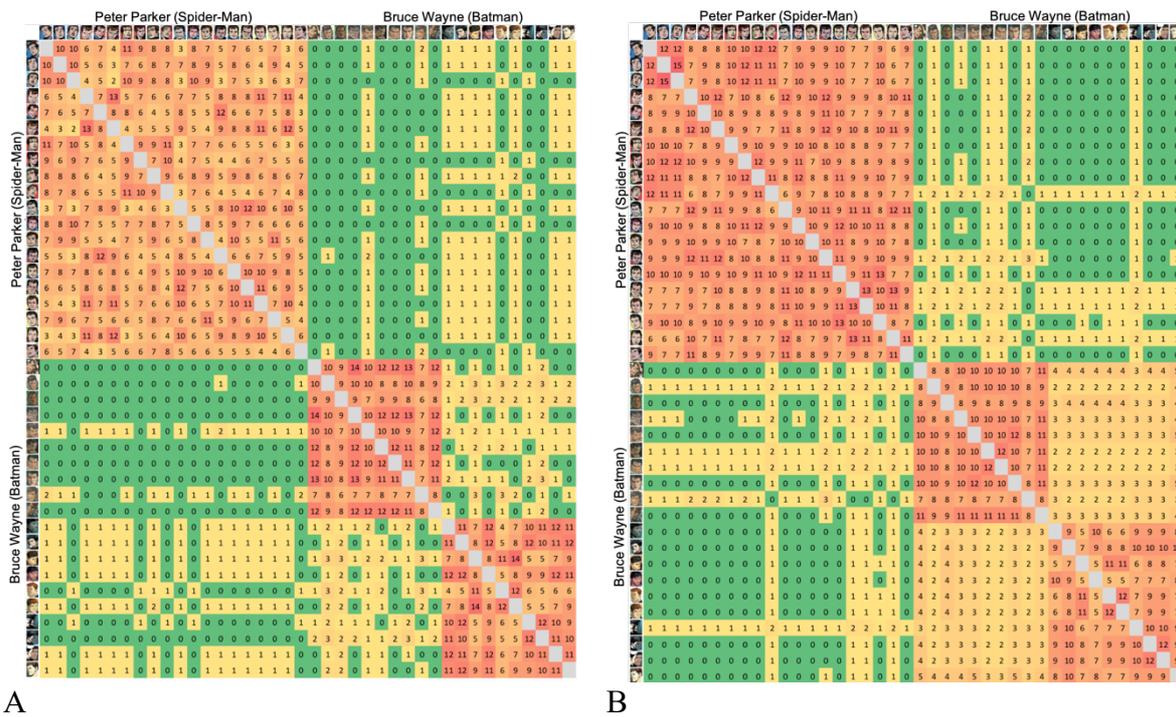


Figure 2.5. Response matrix for (A) low-familiarity and (B) high-familiarity participants in the photo card-sorting task. The x- and y- axes indicate the 20 cards of each of 2 identities (Peter Parker/Bruce Wayne). Each cell in the matrix represents the number of times two different photos were sorted into the same pile across 15 low-familiarity and high-familiarity participants. Therefore, values reflect the number of participants placing the two images together in the same pile Perfect performance would result in scores of 15 in the match quadrants (top-left and bottom-right), and scores of 0 in the mismatch quadrants (top-right and bottom-left). Non-zero values in the latter two quadrants specify misidentification errors, that is, photos of different identities being grouped together. Cell values are highlighted in a

green (low scores) to yellow and then red (high scores) colour gradient. Because an image cannot be matched with itself, blank cells run in a diagonal line from the top-left to bottom-right corners of the matrices.

To compare performance for the two subgroups, a ‘match score’ (averaged values from match cells) and a ‘mismatch score’ (averaged values from mismatch cells) were computed separately for the low-familiarity subgroup and the high-familiarity subgroup (see Table 2.6 for descriptive statistics).

Table 2.6. Descriptive statistics of confusion matrix measures for low- and high-familiarity subgroups.

	Low-Familiarity Participants	High-Familiarity Participants
Match Score ^c		
Mean (SD)	5.82 (.97)	7.43 (1.80)
Median	5.53	7.50
Mode	5.53	5.37, 5.74, 9.32
Range	3.95-7.63	4.79-9.89
Mismatch Score ^d		
Mean (SD)	.38 (.27)	.51 (.40)
Median	.40	.30
Mode	.00	.25
Range	.00-.85	.20-1.12

^c Match score is the averaged values in the cells at the interceptions of a target card and the cards with the same identity as the target card.

^d Mismatch score is the averaged values in the cells at the interceptions of a target card and the cards with a different identity as the target card.

A two-way ANOVA was conducted to examine the effects of identity (match and mismatch) and familiarity (high- and low-familiarity participants) on sorting performance (the number of times two cards has been placed into the same pile).

There was a significant main effect of identity, with higher values for match scores ($M = 6.63, SD = 1.65$) than for mismatch scores ($M = .44, SD = .35$) [$F(1, 39) = 785.50, p < .001, \eta_p^2 = .95$]. This analysis confirms that cards showing the same identity were more frequently placed in the same pile (perceived as the same identity) than cards sharing different identities.

There was also a significant main effect of familiarity, with higher overall scores for high-familiarity participants ($M = 3.97, SD = 3.72$) than for low-familiarity participants ($M = 3.10, SD = 2.83$) [$F(1, 39) = 103.50, p < .001, \eta_p^2 = .73$]. This means that, overall, high-familiarity participants grouped cards together more frequently than low-familiarity participants.

The interaction between these factors was also significant [$F(1, 39) = 63.18, p < .001, \eta_p^2 = .62$]. Simple main effects analyses were conducted to further investigate this interaction. At the levels of familiarity, simple main effects of identity were significant at both high familiarity [$F(1, 78) = 836.07, p < .001, \eta_p^2 = .91$] and low familiarity [$F(1, 78) = 517.63, p < .001, \eta_p^2 = .87$], that match scores were consistently higher than mismatch scores. At the levels of identity, simple main effects of familiarity at mismatch scores were not significant [$F(1, 78) = 1.10, p > .05, \eta_p^2 = .01$], that there were no significant differences between mismatch scores for high or low familiarity. However, simple main effects of familiarity at match score was significant [$F(1, 78) = 162.28, p < .001, \eta_p^2 = .68$], that high-familiarity participants cohered the cards of same identity more frequently than low-familiarity participants.

General Discussion

The present study was designed to investigate the effect of familiarity on face recognition with face photos and face drawings using a card-sorting task (Jenkins et al., 2011). In the original task, Jenkins et al (2011) found that familiar viewers performed almost perfectly

when sorting the cards by identity (i.e., sort the correct photos of two people into two groups), whereas unfamiliar viewers created many different groups, often perceiving images of the same person to be different individuals. Familiarity with a face increases viewers' tolerance level towards image variability, so that more photos are considered as an acceptable representation of the face. In this experiment, I developed variants of the original card sorting task using photos and drawings of well-known fictional characters (Bruce Wayne and Peter Parker).

Familiarity effects in photos of fictional characters

In photo card sorting task of fictional characters, familiar and unfamiliar viewers performed similarly in terms of number of perceived identities and number of misidentification errors. Misidentification errors were rare in both high- and low-familiarity subgroups. The basic familiarity effect reported in Jenkins et al. (2011), whereby unfamiliar viewers create more perceived identities, did not emerge clearly in this version of the task. However, a more sensitive matrix analysis, that compared the precise arrangements of cards, revealed that cards of same identity were more frequently cohered by high- than low-familiarity viewers, whereas, there were no frequency differences in cohering cards of different identities. These findings are consistent with the main findings of Jenkins et al. (2011), but show the effects less strongly. High-familiarity viewers were more able to cohere different images of the same fictional identity than low-familiarity viewers; and for both subgroups, the two different identities were rarely mixed together.

Familiarity effect in drawings of fictional characters

Matrix analysis of the card sorting task for drawings showed a similar familiarity effect to that seen in the photo version of the task. Drawings that shared the same identity were more frequently grouped together by high-familiarity than low-familiarity subgroups. Combining

drawings of different identities was infrequent in both subgroups. These results are especially interesting given that the face drawings contained much less texture information than the photos. Despite the impoverished images, card-sorting performance revealed evidence of face learning, apparently based on exposure to varied images of these fictional faces.

Chapter 3

Social Inference Ratings for Fictional Faces

Introduction

The familiarity effects discussed in the previous chapter raise the question of how identity judgements might be informed by underlying image variability. Burton, Kramer, Ritchie, and Jenkins (2016) investigated this question using Principal Component Analysis (PCA). PCA is a technique used to derive a space of facial images based on the eigenvectors of a PCA-decomposition of a set of faces. It is usually applied to images of many different individuals to extract dimensions along which different faces vary. Burton et al. (2016) performed PCA on images of a single person to understand the variability of that person's face, and hence, to characterize possible images of that person. They found that faces vary in systematic ways, and that the dimensions of variability vary across different faces. Individual faces have their own idiosyncratic variability.

Social categories can be captured from face images as well. In an early example of this, Burton, Bruce, and Dench (1993) used Discriminant Function Analysis (DFA) to establish a metric that can distinguish between female and male faces. Discriminators derived from simple distance measurements in full face, as well as 3D distances of 91 male and 88 female photographs, could generate a single discriminant function that distinguished between sexes with about 95% accuracy, and was strongly correlated with human performance. Linear Discriminant Analysis (LDA) is similar technique that can appearance of faces by establishing the features that best describe a class of stimuli and discriminating it from members of another class (Etemad & Chellappa, 1997; Martinez & Zhu, 2005). In contrast to PCA, LDA is frequently used on different images of the same person which vary only rather slightly. Kramer, Young, Day and Burton (2017c) used everyday ambient face images of few people to train a LDA model on identity. With this identity-trained LDA model, sex and race were classified, even without explicit coding of these two dimensions at learning.

All of these methods of analysis categorize faces based on a set of numbers that quantify the physical appearance of each face, such as metric distances between features ('anthropometry', Kleinberg, Vanezis, & Burton, 2007). However, there are other ways to quantify facial appearance besides anthropometry. One more holistic possibility is social inference ratings, that is, first impressions of character traits such as trustworthiness or dominance. People draw such trait inferences from others' facial appearances in as little as 100-ms exposure (Willis & Todorov, 2006). We automatically draw these inferences from others' faces at the first sight. Given that these rapid social inferences can be drawn from context-free face images, they are presumably based on visual information. In that sense, social inference ratings summarize the appearance of the face. In the previous chapter, I showed that viewers can distinguish fictional identities based on representations that are inherently visual (drawings, or photographed actors). In this chapter, I would like to explore whether highly abstracted summaries of facial appearance, in the form of social inference ratings, are sufficient to distinguish fictional faces. This aim seems ambitious for at least two reasons. First, the faces to be distinguished are quite nebulous, as they only exist in the imagination. This means that the visual categories may be unusually 'fuzzy'. Second, social inference ratings provide extremely sparse information compared with physical images – typically two or three numbers per image (one number for each social dimension). This means that potentially important information may be lost in the encoding.

The general approach involves three phases. The purpose of the first phase is to encode each image as a set of social inference ratings. To achieve this, a group of independent raters will rate the face images on standard social dimensions (*trustworthiness*, *dominance*, *attractiveness*, and *age*). The purpose of the second phase is to establish baseline categorisation performance for humans viewing images. To establish this baseline, a new group of participants will sort the images into exactly two categories (note that this differs

from the previous sorting task, in which the number of categories was not prescribed). The purpose of the third phase is to attempt the same identity categorisation using solely the social inference ratings (that is, the raters' numbers only, without visual images). This will be achieved by submitting the ratings to linear discrimination analysis (LDA). Performance of the LDA will be evaluated by comparison to the baseline performance of human viewers.

Experiment 2

The first step is to examine if there are any consistent differences in social inference ratings (*trustworthiness, dominance, attractiveness, and age*) between the two fictional characters. If there are, then I can proceed to the fundamental question in this chapter - whether instances of the same fictional face could be accurately cohered on the basis of social inference ratings. Social inference ratings on all 80 fictional characters faces instances (40 drawings and 40 photos) were obtained and compared in this initial study.

Methods

Stimuli and Design

All face images (40 drawings and 40 photos) of two fictional characters (40 Bruce Wayne and 40 Peter Parker) used in previous card sorting tasks were rated on *trustworthiness, dominance, attractiveness, and age*. Each of the image was presented with 98mm x 130mm (width x height) in the centre of a 21.5-inch screen. Stimulus presentation and recording of responses were controlled by an Apple Macintosh computer running PsychoPy2 (Peirce, 2007).

Raters

A total of 20 students [12 female, 8 male; mean age (*SD*): 22.80 (2.04); age range: 18 – 25] were recruited from the University of York. All raters were naïve to the stimuli. They gave

written informed consent prior to the experiment and took part in the study in exchange for a small payment or course credit.

Procedure

Raters were instructed to rate *trustworthiness*, *dominance*, *attractiveness*, and *age* of presented faces on a 7-point scale. There were 80 trials in total. In each trial, one face image was displayed on the centre of a 21.5-inch computer screen, with a rating scale from 1 to 7 underneath the image. After observing the image, participants rated one of the four social inferences (*trustworthiness*, *dominance*, *attractiveness*, and *age*) each time by clicking on the rating scale. To ensure all raters had sufficient time for observation, there was no time limitation for task. The task would only proceed when raters gave a response on the current rating.

Results

Social inference ratings (*trustworthiness*, *dominance*, *attractiveness*, and *age*) on the 80 images by 20 participants were averaged across *character* (Peter Parker and Bruce Wayne) and *format* (photo and drawing). Table 3.1 summarises the descriptive statistics.

Comparisons on the four social inference ratings between characters and formats were made to examine whether the two *characters* and *formats* were rated differently, and whether the presentation (format) of the characters influenced ratings.

Table 3.1. Descriptive statistics by character and format on social inference ratings.

Character	Format	Social Inference ratings (<i>max = 7</i>)			
		Trustworthiness	Dominance	Attractiveness	Age

Peter Parker	Photo	4.31 (.48)	3.36 (.69)	4.18 (.98)	3.23 (.18)
	Drawing	3.27 (.45)	3.54 (.78)	2.73 (.40)	3.85 (.29)
	Total	3.79 (.70)	3.86 (.84)	3.45 (1.04)	3.54 (.39)
Bruce Wayne	Photo	3.50 (.59)	4.00 (.68)	3.29 (1.27)	4.62 (.35)
	Drawing	3.36 (.54)	4.53 (.73)	3.14 (.86)	4.14 (.38)
	Total	3.43 (.56)	4.26 (.75)	3.22 (1.07)	4.38 (.44)
Total	Photo	3.90 (.67)	3.68 (.75)	3.74 (1.21)	3.92 (.75)
	Drawing	3.31 (.49)	4.04 (.90)	2.93 (.69)	3.99 (.36)
	Total	3.61 (.66)	3.86 (.84)	3.33 (1.06)	3.96 (.59)

Trustworthiness

A two-way ANOVA was conducted to examine the effects of *character* (Peter Parker and Bruce Wayne) and *format* (photo and drawing) on trustworthiness rating. In this case, there was a significant main effect of *character* on trustworthiness [$F(1, 76) = 9.77, p < .005, \eta_p^2 = .11$], with lower trustworthiness ratings for Bruce Wayne ($M = 4.43, SD = .56$) than for Peter Parker ($M = 3.79, SD = .70$). Overall, Bruce Wayne was seen as less trustworthy than Peter Parker. There was also a significant main effect of *format* [$F(1, 76) = 26.17, p < .001, \eta_p^2 = .26$], with higher trustworthiness rating for photos ($M = 3.90, SD = .67$) than for drawings ($M = 3.31, SD = .49$). Overall, characters presented in photos were seen as more trustworthy than the same characters presented in drawings.

The interaction between these factors was also significant [$F(1, 76) = 15.24, p < .001, \eta_p^2 = .17$]. Simple main effects analyses were conducted to further investigate this interaction.

The simple main effect of character was not significant in drawings [$F(1, 38) = .44, p > .05, \eta_p^2 = .01$], but the simple main effect of character was significant in photos [$F(1, 38) = 35.78, p < .001, \eta_p^2 = .48$]. Specifically, Peter Parker was rated significantly more trustworthy than Bruce Wayne in photos. The simple main effects of *format* was not significant for Bruce Wayne [$F(1, 38) = .78, p > .05, \eta_p^2 = .02$], but it was significant for Peter Parker [$F(1, 38) =$

43.01, $p < .001$, $\eta_p^2 = .53$]. Specifically, photos of Peter Parker were seen as significantly more trustworthy than drawings. One way of summarising this pattern is that Spiderman movies cast actors who look quite trustworthy as Peter Parker, even though Spiderman artists do not draw Peter Parker as a character who looks especially trustworthy. Portrayals of Bruce Wayne are intermediate, both on the screen and on the page.

Dominance

A similar two-way ANOVA was conducted to examine the effects of *character* (Peter Parker and Bruce Wayne) and *format* (photo and drawing) on dominance ratings. There was a significant main effect of *character* on dominance [$F(1, 76) = 25.41$, $p < .001$, $\eta_p^2 = .25$], with higher dominance ratings for Bruce Wayne ($M = 4.26$, $SD = .75$) than for Peter Parker ($M = 3.45$, $SD = .73$). This means that, overall, Bruce Wayne was considered more dominant than Peter Parker. These ratings accord with Bruce Wayne being a masculine entrepreneur and Peter Parker being a young college student. There was also a significant main effect of *format* [$F(1, 76) = 4.89$, $p < .05$, $\eta_p^2 = .06$], with lower dominance rating for photos ($M = 3.68$, $SD = .75$) than for drawings ($M = 4.04$, $SD = .90$). Overall, characters presented in photos were considered less dominant than the same characters presented in drawings, possibly reflecting the artists' propensity to exaggerate facial markers of dominance (e.g. strong brow and jawline).

There was no significant interaction between these factors [$F(1, 76) = 1.24$, $p = .269$, $\eta_p^2 = .02$].

Attractiveness

Another two-way ANOVA was conducted to examine the effects of *character* (Peter Parker and Bruce Wayne) and *format* (photo and drawing) on attractiveness ratings.

There was no significant main effect of *character* on attractiveness [$F(1, 76) = 1.29, p = .261, \eta^2 = .02$], meaning that that Peter Parker ($M = 3.45, SD = 1.04$) and Bruce Wayne ($M = 3.22, SD = 1.07$) were not rated significantly differently on this dimension. However, there was a significant main effect of *format* on attractiveness [$F(1, 76) = 14.82, p < .001, \eta^2 = .16$], with higher attractiveness rating for photo ($M = 3.73, SD = 1.21$) than for drawings ($M = 2.93, SD = .69$). That is, photos were considered more attractive than drawings overall.

The interaction between these factors was also significant [$F(1, 76) = 9.89, p < .005, \eta^2 = .11$]. Simple main effects analyses were conducted to further investigate this interaction. At the levels of format, simple main effects of *character* at drawing were significant [$F(1, 38) = 4.26, p < .01, \eta^2 = .10$] Drawings of Bruce Wayne were rated as significantly more attractive than drawings of Peter Parker. Simple main effects of *character* at photo were also significant [$F(1, 38) = 19.63, p < .001, \eta^2 = .34$]. Photos of Peter Parker were rated as significantly more attractive than photos of Bruce Wayne at photo. At the levels of character, simple main effects of *format* at Bruce Wayne were not significant [$F(1, 38) = .47, p > .05, \eta^2 = .01$]. However, simple main effects of *format* were significant at the character Peter Parker [$F(1, 38) = 40.84, p < .001, \eta^2 = .52$], with photos of Peter Parker rated significantly more attractive than drawings of Peter Parker. As with the Trustworthiness ratings,

Spiderman movies cast unusually attractive actors as Peter Parker, even though these Spiderman artists did not draw Peter Parker as an attractive character. Once again, portrayals of Bruce Wayne appear intermediate, both on the screen and on the page.

Age

As with the preceding social dimensions, two-way ANOVA was conducted to examine the effects of *character* (Peter Parker and Bruce Wayne) and *format* (photo and drawing) on age rating. As expected, there were significant main effects of *character* on age [$F(1, 76) =$

145.00, $p < .001$, $\eta_p^2 = .67$], with higher age ratings for Bruce Wayne ($M = 4.38$, $SD = .44$) than for Peter Parker ($M = 3.54$, $SD = .39$). This means that, overall, depictions of Bruce Wayne were seen as older than depictions of Peter Parker. This is consistent with the characters as written. There was no significant main effect of *format* [$F(1, 76) = .94$, $p = .335$, $\eta_p^2 = .01$], meaning that age ratings were not significantly different for drawings ($M = 3.99$, $SD = .36$) and photos ($M = 3.92$, $SD = .75$).

The interaction between these factors was significant [$F(1, 76) = 61.97$, $p < .001$, $\eta_p^2 = .45$]. Simple main effects analyses were conducted to further investigate this interaction. The simple main effect of *character* was significant in drawings [$F(1, 38) = 10.12$, $p < .005$, $\eta_p^2 = .21$] and in photos [$F(1, 38) = 230.73$, $p < .001$, $\eta_p^2 = .86$], with Bruce Wayne being rated as older than Peter Parker in both formats. The simple main effect of *format* was significant for Bruce Wayne [$F(1, 38) = 37.36$, $p < .001$, $\eta_p^2 = .50$], who was rated as significantly older in photos than in drawings. It was also significant for Peter Parker [$F(1, 38) = 61.32$, $p < .001$, $\eta_p^2 = .62$], but in the opposite direction. Peter Parker was rated as significantly younger in photos than in drawings. Apparently, movie casting decisions exaggerated the difference in character age in the drawings.

Summary

Social inference ratings (*trustworthiness*, *dominance*, *attractiveness*, and *age*) differed between *characters* as well as between image *formats*. Bruce Wayne was rated as more dominant, older, and less trustworthy than Peter Parker. Photos were generally rated more attractive, more trustworthy, and less dominant than drawings. Thus, there are meaningful regularities in the social inference ratings, and these correspond to the fictional characters as written—Bruce Wayne being a relatively senior and imposing entrepreneur; Peter Parker

being a callow college student. Having established these patterns, I would like to proceed to the main question in this chapter – whether different instances of the same fictional character can be accurately cohered on the basis of social inference ratings alone. This question can be addressed by submitting the ratings to linear discriminators analysis.

Experiment 3

Linear Discriminant Analysis (LDA) is a technique to establish the features which best describe a class of stimuli and discriminating it from members of another class (Etemad & Chellappa, 1997; Martinez & Zhu, 2005). This technique is frequently used on face classification in engineering, usually with different images of the same person which vary only rather slightly. Kramer et al. (2017c) used everyday ambient face images of few people to train a LDA model on identity. With this identity-trained LDA model, sex and race were classified, even without explicit coding of these two dimensions at learning. Unlike face shape, eye distance, and other visual information, sex and race are more about the top-down processes that cohere social categories. LDA is particularly useful in modelling the top-down processes that cohere superficially different images of the same person (Kramer, Young, & Burton, 2018), as each face has its own idiosyncratic variability – that is, the dimensions of variability in one face do not generalize well to another (Burton et al., 2016). The selected fictional characters differ significantly in social inference ratings (*trustworthiness*, *dominance*, *attractiveness*, and *age*). The question is whether these differences can be used to discriminate identities accurately.

To evaluate the level of accuracy achieved by LDA, it will be helpful to have a baseline for comparison. Human performance based on viewing the images is a meaningful comparison, because it sets a reasonably high standard. The images contain all of the available information, and the human visual system is a sophisticated processor of such images. Many

previous experiments have tested human participants in card sorting tasks for facial identity. However, very few of these are directly analogous to the present situation. This is because the standard test is unconstrained, in the sense that participants can create as many piles of cards they like. Indeed, the number of piles is a key dependent variable in the standard test. The situation here is rather different. LDA will be constrained to create two categories in its solution (the two facial identities). The measure of accuracy is how well these two categories separate the two identities. The analogous task for humans is a constrained card-sorting task in which participants are required to sort the cards into two piles. Performance on a constrained card-sorting task has been previously reported. Andrews et al. (2015) directly compared free sorting, in which participants were free to create as many piles as they like, to constrained sorting, in which they must create two piles only. Although participants were forced to make fewer piles in the constrained task, the number of misidentification errors was similar across the two conditions. Thus, participants had little difficulty distinguishing between the faces when they were given additional information – the number of identities that were present. The purpose of the current experiment is to establish human performance in a constrained sorting task based on the Bruce Wayne / Peter Parker image set. This will provide a baseline against which to compare LDA performance in the next study.

Methods

Participants

A total of 30 students [15 female, 15 male; mean age (*SD*): 22.85 (3.86); age range: 18 – 26], were recruited from the University of York. Unlike the previous experiment, recruitment did not specifically target volunteers across the familiarity continuum. Instead, it was a convenience sample drawn from the undergraduate population. All participants gave written informed consent prior to the experiment and took part in the study in exchange for a small payment or course credit.

Stimuli and Design

Participants were presented with the all pre-rated faces images (40 “Bruce Wayne” and 40 “Peter Parker”). All these images were printed in their original colour on laminated cards measuring 20 × 29 mm each. For each set of 40 cards, 20 drawings and 20 photographs were presented.

Procedure

Participants were given a shuffled deck of all 80 face cards. They were instructed to sort the cards into piles into two fictional identities regardless of image format. In other words, there would be two piles, each pile could contain both photographs and drawings of the same perceived identity. There was no time limitation on this task.

Results

All participants were required to sort the cards into two piles (i.e. two identities), so the number of piles was not a dependent variable in this analysis. Instead, analysis focused on the number of misidentification errors, and accuracy rate was calculated as the percentage of correct identifications. The accuracy of this two-sort task with fictional character faces was 88.92% (*S.D.*: .11). Six participants sorted the cards perfectly, and 13 achieved an accuracy of over 90%. No participants made piles consisting of 50% of each identity.

Discussion

Human participants rarely confused the two identities in the two-sort task of fictional characters. Overall accuracy was 89%, despite the unusually high diversity of the images (i.e. different image formats). In the Andrews et al. (2015) two-sort task, based on photos of unfamiliar faces, 10 out of 12 participants (83% of participants) sorted the two identities perfectly, and the rest made more than two intrusion errors per identity. Here, only 6 out of 30 participants (20% of participants) sorted the two identities perfectly, consistent with

higher overall difficulty. However, no participants made piles consisting of 50% of each identity, which would be expected by chance.

Experiment 4

Now we can proceed to the Linear Discriminant Analysis (LDA), using the same image set and the same measure of performance. Human performance on this task is highly complex, as it can be determined by a range of factors. Physical similarity of the images presumably plays a role here, as comparison of face shape, skin tone, eye colour, and other aspects of appearance, can inform perceived identity directly. However, they can also inform perceived identity indirectly, by creating a social impression. The LDA presented here focuses solely on social impressions, as it only has access to participants' mean social inference ratings (*trustworthiness*, *dominance*, *attractiveness*, and *age*), and not to the images themselves. In this way, each image is represented by just four numbers. The key question is how accurately the images can be grouped based on these social inference ratings alone.

Methods

Social inference ratings

Social inference ratings (*trustworthiness*, *dominance*, *attractiveness*, and *age*) on the 80 images by 20 participants were averaged separately to obtain 4 ratings for each image.

Procedure

The four social inference ratings (*trustworthiness*, *dominance*, *attractiveness*, and *age*) were analysed with LDA with indicated current classification – *identity* (Bruce Wayne and Peter Parker) and *format* (drawings and photos) respectively. The analysis was carried out using SPSS.

Results

Grouping by identity LDA was used to conduct a multivariate analysis of variance test of the hypothesis that the fictional characters (Peter Parker and Bruce Wayne) would differ significantly on a linear combination of four variables: *trustworthiness*, *dominance*, *attractiveness*, and *age* ratings. The overall Chi-square test was significant [Wilks' $\lambda = .38$, Chi-square = 73.08, df = 4, $p < .001$], confirming the hypothesis. Using the four social inference ratings, 87.5% of the images were correctly classified by the character they represent. Summary statistics are shown in Table 3.2. *Trustworthiness*, *dominance*, and *age* were different for the two groups, so that these variables contributed to the grouping as factors. *Attractiveness* showed no significant difference, and probably do not contribute to the grouping.

Grouping by image format A second LDA was conducted to examine whether the two image formats (drawings and photos) differed significantly on a linear combination of the four variables. The overall Chi-square test was significant [Wilks' $\lambda = .63$, Chi-square = 35.09, df = 4, $p < .001$], confirming the hypothesis. Using the four social inference ratings, 76.3% of the images were correctly grouped by image format. Summary statistics are shown in Table 3.2. *Trustworthiness* and *attractiveness* were different for the two groups, so that these variables contributed to the grouping as factors. In contrast, *Dominance* and *age* showed no significant difference, and probably do not contribute to the grouping.

Table 3.2. Results of tests of equality of group mean, displays the results of one-way ANOVAs for the independent variable using the social inferences as the factors.

Classification	Social Inference							
	Trustworthiness		Dominance		Attractiveness		Age	
	Wilks' λ^a	F (1, 78)						

Character	.92	6.49*	.76	24.13**	.99	1.00	.49	81.42**
Format	.79	20.21**	.96	3.72	.85	13.34**	1.00	.26

^a Wilks' λ is a multivariate test statistic whose value ranges between 0 and 1. Values close to 0 mean that the class means are different and values close to 1 mean that the class means are not different (equal to 1 means all means are the same).

* $p < .05$

** $p < .001$

Summary

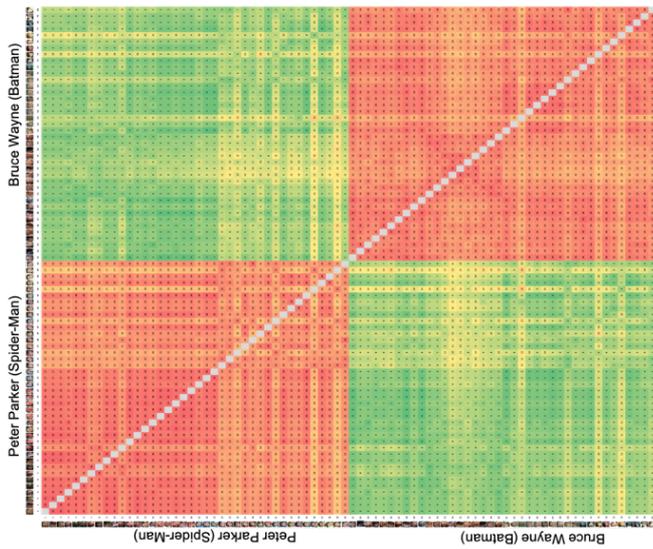
In sum, using LDA with social inference ratings (*trustworthiness, dominance, attractiveness, and age*) as factors, 87.5% of the images were correctly grouped by the fictional character of cards and 76.3% were correctly grouped by the format of cards. Two aspects of these findings are especially interesting. The first is that accuracy for identity is so high. Even though the LDA had access to social inference ratings only, it performed as well as human sorters who viewed the images directly. The second is that accuracy for identity was higher than accuracy for image format. This pattern suggests not only that the two fictional characters create distinct social impressions, but also that these distinct social impressions are conserved across image formats.

Having established overall levels of accuracy for human participants and LDA, the next step is to compare the patterns of performance that give rise to the observed levels of accuracy.

Comparison of Linear Discriminant Analysis (LDA) and Forced Card-sorting

The LDA described above showed how the 80 cards of fictional characters can be grouped into two groups based on the social inference ratings (*trustworthiness, dominance, attractiveness, and age*). Given the similar overall levels of identification accuracy for LDA

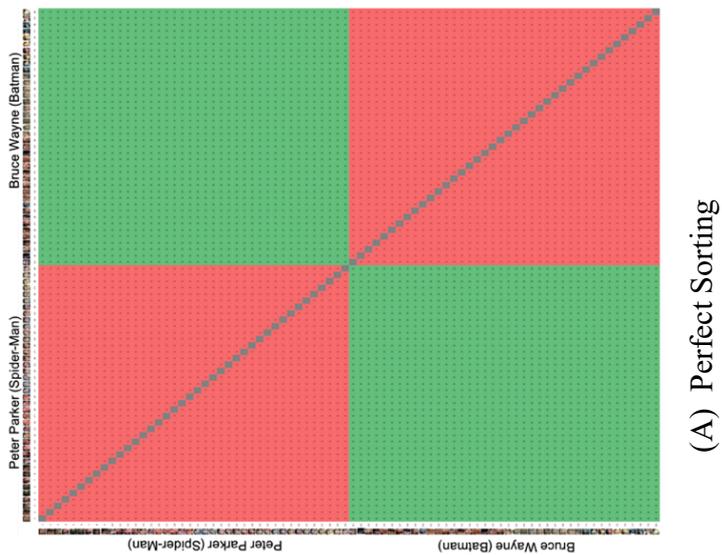
and human viewers, I now ask whether the similar level of accuracy was achieved by similar means, specifically, whether the underlying patterns of performance also similar. To this end, I constructed three confusion matrices to visualize (a) perfect sorting, (b) the suggested grouping by Linear Discriminant Analysis (LDA), and (c) the forced card-sorting task performance across 30 human participants. The three matrices are shown in Figure 3.1.



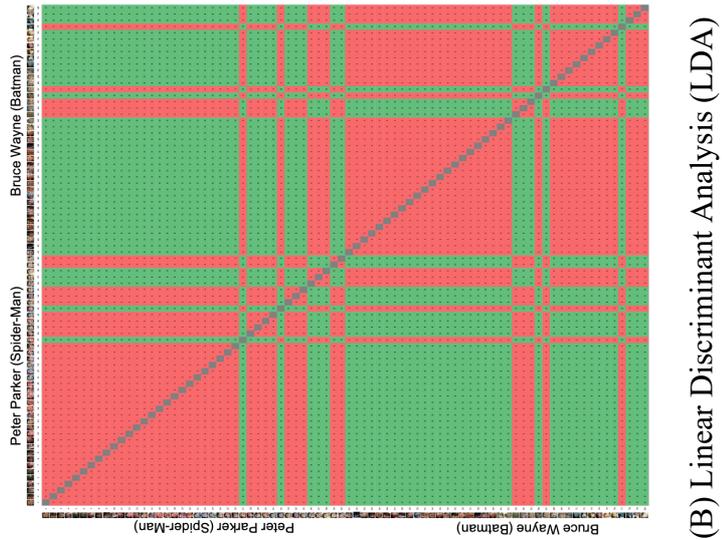
(C) 30 Human Participants

Figure 3.1. Response matrix for (A)

perfect sorting, (B) Linear Discriminant Analysis (LDA) and (C) human participants in the forced card-sorting task. The x- and y- axes indicate the 40 cards of each of 2 identities (Peter Parker/Bruce Wayne). Each cell in the matrix represents the number of times two different images were sorted into the



(A) Perfect Sorting



(B) Linear Discriminant Analysis (LDA)

To assess the similarity of these matrices, I compared perfect performance, LDA suggested grouping, and human performance using Pearson correlation. The results of these correlation analyses are shown in Table 3.3. As the table shows, human performance was strongly correlated with perfect performance ($r = .96$). The LDA solution was also significantly correlated with perfect performance ($r = .55$), even though the LDA had access to only the numerical ratings of images, with no direct visual information. More importantly for this analysis, there was also a significant positive correlation between LDA performance and human performance ($r = .55$). This implies some similarity in the specific patterns of identity grouping, over and above overall accuracy level. Some of this structure can be seen in Figure

3.1. The LDA matrix and human matrix show similar patterns of striping, which correspond to similar patterns of error. To this extent, social inference ratings seem to capture the visual appearance of the two fictional characters and the variability within each identity.

Table 3.3. Correlation matrix of matrices performance (Pearson correlation coefficient, r) of 6320 matrix cells.

	Mean	SD			
1. Perfect performance	.49	.50	1.00		
2. LDA suggested grouping	.49	.50	.55**	1.00	
3. Human performance	14.94	10.19	.96**	.54**	1.00

** $p < .001$

To further compare the LDA and human solutions, I first computed a ‘match rate’ (averaged ‘per-participant’ values from match cells) and a ‘mismatch rate’ (averaged ‘per-participant’ values from mismatch cells) separately for matrices of perfect performance, LDA suggested grouping (by character), human performance (see Figure 3.2), as well as a LDA suggested grouping matrix with x- and y- axes indicate the 40 cards of each of 2 formats (drawing/photo). The two LDA matrices contain exactly the same data, but arranged into character-based and format-based matrices, resulting two matrices with different patterns.

For both match and mismatch rate, the LDA grouping is closer to perfect performance when arranged into a character-based matrix than when arranged into a format-based matrix.

Specifically, the character-based LDA matrix has a higher match rate, and a lower mismatch score than the format-based LDA matrix, confirming that the character-based LDA matrix is more accurate. This implies that the social inference ratings supported differentiation of the two characters (Bruce Wayne and Peter Parker) better than they supported differentiation of the two image formats (drawings and photos) in this case.

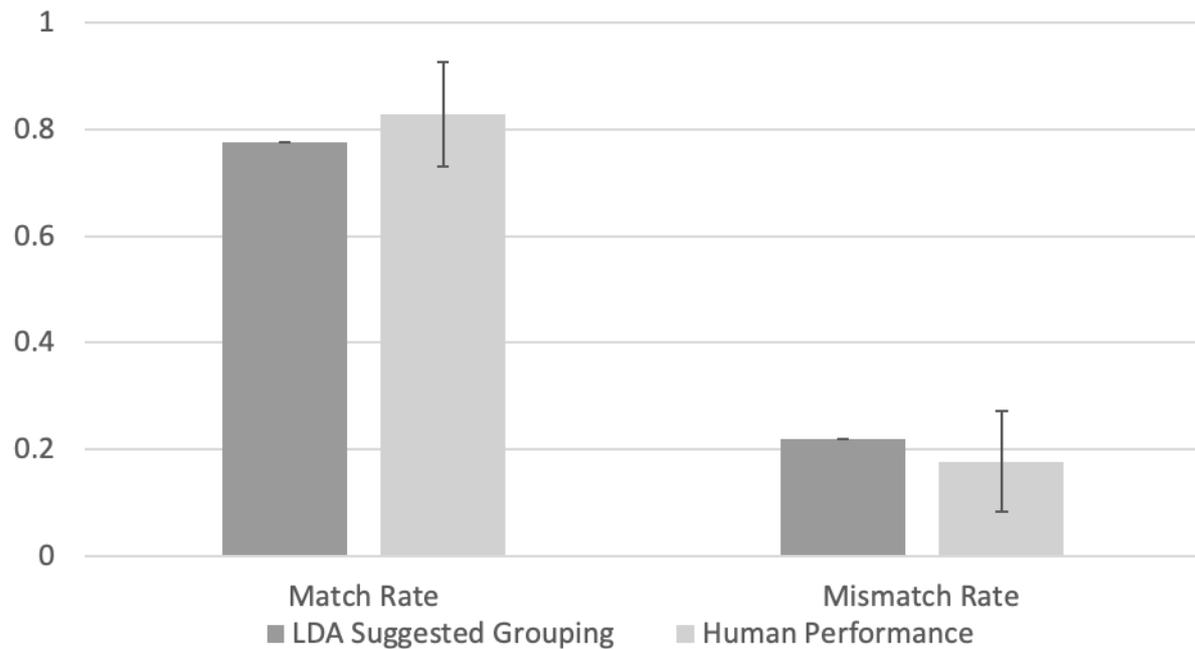


Figure 3.2. Means of match rate (averaged ‘per-participant’ values from match cells) and mismatch rate (averaged ‘per-participant’ values from mismatch cells) for matrices of LDA suggested grouping, and human performance. Vertical bars represent the standard error of the means.

Summary

LDA suggested grouping created with social inference ratings (*trustworthiness, dominance, attractiveness, and age*) achieved similar accuracy to human performance in a constrained card-sorting task for fictional faces. LDA was able to cohere images of the same fictional character across image format, which implies that (i) idiosyncratic aspects of the character’s face (Burton et al., 2016) were consistent across different image formats, and (ii) social ratings of fictional characters’ faces captured these idiosyncratic aspects of appearance.

General Discussion

The present study investigated whether varied images of two fictional faces could be accurately grouped on the basis of social inference ratings. Analysis of social inference ratings (*trustworthiness, dominance, attractiveness, and age*) revealed significant differences

between characters and between image formats. Linear Discriminant Analysis (LDA) on these social inference ratings grouped the images by identity with 87.5% accuracy and grouped the images by format with 76.3% accuracy. This level of accuracy was similar to human accuracy in a constrained card sorting task for identity (89%). Unsurprisingly, human viewers performed perfectly when sorting the images by format (100%). In addition, in a matrix analysis, the LDA grouping based on social inferences achieved similar accuracy to human performance in a constrained card-sorting task for identity, regardless of image format. Together, these findings imply that even sparse ratings about social impressions from face images (four digits) can preserve a surprising amount of information about facial appearance. Human viewers could use not only social impressions inferred from the images, but also background knowledge and comparison of visual details when distinguishing the two characters. However, LDA had access only to the social inference ratings. The relatively high accuracy of the LDA solution shows that these social inference ratings captured idiosyncratic aspects of facial appearance and that these idiosyncracies are consistent across image formats.

One limitation of the current study is the sampling of stimuli. Given the exploratory nature of the research, focusing on two fictional characters seems a sensible starting point. It also allows meaningful comparison with previous card-sorting tasks that were based on two identities. Although these studies illustrate how progress can be made studying fictional faces experimentally, they do not address generalisability to other fictional faces, or the issue of discriminating greater numbers of identities. Future studies could also examine a greater range of within-person variability. Although the current studies go further than most in presenting photos and drawings of the same fictional faces, these portrayals were limited to two actors and two artists for each identity. The full breadth of visual representations extends much further. In future work, it would be interesting to bring compare portrayals from dozens

of artists, and especially, different characters drawn by the same artist. A more comprehensive analysis of this type would provide greater insight into the mappings between facial appearance and identity. The preceding chapter and the current chapter offer a framework for such research.

One of the overarching messages from Chapters 2 & 3 is that impoverished visual representations of faces (drawings), can engage the cognitive apparatus of face perception to a surprising degree. The next chapter develops this theme by examining a different style of drawing in which visual representations of faces are simplified further still.

Chapter 4

Identification from Sparse Drawings

Introduction

The preceding experiments showed that faces on drawings from comic book series could be learned and familiarized by readers. Although the drawings contained impoverished information (relative to photographic images), they still contained numerous cues to shape and texture—and these were often rendered in a moderately realistic style. These selected drawings from superhero comics were only one type of the representations. McCloud (1993) introduced a scale of progression showing two sides of representation, from realistic to abstract (see Figure 4.1). Realistic representation includes photographic-like images, which look like the ‘real life representation’, whereas, abstract representation is a more general representation, which the image has a greater room for interpretation. Realistic representation is more representational and most dedicated to realism, that faces have consistent depictions and little symbolic purpose, therefore the face would be easily identifiable. The abstract representation is more simplistic, symbolic and impressionistic, that faces are more stylized to create an impression of the character, which allows readers easily identify with the faces.

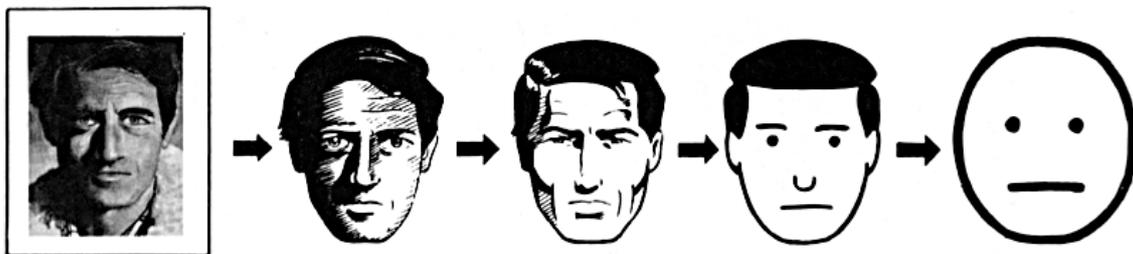


Figure 4.1. The scale of face drawing progression from realistic representation (left) to abstract/iconic representation (right) (McCloud, 1993).

Artists can further reduce the amount of visual information that the drawings provide, by moving further away from visual realism. Face drawings from a comic column, called *Make Me A Menace*, from British children’s comic magazine *The Beano*, are good examples of this

(Beano Studio, 2018). This column publishes a one-page story based on a child's photo and a 'menace name' (usually the name of the child with a descriptive word describing the child's personality or interests, e.g., Fidgety Florence, Olive-Eating Olivia, and Lazy Lila) sent by the child. Multiple scenes with drawings of the child would be drawn based on that photo, with a short story plot fits the descriptive word. Drawings in *Make Me A Menace* are relatively less representational, but more stylized. They are not capturing the appearance in literal sense, but creates impression of the character, that there are plenty of expressive but non-realistic expressions (e.g., rosy-cheeks, dot eyes, or super wide mouth). As opposed to faces in superhero comic book series, these faces appear in only one single story, which do not need to be consistently recognizable across different stories. Shape and texture are further reduced in these face drawings. These images do not lend themselves to a card sorting task, because the number of images for each identity is severely limited - one photographic image and typically three to five useable drawings. However, they do lend themselves to a perceptual matching task for identity.

Bruce et al. (1999) pioneered an experimental method using an array of face images to study how did changes in viewpoint and expression between the target and distractor faces affect the accuracy of matching unfamiliar faces. A single still target image captured from a video was presented along with an array of 10 high-quality photographs of men with similar physical appearance to the target. Observers were required to decide if the target was present in the array, and if yes, which one was the target. The average accuracy (i.e., correct hit or correct rejection) for the task was 70% in the matched viewpoint and expression condition for both target-present and target-absent arrays. The task was further simplified with target-present arrays that viewers were forced to choose one answer among the 10 photos in the array, performance improved slightly. Performance of this 1-in-10 line-up task affected not purely by the appearance of the target, that participants make decisions by simply confirming

the target, foils can provide hints to rule out the images must not be the target. Henderson et al. (2001) further simplified into a single-item verification task that observers have to decide whether the two photos showed the same person. Accuracy was still low in this task with only two images. However, the result revealed the performance driven purely by the target by eliminating the possible effects of the presence of foils. These perceptual matching tasks for identity could also be applicable to non-photographic faces. Bruce, Ness, Hancock, Newman and Rarity (2002) conducted perceptual matching tasks with face composites constructed by viewers using Pro-Fit: presenting a line-up of six faces of similar hairstyle and approximate age with either the best single composite, the worst single composite, the morphed image of four composites contributed by the four different participants, or a set of four individual composites. Participants were told the composite(s) were attempts to generate a likeness of one of the six face and asked to indicate the face. The morphed images were considered best in likeness that participants performed the best in line-up tasks, which they provided a better impression as they weighted accurate over inaccurate details, assuming that different witness errors are uncorrelated.

I would like to conduct perceptual face matching tasks with face images from *The Beano* (Beano Studio, 2018): a photo and several drawings, which contain further impoverished information compared to superhero comics used in our previous experiment, or face composites in the studies of Bruce et al. (2002). These images capture further less of the physical appearance but more of the impression of a target photo with help of a name. It would be difficult to identify the particular target with the images on the basis of purely physical appearance, that the 'impression' provided by the drawing would be relatively essential in matching the images to the target photo. These would allow us to advance our understanding in recognizing more abstract representations of faces.

Experiment 5

Bruce et al. (1999) conducted face matching tasks with one target against an array of 10 possible images with neutral expression which simulate police line-up tasks. They examined the effect of target pose (full-face neutral expression, full-face smiling, and neutral expression photo taken with an angle of 30-degree) on matching unfamiliar faces. Half of the task were target-absent, and this was told to the participants. Their participants performed significantly better in full-face with neutral expression compared to other poses, with the highest percentage of hits (70%) and lowest percentage of incorrect (12%), and miss (18%) in a target-present trials, and highest percentage of correct rejection rates (70%) in target-absent trials. The other two poses did not differ significantly from each other. They found that these superficial impressions of resemblance or dissimilarity between face images would be misleading in identification.

Line-up identification task was also conducted with synthesized images. Bruce et al. (2002) conducted perceptual matching tasks with face composites constructed by viewers using Pro-Fit: presenting a line-up of six faces of similar hairstyle and approximate age with either the best single composite, the worst single composite, the morphed image of four composites contributed by the four different participants, or all four individual composites. Participants were told the composite(s) were attempts to generate a likeness of one of the six face and asked to indicate the face. On half of the arrays, no target image was present. Participants generally made incorrect decisions more frequently for all composite types. The number of times of making a correct decision was significantly higher with the morphed image of four composites in both target-present and absent trials. The morphed image weighted accurate over inaccurate details, assuming that different witness errors are uncorrelated. They provided a better impression of the to-be-identified target.

The focus of the current study is rather different, in that the purpose of the materials was to convey a sense of character, rather than to support identification. Nevertheless, the artist is obliged to produce a certain likeness to the photographic subject. Interestingly, in this situation, the artist has only a single photograph to work with, and so the resemblance will presumably rely heavily on that photograph—perhaps supported by inferences drawn from that photograph and the name.

For this reason, we might expect the drawings to emphasise ‘external’ features such as hair colour and hair style. Viewers are known to rely on such external features when matching unfamiliar faces, and so these cues are likely to be salient to the artist when generating the drawings, and salient to the reader when viewing the drawings.

Our hypothesis is that participants would perform better in drawing-to-drawing than photo-to-drawing condition, as image dissimilarity increases in the later condition. Moreover, the performance in drawing-to-drawing condition would resemble the pattern of full-face neutral expression pose in Bruce et al. (1999) as the resemblance or dissimilarity between face images were at a similar level, regardless of the image formats are different for the two experiments.

Methods

Participants

Participants were 60 students from University of York [mean age (SD): 23.52 (3.44); age range: 18 – 28; 30 male, 30 female]. All participants gave written informed consent prior to the experiment and took part in the study in exchange for a small payment or course credit.

Stimuli and Design

The experiment had a 2×2 factor within-subjects counterbalanced design, with 15 participants tested in each condition. The first variable was target format (drawing vs. photograph), the second was array condition (present vs. absent). All face images were chosen from a comic column, called *Make Me A Menace*, from British children's comic magazine *The Beano*. Children sent *The Beano* photos of themselves with a menace name, and a story would be drawn based on these. As there was only one photograph published in each story, photos could only appear as the target in this task. All target faces were tested in all conditions of the experiment.

Prior to our experiment, I constructed for each target face (two drawing and one corresponding photograph) a set of five other face drawings that resembled the targets based on features such as (in order of priority) sex, skin tone, hair colour, to hairstyle. The array was constructed with one drawing of the target face and four high resemblance face drawings. The same array was used in all four conditions, with target image changed in each condition. A target-present trial was constructed with the drawing (which not present in the array) or photo of the target, whereas a target-absent trial consisted of the resembling face drawing other than the four in the array, or its corresponding photo (see *Figure 4.2*).

In each trial, a target image was displayed on the upper centre of a 21.5-inch computer screen, with five numbered images in a horizontal array underneath. Each image measured approximately of 10 cm height on the screen. Stimulus presentation and recording of responses were controlled by an Apple Macintosh computer running PsychoPy2 (Peirce, 2007).

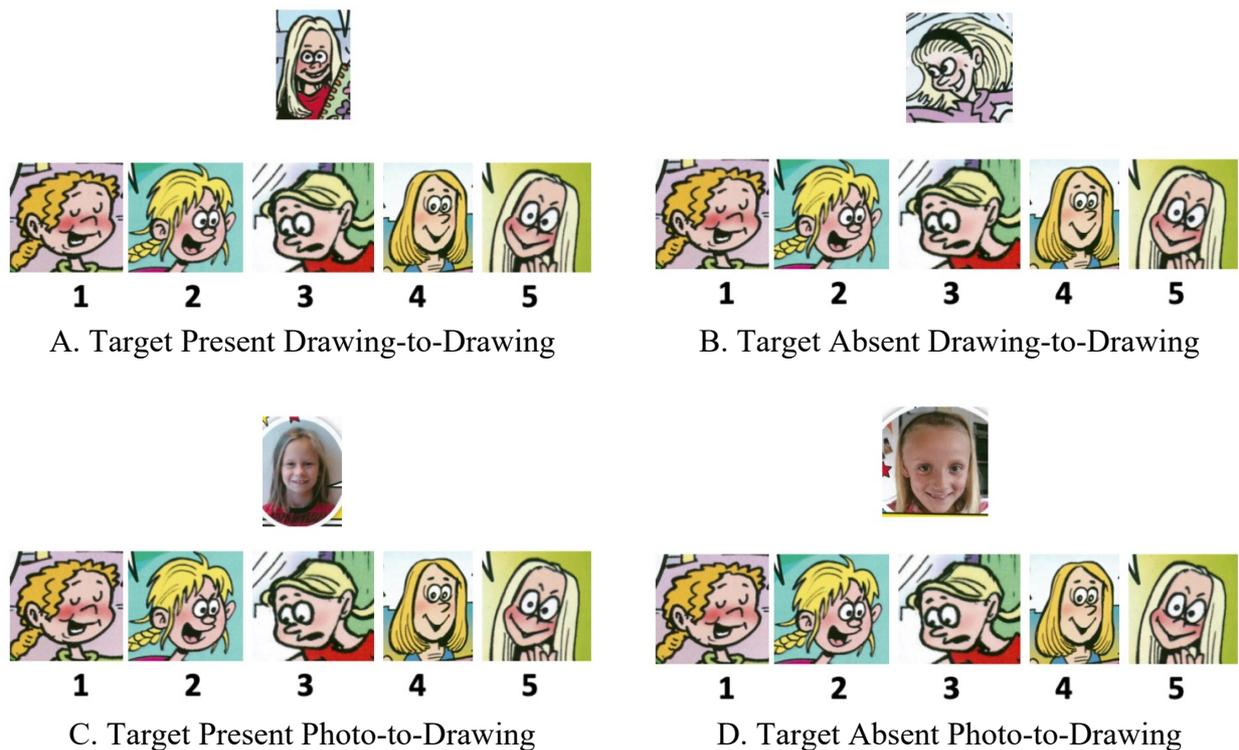


Figure 4.2. Examples of a target with an array used in the experiments. The target image is either a drawing or a photo of the same child, accompany either a target-present or a target-absent array. Drawing 5 is the solution in for target-present array.

Procedure

Participants were instructed that the experiment is similar to a police line-up task. Each participant completed 30 consecutive trials. In each trial, they would see a target face at the top of the display, and a line-up of five faces in an array below. Their task was to inspect each array and indicate the number of the face they thought matched that target, or to state that the target was absent. There was no time limit for the comparison made on each trial. They were asked to guess for trials on which they felt unable to reach a decision.

Results

Table 4.1 shows the mean percentages of each possible type of response for *drawing-to-drawing* and *photo-to-drawing* trials in this task. A one-way ANOVA of the correct hits on

target-present trials revealed a main effect of target format, $F(1, 118) = 103.56, p < .001$, such that there were significantly more hits in the *drawing-to-drawing* trials (72%) compared with the *photo-to-drawing* condition (36%). Analysis of correct rejection in target-absent arrays also showed a main effect of target format, $F(1, 118) = 7.15, p < .01$, such that there were significantly more correct rejections in the *drawing-to-drawing* trials (52%) compared with the *photo-to-drawing* condition (41%).

Table 4.1. Summary of the mean percentages of each possible type of response in drawing-to-drawing and photo-to-drawing trials.

Target Format	Target Present						Target Absent			
	Hits	SD	Incorrect	SD	Miss	SD	Correct Rejection	SD	False Alarm	SD
Drawing-to-Drawing	72	17	14	14	15	11	52	24	48	24
Photo-to-Drawing	36	21	39	21	25	17	41	23	59	23

Analysis of Miss errors (a target was present but the response was Absent) similarly showed a main effect of target format, $F(1, 118) = 14.97, p < .001$, with fewer Miss errors in the *drawing-to-drawing* condition (15%) than in the *photo-to-drawing* condition (25%). There was also a main effect of target format in Another Person errors (choosing a foil instead of the target), $F(1, 118) = 61.05, p < .001$; with fewer errors in *drawing-to-drawing* condition (14%) compared to *photo-to-drawing* condition (39%).

A 2 (target format) \times 2 (target present or absent) ANOVA was conducted on the percentage of correct responses (hits or correct rejection). There was a significant main effect of target format, with significantly more correct responses in the *drawing-to-drawing* condition ($M = .62, SD = .23$) than in the *photo-to-drawing* condition ($M = .38, SD = .22$), $F(1, 59) =$

75.57, $p < .001$, $\eta_p^2 = .56$. There was also a significant main effect of target present or absent [$F(1, 59) = 6.33$, $p < .01$, $\eta_p^2 = .10$], with significantly more correct responses in target-present trials ($M = .54$, $SD = .26$) than target-absent trials ($M = .47$, $SD = .24$).

The interaction between these factors was also significant, $F(1, 59) = 26.61$, $p < .001$, $\eta_p^2 = .31$. Simple main effects analyses were conducted to further investigate this interaction. The effect of *format* was significant for target-present trials [$F(1, 118) = 99.00$, $p < .001$, $\eta_p^2 = .46$], and for target-absent trials [$F(1, 118) = 10.26$, $p < .005$, $\eta_p^2 = .08$], reflecting more hits (target-present) and correct rejections (target-absent) in the *drawing-to-drawing* trials compared with the *photo-to-drawing* condition. The effect of target presence or absence was significant for the *drawing-to-drawing* task [$F(1, 118) = 27.49$, $p < .001$, $\eta_p^2 = .19$], reflecting more hits (target-present) than correct rejections (target-absent). However, this effect was not significant in the *photo-to-drawing* task [$F(1, 118) = 1.93$, $p > .05$, $\eta_p^2 = .02$].

Discussion

In this face matching task, performance was significantly better when the format of target and array images matched (*drawing-to-drawing* condition) than mismatched (*photo-to-drawing* condition). It is easier to do within format comparison.

Participants more frequently chose the correct target when there was a target or made correct rejections when the target was absent in *drawing-to-drawing* condition. The performance in *drawing-to-drawing* trials in the target-present condition was similar in pattern to the full-face neutral target-present condition in the 1-target-in-10-array task (Experiment 5) of Bruce et al. (1999). Specifically, the percentage of hits was distinctively high compared with percentage of incorrect and miss responses, which were similar to each other. The correct rejection rate in Bruce et al. (1999) was about 70%, whereas our correct rejection rate was

about 52%, which means that, in about half of the target-absent trials, participants selected an image that they thought was the target. Part of the difficulty in this task could be attributed to the other-age effect (Anastasi & Rhodes, 2005), as the participants were of a different age to the faces they were comparing. However, a more important determinant is that the visual information available in the drawings is much reduced compared with photos, such that visual comparison could only be made based on limited cues.

In terms of target absent trials, participants were more likely to select one of the images in the array than report the target was not there. It is understandable that it would be hard to be certain that the target is definitely not in the array. The tendency to select someone who is not the target (in both target absent and present trials) in photo-to-drawing condition implies that a photo target could possibly fit with several of the drawings. Thus, participants were more likely to find a foil that meets the same general description as the photo target.

It is difficult to evaluate and further interpret the performances in current line-up task. Therefore, I would like to repeat the line-up task with only target-present trials, which evaluations on performances could be made on the basis of a chance performance.

Experiment 6

In Experiment 5, target-present and target-absent trials were presented intermixed. This design made task performance hard to evaluate, specifically comparisons against chance. The current experiment uses a simplified design based on target-present trials only. This design should allow direct comparison between participant performance and a known level of chance performance.

Methods

Participants

Participants were 60 students, who are naïve to the stimuli, from University of York [mean age (*SD*): 22.40 (3.09); age range: 18 – 27; 30 male, 30 female]. All participants gave written informed consent prior to the experiment and took part in the study in exchange for a small payment or course credit.

Stimuli and Design

The method was the same as for Experiment 1, except that only target-present trials were used, so that each participant was presented with 30 trials (15 drawing-to-drawing and 15 photo-to-drawing). As before, the target could be presented as a photograph or as a drawing. Participants were told that the target was present in every array, and that their task was to pick out the match.

Results

Accuracy rates of 68% and 43% were recorded for *drawing-to-drawing* and *photo-to-drawing* conditions respectively. A within-subjects t-test confirmed that this difference was significant [$t(59) = 6.81, p < .001$]. Separate one-sample *t*-tests against chance (20%, as there is one target out of five available options in array) were carried out to compare these performance levels against random guessing. These analyses confirmed that there were significantly more hits in both the *drawing-to-drawing* condition [$t(59) = 16.48, p < .001$] and the *photo-to-drawing* condition [$t(59) = 12.374, p < .001$] than would be expected by chance alone.

Discussion

In this face matching task, participants performed better in drawing-to-drawing trials than in photo-to-drawing trials, even though the photos contained more face information than the drawings. Evidently, this information did not facilitate recognition. In previous work (White, Burton, Jenkins, & Kemp, 2014), matching accuracy improved when participants compared two photos against one, but no further improvement was seen when they compared three or four photos against one. This finding demonstrates recognition performance is not always better with more available information. Ritchie and Burton (2017) found that exposure to high-variability images of a person (photos taken on different occasions) compared to low-variability images of a person (photos taken on the same occasion) enhanced learning of new identities. Together, these findings suggest that the degree of image variability is important for acquiring representations of novel faces. The increase in variability enables the viewer to extract information about the idiosyncratic ways in which people vary, leading to representations that better generalise across images. The increase of information in the current task may be working in the opposite direction. If a photo can match more than one possible drawing that meets the similar description, it may be difficult to carry over any idiosyncratic information between the drawings and photos.

In a line-up task, responses depend partly on the target, but also partly on the foils. The selection of foils largely determines the task difficulty. For example, for a target with blonde hair, an array containing foils with blonde hair would make a more difficult task than an array containing foils with brown hair. In the experiments conducted here, it is unclear how much the performance was driven by similarity among items in the arrays. Therefore, I would like to further simplify the task into a pair matching task.

Experiment 7

In a paired matching task, only two images (target-target, or target-nontarget) are displayed in each trial. The participants' task is simply to decide whether the two images show the same person or two different people. The advantage of this task is that any possible effects of foils are eliminated. In this situation, the required identity judgements must be based solely on the target and the comparison image.

Methods

Participants

Participants were 30 students, who are naïve to the stimuli, from University of York [mean age (*SD*): 21.10 (2.35); age range: 18 – 25; 15 male, 15 female]. All participants gave written informed consent prior to the experiment and took part in the study in exchange for a small payment or course credit.

Stimuli and Design

Participants were presented with 120 consecutive random-ordered trials. In each trial, they were asked to decide if the two face images presented on the screen show the same person or two different people. There were four types of trials, 30 of each type: (i) drawing-to-drawing match trials (25%): a drawing of the target plus another drawing of the target (Figure 4.3.A); (ii) drawing-to-drawing mismatch trials (25%): a drawing of the target drawing plus a nontarget drawing from the array used in Experiment 5 (Figure 4.3.B); (iii) photo-to-drawing match trials (25%): a photo of the target plus a drawing of the target (Figure 4.3.C); (iv) photo-to-drawing mismatch trials (25%): a photo of the target plus a nontarget drawing from the array used in Experiment 5 (Figure 4.3.D).



A. Drawing-to-Drawing Target-Target Match



B. Drawing-to-Drawing Target-Nontarget Mismatch



C. Photo-to-Drawing Target-Target Match



D. Photo-to-Drawing Target-Nontarget Mismatch

Figure 4.3. Example trials in Experiment 7. (A) Drawing-to-drawing target-target match pair. (B) Drawing-to-drawing target-nontarget mismatch pair. (C) Photo-to-drawing target-target match pair. (D) Photo-to-drawing target-nontarget mismatch pair.

The paired images were displayed side by side at the centre of a 21.5-inch computer screen. Each image measured approximately 10×15 cm height on the screen. Stimulus presentation and recording of responses were controlled by an Apple Macintosh computer running PsychoPy2 (Peirce, 2007).

Procedure

Participants were instructed that they would see pairs of images, which could be photos or drawings, and that their task was to decide whether or not they are the same person.

Participants were presented with 120 consecutive trials in random order. There was no time limit for the comparison made on each trial. Participants were asked to guess for which they felt unable to reach a decision.

Results

Table 4.2 shows the mean percentages correct for each of the four conditions. In the match condition, participants performed almost perfectly (99% accuracy) with drawing-to-drawing pairs, whereas accuracy was only 56% with photo-to-drawing pairs.

Table 4.2. Summary the mean percentage corrects of drawing-drawing and photo-drawing pairs in target-target match and target-nontarget conditions.

Target Format	Target-target Match		Target-nontarget Mismatch		Total	
	Correct	SD	Correct	SD	Correct	SD
Drawing-to-Drawing	99	2	83	25	91	19
Photo-to-Drawing	56	19	83	25	70	26
Total	78	25	83	25		

Separate one-sample t-tests against chance level (50%) were conducted for each condition. In drawing-to-drawing conditions, participants' performance was significantly above chance in both match trials [$t(29) = 167.55, p < .001$] and mismatch trials [$t(29) = 7.35, p < .001$]. In photo-to-drawing tasks, performance was above chance on mismatch trials [$t(29) = 7.35, p < .001$], but not in match trials [$t(29) = 1.75, p > .05$].

A 2 (target format) \times 2 (match or mismatch pair) ANOVA was conducted on the percentage accuracy data. This analysis revealed a significant main effect of target format, with higher accuracy in the drawing-to-drawing condition ($M = .91, SD = .19$) than the photo-to-drawing condition ($M = .70, SD = .26$), [$F(1,29) = 152.08, p < .001, \eta_p^2 = .84$]. There was no significant main effect of pairing, with similar overall levels of performance for match ($M = .78, SD = .25$) and mismatch trials ($M = .83, SD = .25$) [$F(1, 29) = 1.13, p > .05, \eta_p^2 = .04$].

The interaction between these factors was significant [$F(1, 29) = 152.08, p < .001, \eta_p^2 = .84$]. Simple main effects analyses showed that the effect of pairing was significant in photo-to-drawing pairs [$F(1, 58) = 22.39, p < .001, \eta_p^2 = .28$], with higher accuracy on mismatch trials than in match trials. The effect of pairing was also significant in drawing-to-drawing pairs [$F(1, 58) = 7.31, p < .01, \eta_p^2 = .11$], but in the opposite direction.

The effect of target format was significant for match pairs [$F(1, 58) = 304.16, p < .001, \eta_p^2 = .84$], with higher accuracy for drawing-to-drawing trials than for photo-to-drawing trials. The effect of target format for mismatch pairs was not significant [$F(1, 58) = .00, p > .05, \eta_p^2 = .00$].

Discussion

Overall accuracy rates of drawing-to-drawing (91%) and photo-to-drawing (70%) trials were both significantly above chance, confirming that participants could identify targets in both conditions, although and change in image format significantly reduced accuracy. Specifically, drawing-to-drawing comparisons were easier than photo-to-drawing comparisons.

For drawing-to-drawing trials, accuracy rates of approximately 99% and 83% were recorded in target-target match and target-nontarget mismatch trials respectively. These accuracy rates are high, compared with matching photos of unfamiliar faces (e.g., Megreya & Burton, 2006, 2007). One possible explanation for high accuracy in current task is that the drawing-to-drawing pairs were more akin to image matching (Bruce, 1982), with little variability among different drawings of the same person. This is perhaps not surprising, given that all of the drawings of a given face were based on a single photograph. In addition, several sources of image variability that arise in photographs do not apply to this type of cartoon (e.g. changes in environmental lighting).

In the photo-to-drawing trials, accuracy rates of 56% and 83% were recorded for match and mismatch trials respectively. Although accuracy for mismatch pairs was significantly better than chance, participants did not perform above chance in match trials. One possible explanation for this pattern is that the change in image format exaggerated visual differences between the two images, making it harder to say ‘yes, they are the same person’. This is different to the previous line-up experiments. In Experiment 5, participants compared the target to a whole array of options. In that situation, it is difficult to say ‘no, the target is definitely not there’. In Experiment 6 (target-present only), participants were not offered a chance to say no. Solving the identity puzzle through a process of elimination is not possible in a paired matching task, meanwhile the image similarity is low in this photo-to-drawing condition. Both of these factors likely contribute to the observed pattern of performance.

General Discussion

In this series of experiments, I conducted different identification tasks with photos and simple line drawings (based on the same photos) from *The Beano* (Beano Studio, 2018). In these comic strips, the artists created stories based on a single face photo and a name. A certain likeness to the photographic subject was maintained. However, it is important to acknowledge the purpose of the drawings was to convey a sense of character, rather than to support precise identification (cf. police sketches). Of particular interest for these studies, the drawings were very simple compared with photographs, in terms of shape and texture.

In Experiment 5, I used a line-up identification task based on Bruce et al. (1999), with one target accompanied by an array of five drawings. Participants were required to identify the target from the array or confirm the target was not in the array. The target was either a photograph or a drawing, and the corresponding target was either present or absent. I found that it was easier to carry out within-format comparisons in this situation, as performance was

higher for drawing-to-drawing pairs than for photo-to-drawing pairs. In target-absent trials, participants were more likely to select one of the images in the array than to report the target was not there.

Experiment 6 repeated the experiment with target-present line-ups only. This allowed comparison against chance performance. Accuracy in both conditions (drawing-to-drawing, and photo-to-drawing) was significantly better than chance level, confirming that some identifiable information could be encoded in the drawings and decoded by the viewers. In addition, participants still performed better in drawing-to-drawing trials than in photo-to-drawing trials, replicating the finding that within-format comparisons were easier than between-format comparison, despite containing less information. As it was unclear how far responses were influenced by foils in this line-up task, I carried out a final experiment using a paired matching task.

In Experiment 7, only two images were presented in each trial, and participants had to decide whether or not they showed the same person. The images were either two drawings or one drawing and a photo. Accuracy was high in the drawing-to-drawing condition, especially for match trials, possibly because different drawings of the same face resulted in very similar images (cf. different photos of the same face). Accuracy was much lower in the photo-to-drawing condition, especially for match trials. This aspect of the findings accords with previous demonstrations that viewers find it difficult to integrate dissimilar images (here, a cartoon and a photo) into the same identity.

The very sparse drawings used in these experiments raise the question of whether identity judgements were based solely on visual comparison. The reduction in information presumably made the cartoon drawings more abstract and more ambiguous representations of the photographic subjects, compared with photographic images. Given the low-level

dissimilarity in the images (e.g. differences in contrast, brightness, contours, etc.) it is possible that visual similarity was not the most important determinant of participants' responses. An alternative possibility is that participants made their identity decisions at a more abstract level than physical appearance, perhaps by comparing the overall impressions that the two images make.

McCloud (1993) claimed that 'the ultimate simplification of a symbol for face is language'. A drawing of a face and a description of a face can both be said to approximate an imagined face in mind of the artist or author. Any face drawing, however sparse, contains some visual information that could be used to support identity judgements. That is not the case for face descriptions, where direct visual comparison (i.e. words-to-photo) breaks down completely. Face descriptions also allow the possibility of separating information about physical descriptions concerning facial features (e.g., bushy eyebrows, broad nose) from character descriptions concerning the impression that the face makes on the viewer, (e.g., curious eyes, imperious face). The remaining experimental chapters examine this next level of abstraction for fictional faces.

Chapter 5

Matching Photos to Written Descriptions

Introduction

Anecdotal evidence suggests that the following experience is common among people who watch film adaptations of novels that they have read: a character appears on screen for the first time, prompting an immediate reaction in the viewer, “That’s not what the character looks like!” This reaction is intriguing, as it implies a mismatch between the onscreen appearance of the character and the viewer’s expectations. Apparently, the viewer’s mental picture of the character was sufficiently specific that it could clash with the person who was cast in the role. Where did this mental picture come from? In the case of a literary character, it can only have come from the text. Somehow, written descriptions of the character’s appearance (or perhaps their setting or deeds) can give rise to a visual representation of the person described. The experiments in this chapter address this process, by examining connections between written descriptions of faces and visual images of faces.

Previous studies of face description have tended to focus on applied questions related to eyewitness testimony. Schooler and Engstler-Schooler (1990) found that describing a previously seen face could hinder subsequent identification, and termed this phenomenon ‘verbal overshadowing’. Verbal overshadowing was observed when participants were provided with a description, and also when they generated their own description of an earlier seen face (Dodson, Johnson, & Schooler, 1997). However, Brown & Lloyd-Jones (2005) found that verbally describing a face could facilitate memory in an old/new face recognition task in which participants discriminated the viewed and described faces from distractors. This did not replicate the expected verbal overshadowing effect. Meissner and Brigham (2001) conducted a meta-analysis of verbal overshadowing in face identification and concluded that the effect occurred under certain conditions. Specifically, verbal overshadowing occurred when participants were required to perform the identification immediately after describing

the target, and when participants were engaged in forced recall, as opposed to a free recall, during the description task.

Todorov et al. (2008) offered a very different approach to face description, emerging from theoretical work on first impressions, rather than from applied face identification. They presented a two-dimensional space that captures the main dimensions of social evaluation of faces. Subsequent development by Sutherland et al. (2013) extended this space to three dimensions (trustworthiness, dominance, and attractiveness) based on analysis of natural face photographs. Each point in the constructed space represents a different facial image with particular loadings on each of the dimensions. Although this approach emerged from face perception research, nothing in its construction restricts it to use with face images. In principle, anything can be projected into the same social inference space as long as it can be rated on the underlying social dimensions. This is the contention at the heart of this chapter. If written descriptions of fictional faces can be rated for trustworthiness, dominance, and attractiveness, then those descriptions can be located in the same space as photographs of faces. A common metric space for descriptions and images should allow meaningful comparisons across these very different types of representation. For example, it should be possible to specify whether a description and an image are close together (similar social inference ratings) or far apart (dissimilar social inference ratings).

The experiments in this chapter develop this idea in an effort to make sense of the film-goer's experience. How could reading a description of a fictional face put a picture in the reader's mind?

Experiment 8

The purpose of this experiment was to obtain social inferences ratings for different face descriptions. Our main interest was readers' impressions of these descriptions. Participants

were asked to read a series of written descriptions, and to generate a mental picture of each face based on the written description. For each item, they were asked to evaluate the social dimensions of attractiveness, dominance, and trustworthiness, using separate Likert scales. For pictorially presented faces (e.g. photos or computer-generated images), evaluations on these dimensions appear to require little mental effort (Willis & Todorov, 2006), and tend to be rather consistent across viewers (Sutherland, Young, & Rhodes, 2017). Our intention here was simply to establish whether the same framework could be used to structure the impressions made by written descriptions too. If successful, we expected that participants would have no difficulty providing ratings for these materials, and that the resulting data would allow us to distinguish between different characters.

Previous studies have also reported consistent sex differences in social inference ratings for faces. Generally, female faces are rated high on attractiveness and trustworthiness, and low on dominance, whereas male faces show the opposite pattern—low on attractiveness and trustworthiness, high on dominance. Although these sex differences are not the main interest for this study, they do provide a useful point of comparison. If rating written descriptions engages similar processes to rating face images, then we should expect the same pattern of sex differences to emerge here.

Finally, using written materials provides an opportunity to investigate what *types* of descriptive information drive the reader's impressions. By selectively redacting references to *physical* attributes versus references to *character* attributes, it was possible to present these types of information independently or in combination.

Methods

Stimuli

We selected twenty-one descriptions of different minor characters (8 female, 13 male) from various Sherlock Holmes short stories (Doyle, 1890, 1895, 1998, 1999, 2001a, 2001b, 2004, 2007, 2009). Each extract consisted of contiguous text surrounding the character's first occurrence in the story, and included both *physical* description concerning attributes such as the shape and colouration of specific facial features (e.g., bushy eyebrows, broad nose) and *character* description concerning the impression that the face makes on the viewer, e.g., inquisitive eyes, haughty face). Each extract was modified to form three different versions—*Full* (Figure 5.1.A), *Physical* (Figure 5.1.B), and *Character* (Figure 5.1.C) descriptions. The *Full* description refers to the complete extract in its original form. To create the *Physical* descriptions, we redacted all elements of character description, leaving physical description intact. For *Character* descriptions, we took the opposite approach, redacting all elements of physical description and leaving only character description intact. The mean word counts for *Full*, *Physical*, and *Character* descriptions were 117 (*S.D.* = 40), 67 (*S.D.* = 31), and 79 (*S.D.* = 31) respectively. A t-test comparing word count for the *Physical* and *Character* descriptions found no significant difference in the lengths of these extracts [$t(40) = 1.25$, $p > .05$], indicating that the quantity of physical and character information was balanced overall.

She was very young - at the most nineteen, with a pale somewhat refined face, yellow hair, merry blue eyes, and shining teeth. Her beauty was of an ethereal type. She looked so white and light and fragile that she might have been the spirit of that storm-foam from out of which I plucked her. She had wreathed some of Madge's garments round her in a way which was quaint and not unbecoming.

A. Full Description

She was very young - at the most nineteen, with a pale [REDACTED] face, yellow hair, [REDACTED] blue eyes, and shining teeth. [REDACTED] She looked so white and light [REDACTED] [REDACTED] She had wreathed some of Madge's garments round her [REDACTED]

B . Physical Description

She was very young [REDACTED] with [REDACTED] [REDACTED] merry [REDACTED] eyes [REDACTED] Her beauty was of an ethereal type. She looked so [REDACTED] fragile that she might have been the spirit of that storm-foam from out of which I plucked her. She had wreathed some of Madge's garments round her in a way which was quaint and not unbecoming.

C. Character Description

Figure 5.1. Examples of (A) Full, (B) Physical, and (C) Character descriptions of one character.

Manipulation check To verify that the physical descriptions were seen as containing mainly physical information, and that the character descriptions were seen as containing mainly character information, independent readers were asked to categorise the written descriptions in a manipulation check. Sixty native English-speaking students (30 female, 30 male; mean

age 22.82; age range 20–26) were recruited from the University of York in exchange for a small payment or course credit. The three versions of the written descriptions (21 *full*, 21 *physical*, and 21 *character*) were presented individually in a random order on a computer screen (12-point Times New Roman font, double-line spaced). Presentation of stimuli and recording of responses were controlled by an Apple Macintosh computer running PsychoPy2 (Peirce, 2007). The readers' task was to simply read each extract in turn, and then indicate via keypress response the description contained more physical or character information.

A one-sample *t*-test against chance (50%) confirmed that a *full* description contained more physical information ($M = .72, S.D. = .23$) [$t(20) = 4.33, p < .001$]. Thus, physical information seemed to dominate participants' impressions, even though it did not dominate word count. Categorisation responses were consistent with the intended manipulation. After manipulation, a *physical* description contained significantly more physical information ($M = .94, S.D. = .08$) than chance [$t(20) = 26.29, p < .001$]; whereas a *character* description contained significantly less *physical* information ($M = .26, S.D. = .23$) than chance [$t(20) = 4.52, p < .001$].

The separation between *physical* descriptions and *character* descriptions forms the basis the rating study that follows.

Participants

Thirty native English-speaking students (15 female, 15 male; mean age 25.57; age range 22–30) were recruited from the University of York in exchange for a small payment or course credit.

Design

Each participant rated 7 *full* descriptions, 7 *physical* descriptions, and 7 *character* descriptions. The fictional characters were counterbalanced with respect to information condition so that, across the whole experiment, each character appeared in each condition an equal number of times, and each participant saw each character exactly once. The 21 trials were presented in a random order. The measures of interest were the three social inference ratings (*trustworthiness*, *dominance*, and *attractiveness*). Within each trial, these social inferences were rated in a random order.

Procedure

Participants were instructed to read each of the 21 written descriptions carefully. They were informed that some of the text could be redacted, and to make use of whatever information was available.

In each trial, a single paragraph of text was displayed on the centre of a 21.5-inch computer screen, with a rating meter from 1 to 7 presented beneath the paragraph. After reading the paragraph, participants were asked to form a mental picture of the fictional character. Based on this mental picture, they rated each of the three social inferences (*trustworthiness*, *dominance*, and *attractiveness*) in turn by clicking on the rating meter. Social inferences were rated one by one in a random order. To ensure that participants had sufficient time for reading, there was no time limit. The task only advanced when the participant made a response.

Results and discussion

Social inference ratings for the written descriptions are summarised in Table 5.1. Ratings of male and female characters were examined to allow comparison with previous research on social inferences from face images.

Table 5.1. Means and SDs of social inferences ratings for the full, physical and character descriptions in Experiment 8.

Social Inferences Rating (Out of 7)		Trustworthiness		Dominance		Attractiveness	
Description Version		Mean	S.D.	Mean	S.D.	Mean	S.D.
Full	Female Characters	4.87	.68	3.79	.93	5.56	.74
	Male Characters	3.70	.53	4.98	.76	3.75	.59
	Combined	4.15	.45	4.46	.56	4.26	.26
Physical	Female Characters	4.64	.64	4.06	.81	4.93	.90
	Male Characters	3.80	.56	4.56	.70	3.61	.61
	Combined	4.16	.41	4.46	.69	4.25	1.43
Character	Female Characters	4.79	.61	4.08	1.21	4.98	.79
	Male Characters	3.52	.55	4.54	.79	3.49	.79
	Combined	4.02	.40	4.35	.81	4.06	.71

Separate *t*-tests for each social dimension revealed that participants rated female characters as more trustworthy [$t(29) = 12.70, p < .001$], more attractive [$t(29) = 12.81, p < .001$], and less dominant [$t(29) = -5.69, p < .001$] than male characters. These significant sex differences confirm that the rating measures are sensitive enough to detect psychologically meaningful effects when using written materials. Moreover, for each dimension, the direction of the difference accords with previous work based on face images (e.g., Becker, Kenrick, Neuberg, Blackwell & Smith, 2007; Hess, Adams, Grammer & Kleck, 2009; O'Doherty, Dayan, Friston, Critchley & Dolan, 2003). There were no significant differences between description versions (*Full*, *Physical*, *Character*) for trustworthiness [$F(2, 87) = .94, p > .05$], dominance [$F(2, 87) = .25, p > .05$], or attractiveness [$F(2, 87) = 1.07, p > .05$], presumably because the *Physical* and *Character* information was 'pulling in the same direction', in the

sense that both types of information were supporting a coherent role. Overall, the structure in the data suggests that participants had no difficulty providing social inference ratings for the written descriptions, and that different roles can be distinguished on the basis of these ratings. The next experiment tests whether social inferences, as an abstract level of representation, can be used to connect written descriptions of individuals to images of individuals.

Experiment 9

The purpose of this experiment was to examine mental translation between written descriptions of faces and visual images of faces, using social inferences as a medium of exchange. Previous studies have investigated translation in both directions between these very different representational codes. For example, eyewitness testimony may require us to go from a visual image (the seen suspect) to a verbal description (a police statement) and back to a visual image (an artist's impression or photofit). This translation between codes implies that there must be a more abstract level of representation that can mediate between them, analogous to the propositional level in translation between languages (see *Figure 5.2*).

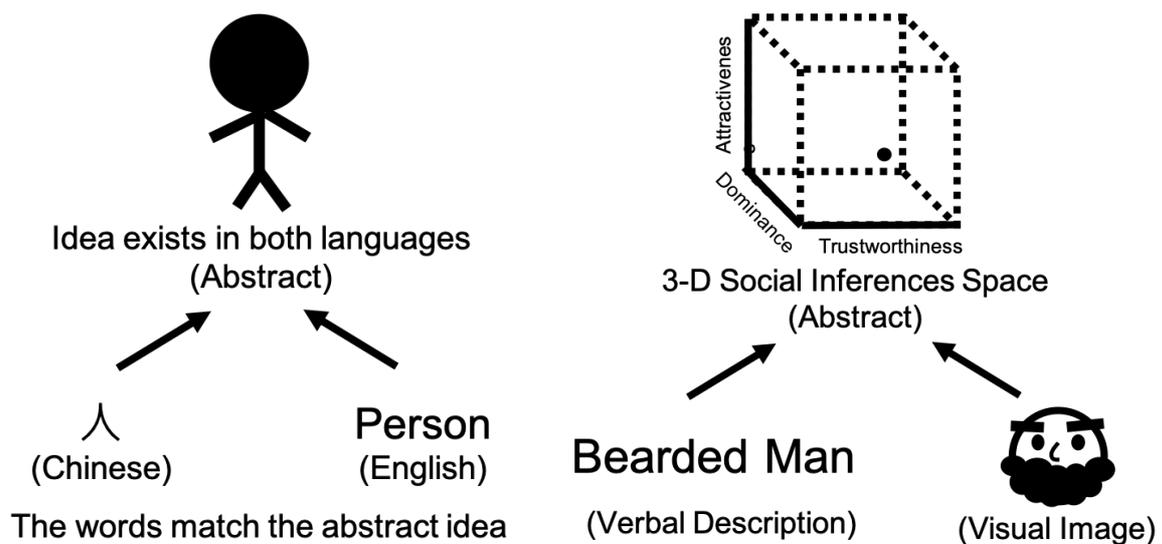


Figure 5.2. Different representations map onto the same abstract idea.

In the case of faces, the nature of this abstract level is not well understood, but it is possible to specify some criteria. First, the abstract representation cannot itself be visual or verbal, as that would only push the problem back further. Second, the abstract level must be accessible to both visual and verbal representations, in the sense that input of either type can be recoded into the required form.

One form of abstract representation that meets these criteria is the space defined by social dimensions (e.g. *trustworthiness*, *dominance*, and *attractiveness*). Social impressions are themselves neither visual or verbal, but they can be accessed via images and words. Indeed, face images seem to evoke social impressions very rapidly (Willis & Todorov, 2006). The preceding experiment shows that readers can easily generate social impressions from written descriptions (though presumably more slowly).

This framework provides one possible mechanism by which written descriptions of faces could give rise to mental images of faces: the written description evokes a certain pattern of social inferences, and that pattern of social inferences evokes a certain pattern of visual attributes. To be clear, the proposal is not that this is the *only* process involved. Rather, social inferences provide one possible bridge between different types of face representation.

The current experiment is based on a 3D social inference space along the dimensions of attractiveness, dominance, and trustworthiness. By projecting written descriptions of fictional characters and photographic images of real faces into this 3D social inferences space, it should be possible to make direct metric comparisons between these types of representation. For example, it should be possible to measure which of the images are closest to a given description in this common space (see Figure 5.3).

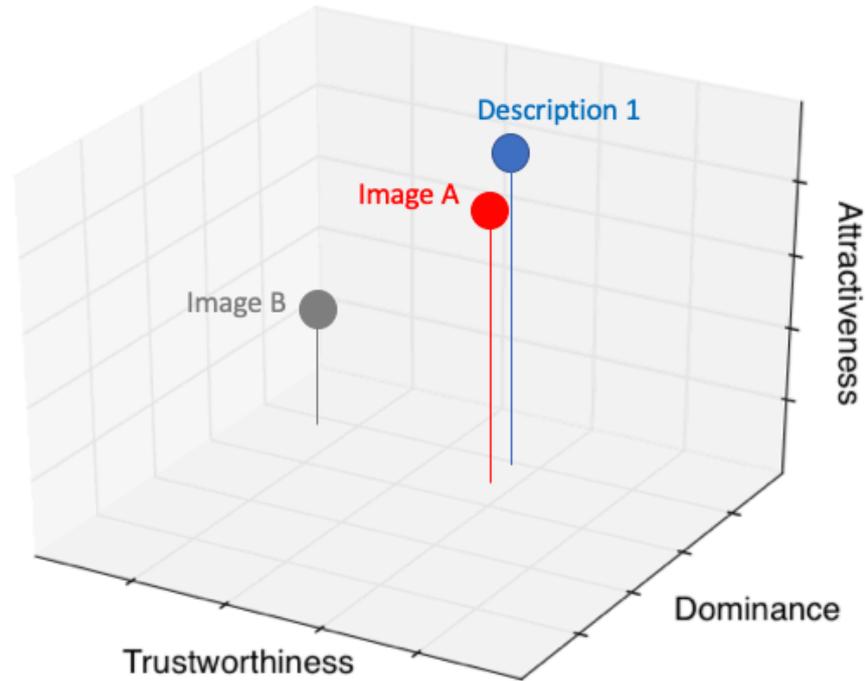


Figure 5.3. A 3D space along the dimensions of attractiveness, trustworthiness, and dominance, where both descriptions and images of face could be projected into this space. Absolute distances between descriptions and images can then be compared directly. In this space, Image A is the closest image to Description 1. Therefore, it would be regarded as the target face for Description 1. Image B is a distractor, as it is not the closest image to Description 1.

This arrangement opens a number of research questions. The current experiment focuses on two questions in particular. The first is whether different people reading the same description arrive at similar mental pictures. The second is how closely the mental picture and the description coincide at the level of social inferences. To answer these questions, I developed a new casting task based on pre-rated face descriptions (from Experiment 8) and pre-rated face photographs (from a published source). Participants were recruited as ‘casting agents’ who were tasked with reading written descriptions of fictional characters, and then choosing from an array of photos the best match for each role. Separate analyses address (i) the level of

agreement between participants, and (ii) whether the preferred face is close to the description in social inference space.

Methods

Participant

Sixty native English-speaking students (38 female, 22 male; mean age 26.13; age range 19–33), who were naïve to the stimuli, were recruited from the University of York in exchange for a small payment or course credit.

Stimuli

Written descriptions The written descriptions of fictional characters from Sherlock Holmes stories were presented as roles to be cast in this experiment (*physical*, *character*, and *full* descriptions for each of 21 fictional characters; 63 descriptions in total). All of these descriptions had already been rated on *trustworthiness*, *dominance*, and *attractiveness* by participants in Experiment 8.

Face photographs Rated photographs of unfamiliar faces were sampled from a published database. Sutherland et al. (2013) collected *trustworthiness*, *dominance*, *attractiveness*, and *age* ratings for 1000 adult Caucasian faces (500 male, 500 female) gathered from the internet (Santos & Young, 2005, 2008, 2011). The photographs in this database were standardised to a height of 150 pixels (approximately 5 cm on screen) and were cropped around the head and shoulders to minimise extraneous background. Non-Caucasian faces were excluded to avoid the impact of other race effects (Hugenberg, Young, Sacco, & Bernstein, 2011; Rossion & Michel, 2011), which could potentially influence facial impressions. All other image variables were deliberately left unstandardised, to capture a naturalistic representation of sources of variation that might contribute to first impressions. These include facial

characteristics such as age, expression, pose, health; facial hair, glasses and piercings; and image characteristics including lighting, background, camera type and angle.

Pairing faces and descriptions For each of the 63 written descriptions, I established which of the 1000 rated photographs was the best match in the 3D social inference space. The best match was defined as the same-sex image with the shortest linear distance to the target description. For descriptions that included age information, the best-match image was constrained to be of congruent age. In a few cases, two similar descriptions mapped onto the best-match image. Thus, 31 male best-match images were identified for the 39 male descriptions (13 male characters with 3 types of description), and 22 female best-match images were identified for the 24 female descriptions (8 female characters with 3 types of description). These selected photos were arranged into a male array of 31 faces and a female array of 22 faces (see Figure 5.4). Each face was assigned an identifying number to allow responses to be coded. Constructing the photo arrays in this way ensured that each of the photos was the best match for a particular description. For any given description, the array contained a single best-match photo (the ‘target’) among a number of other photos (‘distractors’).

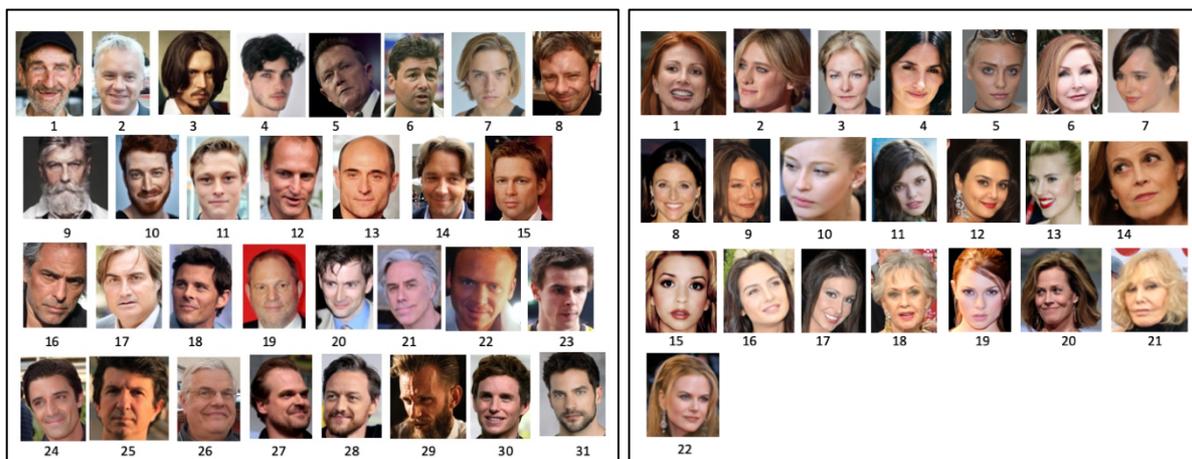


Figure 5.4. Printed photo arrays available for inspection throughout the task (left: for male characters; right: for female characters).

Design

The 63 written descriptions (3 information conditions for each of 21 fictional characters) were organised into three booklets, each containing 7 *full* descriptions, 7 *physical* descriptions, and 7 *character* descriptions in a random order. Fictional roles were rotated around information conditions so that, across the whole experiment, each fictional role was seen in each condition an equal number of times, and each participant saw each character exactly once. For each role, the participant's task was to choose the photo that best fits that role by writing the identifying number in the booklet.

Procedure

Participants were presented with the following written instructions:

“Imagine that you are a casting agent for a theatre company. Your task is to select the best actors and actresses for particular roles.

Before deciding which actors to audition, you will first read a short descriptive paragraph about the character being cast. After reading the description, you should try to form a mental picture of the character. Your task is to select the best-fit actor or actress from the portfolio photos provided. Makeup and hair artists can be recruited at a later stage to adjust makeup and hairstyles.

You do not have to audition all the actors.

You are free to audition the same actor or actress for more than one role.

There is no time limitation for the task.”

Participants were asked to work through the entire booklet, reading each of the 21 written descriptions carefully. They were informed that some of the text could be redacted, and to make use of whatever information was available. Printed photo arrays (Figure 5.4) were

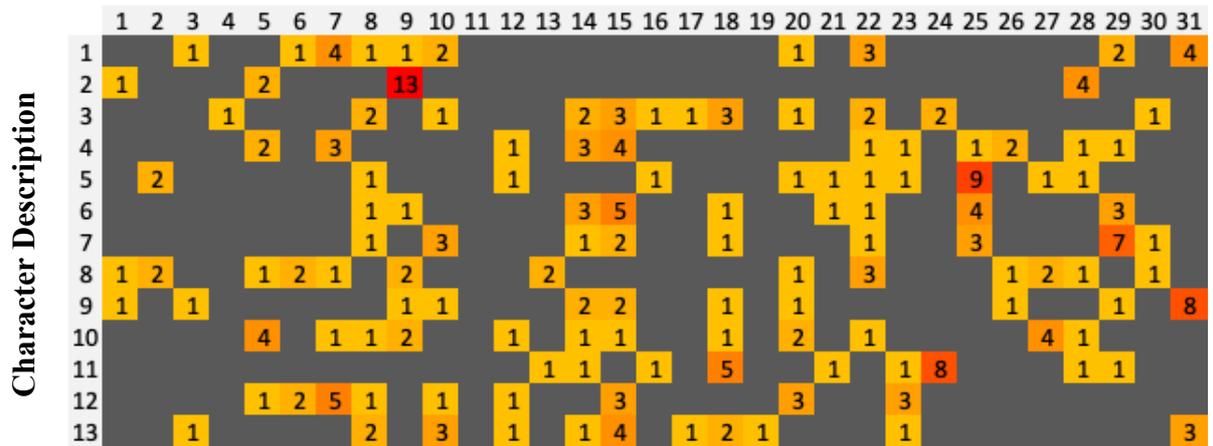
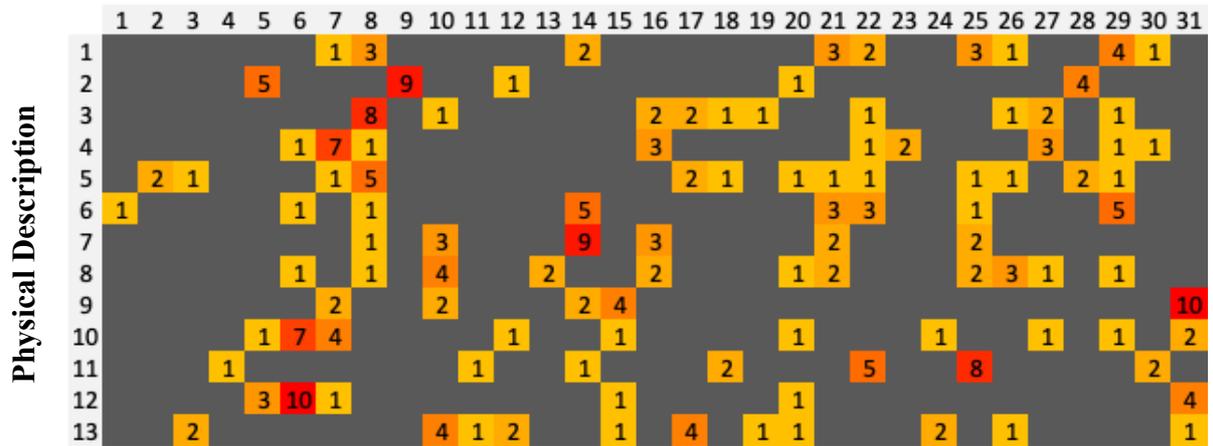
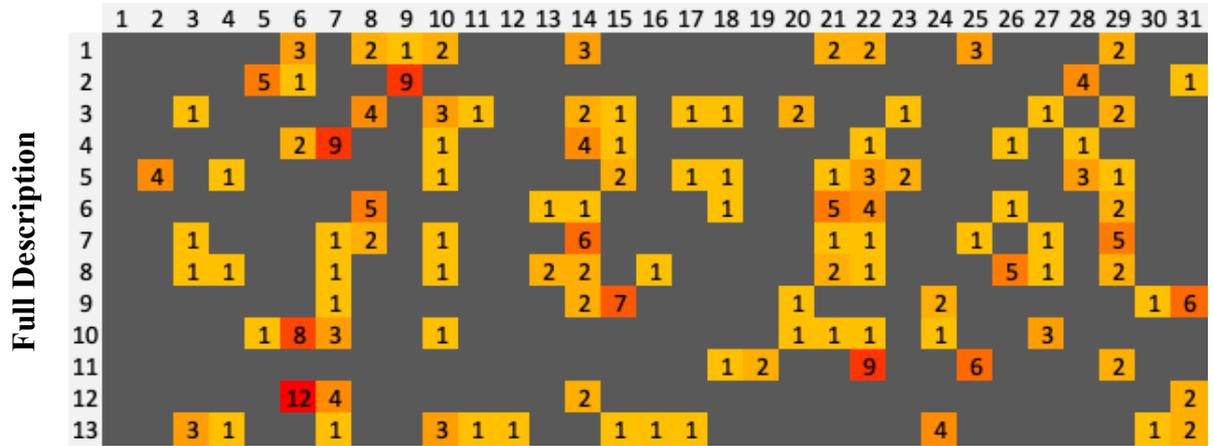
available for inspection throughout the task. Participants recorded their casting decisions by writing the identifying number for the chosen photograph next to the written description. To ensure all participants had sufficient time for reading, and examining the photographs, no time limit was imposed. The entire task took approximately 40 minutes to complete.

Results and Discussion

Agreement between participants

Participants' casting decisions are summarised in Figure 5.5. In each matrix, rows represent descriptions, and columns represent photos. Clustering of responses indicates agreement among participants. A completely even distribution across each row would indicate no agreement. A single hotspot in each row would indicate total agreement. The actual data show intermediate levels of agreement. For each description, casting decisions cluster around a subset of the available options (coloured cells). Importantly, many of the options are chosen by none of the participants (grey cells).

Male Profile Picture



Female Profile Picture

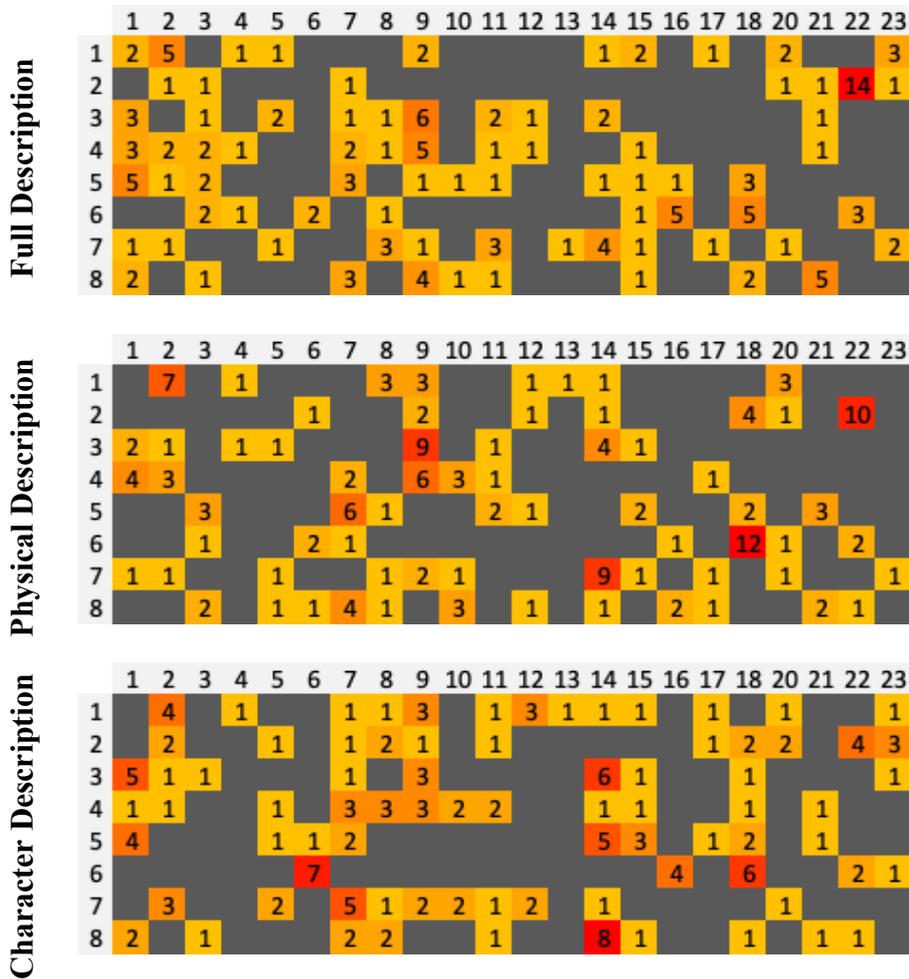


Figure 5.5. Number of times that specific profile pictures being chosen as the best-match for each description (i.e. participants' agreement) in Full, Physical, and Character versions in Experiment 9. Each column represents a profile picture in the array, and each row represents a verbal description. Each description was cast by 20 participants, so that a single profile picture can be picked a maximum of 20 times. Photos that were never chosen are shown in grey. Among photos that were chosen at least once, popularity is shown by the yellow–red gradient (yellow = 1, red = max).

To quantify the levels of agreement among the 20 participants, Fleiss’s Kappa (κ) was calculated separately for male and female characters in each information condition. The results of this analysis are summarised in Table 5.2.

Table 5.2. Fleiss’s κ statistic indicating interrater reliability among 20 participants’ judgements of best-fit actors and actresses in three information conditions.

		Male Characters	Female Characters
Physical Description	K	.125*	.130*
	95% CI	.116 to .134	.116 to .144
Character Description	K	.092*	.052*
	95% CI	.083 to .101	.011 to .093
Full Description	K	.121*	.085*
	95% CI	.111 to .130	.072 to .098

* $p < .001$

In all of these conditions, the agreement between raters is significantly higher than would be expected by chance. Kappa values below 0.2 are characterised as showing ‘slight agreement’ in the scheme proposed by Landis & Koch (1977). However, raw Kappa values can be difficult to interpret when the number of categories is high, as there is greater potential for disagreement (Sim & Wright, 2005). In general, two aspects of the data contribute to the statistically significant pattern. First, a few of the options were endorsed by multiple participants (partial agreement). Second, many of the options were rejected by all of the participants (total agreement).

Proximity of chosen photos to written descriptions

The preceding analysis revealed some agreement among participants as to which face was the best fit for the role. This agreement allows us to identify the most commonly chosen photo or photos for each description (as shown in Figure 5.5). By combining these behavioural data with the metric data from Experiment 8, we can ask whether the preferred photos were those

that were close to the description in 3D social space. If so, then representations based on social impressions could plausibly mediate the translation from written descriptions to mental images. On the other hand, if there is no relation between the proximity of a photo to the description and the frequency with which it is chosen, then representations based on social impressions probably do not mediate this translation.

For each of the 63 descriptions, I first calculated the average of the absolute distances in 3D social space between each of the array photos and that description. For comparison, I then calculated the weighted average of the absolute distances between just the *chosen* photos and the written descriptions. This comparison revealed that chosen photos (Weighted average distance = 1.49, *S.D.* = .72) were significantly closer to the description than were the array photos overall (Average distance = 1.80, *S.D.* = .75), [$t(62) = 3.10, p < .05$]. Specifically, for the 21 *full* descriptions, chosen photos (Weighted average distance = 1.54, *S.D.* = .55) were significantly closer to the description than were the array photos overall (Average distance = 2.13, *S.D.* = .52), [$t(20) = 7.17, p < .001$]. For the 21 *physical* descriptions, chosen photos (Weighted average distance = 1.35, *S.D.* = .78) were again significantly closer to the description than were the array photos overall (Average distance = 1.95, *S.D.* = .55), [$t(20) = 2.78, p < .05$]. For the 21 *character* descriptions, chosen photos (Weighted average distance = 1.58, *S.D.* = .90) were significantly closer to the description than were the array photos overall (Average distance = 2.11, *S.D.* = .64), [$t(20) = 4.96, p < .001$].

In other words, the probability that a particular photo will be chosen can be predicted in part from its proximity to the written description. Given the complexity of the casting decision in this experiment, I next devised a more straightforward casting task, based on the same stimulus set, that was designed to give the proximity effect its best shot.

Experiment 10

In Experiment 9, there was some agreement among participants as to the faces that captured the appearance of the fictional character, as well as faces that did not. In addition, the chosen faces tended to be close to the written description in 3D social space, relative to the faces as a whole. The current experiment focuses on this proximity effect in a simplified casting task in which the number of casting options was reduced.

Several previous studies have shown that it is difficult to process multiple faces. For example, processing a face has been shown to block response competition effects (Bindemann, Burton, & Jenkins, 2005) and repetition priming effects (Bindemann, Jenkins, & Burton, 2007) from a second face in the display. Such findings have been interpreted as evidence for strict capacity limits in face perception, with faces being processed one at a time. On this account, multiple faces cannot be processing in parallel, and must be processed serially instead. For tasks that involve comparisons of multiple faces, this constraint presumably imposes additional demands on memory.

Identity decisions in line-up tasks are a good example of this situation. Bruce et al. (2001) presented one full-face target image captured from high-quality video alongside a photographic array of 10 potential matches (1-in-10 line-up). The task was to decide whether or not the target was present in the line-up, and if so, to pick out the culprit. Participants performed at about 70% accuracy for those arrays in which the target was present (with subjects claiming no match on roughly 20% of trials, and choosing the wrong face on roughly 10% of trials). However, the potential influence of non-target faces in these line-ups—particularly their similarity to the target—is difficult to interpret. Partly in response to this difficulty, later face identification experiments simplified the task by reducing the number of images involved. For example, Megreya and Burton (2006) introduced a much simpler 1-in-1

‘line-up’ in which an image of the target was presented alongside a photo of either the target (target-present) or a distractor (target-absent). Participants still found this task difficult, performing at about 82% accuracy in the target-present condition. However, in this simplified task, questions concerning the influence of other faces in the display do not arise, making interpretation more straightforward.

In the present experiment, simplifying the casting task served two purposes. The first was to reduce the cognitive burden of selecting the best match. The second was to test the strongest possible manipulation of proximity with these materials (i.e. the written descriptions and face photographs used in Experiment 9). In the new version of the task, participants read the written description of a fictional character as before. This time however, they chose the best match face from just two options. One was the photo that was closest to the description in 3D social space. The other was the photo that was furthest from the description (with the constraint that it must be the same sex and same age band). If proximity of the photo to the description predicts casting decisions, then participants should choose the nearest photo more frequently than one would expect by chance (>50%).

Methods

Participants

Sixty native English-speaking students (30 female, 30 male; mean age 23.15; age range 20–26), who were naïve to the stimuli, were recruited from the University of York in exchange for a small payment or course credit.

Stimuli and design

Images with the nearest and the furthest social distances to each description were selected from the profile arrays used in Experiment 9. There were 21 descriptions with 3 versions (*full*, *physical*, and *character*) of each description, resulting 63 pairs of images. The

experiment was conducted in a counterbalanced design that each participant read 7 *full*, 7 *physical*, and 7 *character* descriptions.

Procedure

Participants were asked to work through the entire booklet, reading each of the 21 written descriptions carefully. They were informed that some of the text could be redacted, and to make use of whatever information was available. In each trial, participants read a description and were instructed to form a mental image of the role being described. Participants were to select the better-fit actor/actors from the two images presented and record their casting decision by circle the number under the profile picture, see Figure 5.6 for a task example. To ensure all participants had sufficient time for reading, and examining the photographs, no time limit was imposed. The entire task took approximately 40 minutes to complete.

The first impression left by Mrs. Lyons was one of extreme beauty. Her eyes and hair were of the same rich hazel colour, and her cheeks, though considerably freckled, were flushed with the exquisite bloom of the brunette, the dainty pink which lurks at the heart of the sulphur rose. [redacted]
[redacted] There was something subtly wrong with the face, [redacted]
[redacted]
[redacted] At the moment I was simply conscious that I was in the presence of a very handsome woman, [redacted]
[redacted]



Actress 1

Actress 2

Figure 5.6. An example of the experiment task. Participants read the paragraph showed at the left and select one actress from the two profile pictures at the right. Their casting decision was recorded by circle the number under the profile picture.

Results and Discussion

Overall, participants chose the nearest photo on 72% of trials, and the furthest image on 28% of trials. A one-sample *t*-test against chance (50%) confirmed that this was a statistically significant difference [$t(59) = 20.36, p < .001$]. Separate *t*-tests revealed a preference for the nearest photo in the *full* condition ($M = 77%$) [$t(59) = 12.08, p < .001$], the *physical* condition ($M = 71%$) [$t(59) = 10.75, p < .001$], and the *character* condition ($M = 66%$) [$t(59) = 7.87, p < .001$].

Experiment 9 showed that chosen photos were closer to the descriptions in 3D social space than the photos as a whole. In this simplified experiment, participants chose photos that were near the description over photos that were far from the description. The important point of this finding is that these distances were defined solely in terms of social inference ratings, whereas participants' judgements in this task were based on visual comparison. The convergence between metric and visual comparisons in this situation shows that social inference ratings can mediate between written and visual representations of faces.

Another interesting aspect of these results is that *physical* description and *character* description appear to provide additive information. Either type of description alone led to above chance performance (66% or 71%). However, performance was higher still when both types of description were present (77%). I return to the issue of physical and character information in the next experiments.

General Discussion

The experiments in this chapter were designed to investigate translation between textual and visual representations of faces using social inference ratings (trustworthiness, dominance, and attractiveness) as mediators. Ratings on these dimensions were used to construct an 3D social

inference space that could accommodate both verbal descriptions and visual images of faces, so that direct metric comparisons between these representations could be made.

In Experiment 8, written descriptions of fictional characters were selectively redacted to create *physical*, *character*, and *full* versions of the descriptions, as verified by a subsequent manipulation check. Interestingly, physical information seemed to dominate readers' impressions of these roles, even though it did not dominate by word count. A separate group of participants was recruited to read the text descriptions, and to form a mental image of each fictional face in turn. The participants were asked to rate each of their mental images on the social dimensions of attractiveness, dominance, and trustworthiness. These ratings produced the standard pattern of sex differences seen in ratings of visual face images, and were statistically equivalent across *physical*, *character*, and *full* description conditions. The coherence of these findings suggests that participants had no difficulty complying with the task instructions, and that imagined faces, based on written descriptions, can meaningfully be located in 3D social inference space. The next step was to approximate these imagined faces through comparisons with photographic images.

In preparing Experiment 9, I paired each rated description with a rated face photograph from Sutherland et al.'s (2013) database. In each case, the paired photo was the nearest neighbour in 3D social inferences space. These materials formed the basis of a new 'casting task' in which participants read each description in turn, and chose from a photo array the most suitable actors or actresses for the role. There was substantial agreement in casting decisions across participants. For any given description, participants tended to choose one of a small number of options. Moreover, chosen faces tended to be near the written description in social space (proximity effect). This clustering could indicate that each of the popular options contained some of the attributes that participants were seeking.

Experiment 10 involved a simplified casting task that was designed to maximise the proximity effect—the tendency for participants to cast faces that were near the description. In this version of the task, the face array was reduced to just two options—the face that was nearest the description, and the face that was furthest from the description. Participants showed a strong preference for the near face over the far face, even though all of the distance metrics were based on other people’s ratings.

I started this chapter by pointing out that casting decisions in film adaptations are sometimes jarring to viewers who have read the novel. The experimental findings go some way to fleshing out this phenomenon: physical and character descriptions both provide cues to facial appearance—either directly (by specifying them), or indirectly (via stereotypical associations). These cues can be assembled into a facial impression through mental imagery. At the same time, the description, or the imagined face, or both, evoke a social impression. An unfortunate actor or actress could mismatch the viewer’s mental representation either at the level of imagery, or at the level of social impression, and these two levels interact. The fact that such mismatches seem to be the exception rather than the rule suggests considerable agreement among authors, readers, and real casting agents.

In these casting experiments, visualisations of readers’ mental images have been limited to face photographs that best approximate the imagined appearance. The experiments in the next chapter use an alternative visualisation that combines information from multiple photos.

Chapter 6

Mental Imagery from Written Descriptions

Introduction

The casting task (Experiment 9) required participants to read a description of a fictional face, and then select the best match for that face from an array of photographs. By projecting the written descriptions and the comparison photographs into a common metric space, distances between photos and descriptions could be compared directly. A key finding from this analysis was that selected photographs were (on average) closer to the target description than were non-selected photographs (on average).

As well as being interesting in its own right, this finding provides a baseline against which to compare further progress in divining the appearance of imagined faces. Participants in the casting task chose the best match *from a fixed set of options*. Given the limited options, it seems likely that a chosen image deviates from the imagined appearance to some extent; and that different choices deviate from the imagined appearance in different ways. For instance, one participant might read the description and settle for a particular photo, even though the depicted face looks a bit too broad. Another participant might read the same description and settle for a different photo, even though the depicted face looks a bit too weathered. If we assume that deviations of the chosen photos from the imagined face are uncorrelated, then it should be possible to approximate an imagined face more closely by averaging the chosen photographs together (Galton, 1907).

Image averaging has previously been used to refine visual representations of facial identities. For example, Burton, Jenkins, Hancock, and White (2005) showed that different photos of the same face were more similar to the identity average than they were to each other (on average). Identity averages are generally better recognised by human viewers (Burton et al., 2005) and by automatic face recognition systems (Jenkins & Burton, 2008; Robertson, Kramer, & Burton, 2015), compared with constituent photographs. The implication is that an

average image better approximates the stored representation of the person's face. In essence, the averaging process preserves aspects of facial appearance that are consistent across photographs, while washing out aspects of appearance that vary from one photo to the next.

The experiments in this chapter apply the same logic to fictional faces. The social inferences framework described in Chapter 5 is again useful here, as it provides a common space for comparing different types of face representation. In particular, it allows us to compare the locations of single photos and average images to the location of the description (Experiment 11). A basic behavioural test of these average images is whether new readers can tell which average matches a given description (Experiment 12). However, the separation of *physical* description and *character* description in Chapter 5 allows for more sophisticated questions too. Experiment 13 addresses the question of whether an average based on *physical* description and an average based on *character* description converge, in the sense that they are seen as the same person. Experiment 14 tests whether *physical* or *character* information plays a stronger role in defining the imagined face.

Experiment 11

The previous chapter was designed to connect existing faces photographs with written descriptions. Participants showed considerable agreement on which photos could plausibly match the person in description, and which faces could not. For any given description in Experiment 9, participants tended to choose one of a small number of options. One possible explanation of this clustering is that each of these popular choices contained some of the attributes that participants were looking for. If so, then combining those popular choices by image averaging should preserve the sought after attributes while diluting attributes that do not fit.

To test this possibility, I created a set of average images, and asked a new group of participants to rate the average images on the same social dimensions that were used in Experiment 9 (*trustworthiness, dominance, and attractiveness*). In this way, the average images could be projected into the same metric space as the written descriptions and the chosen photographs. To ensure that each average image reflected the frequency with which constituent photos were chosen, weighted averages were used, in which the contribution of each photo to the average was proportional to the number of participants who chose it. To track the influence of different types of descriptive information, separate weighted average images were created for each of the three information conditions in the casting study—*Full* description, *Physical* description, and *Character* description.

If the average image is closer to the description than are the chosen photographs (on average), this would suggest that chosen photographs approximate an imagined appearance that is shared across participants. In other words, reading a description of a face evokes similar mental images in different readers. On the other hand, if the average image is not closer to the description than the selected photographs (on average), this would suggest that chosen photographs do not approximate an imagined appearance that is shared across participants (or that standard assumptions about the metric space do not hold). Comparing the different information conditions should indicate the relative importance of physical and character information in shaping mental pictures of faces.

Methods

Participants

Sixty participants (31 female, 29 male; mean age 21.42; age range 18–25), who were naïve to the stimuli and the unmorphed original images, were recruited from the University of York in exchange for a small payment or course credit.

Stimuli

Weighted averaged face images Identity averages were created for each description by morphing together chosen photos from the casting task in Experiment 9. These constituent photos contributed to the average image in proportion to the frequency with which they were chosen (weighted averages). For example, if 17 participants chose Photo A as the best match for a given description, and 3 participants chose Photo B, the weighted average for that identity would contain 17 copies of Photo A and 3 copies of Photo B. Figure 6.1 illustrates this process for an actual case.

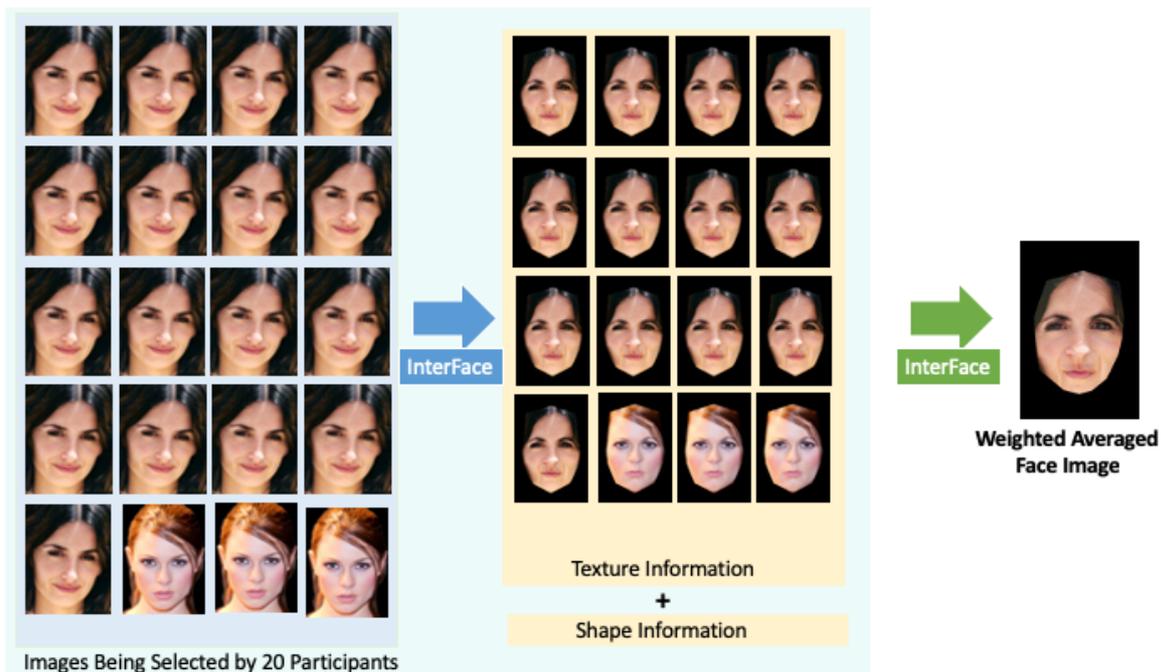


Figure 6.1. Process of creating a weighted averaged face image using InterFace (Kramer et al., 2017a). Images at the left are the unprocessed images that was selected as the best actress for a description by 20 different participants in Experiment 9. InterFace first separated these images into texture and shape information (middle), and then a weighted averaged face image could be generated (right).

InterFace (Kramer et al., 2017a), a MATLAB application for face image manipulation and analysis, was used to create weighted average face images for each description. Each array photo was first mapped and standardised to create a standardized face image for further processing. To generate the weighted average images, the standardized images were then combined in proportion to number of times they had been selected in Experiment 9. Thus, each weighted average face image summarised 20 choices, as each character in Experiment 9 was cast by 20 participants. Three separate averages (*Full*, *Physical*, and *Character*) for each of the 21 characters (8 female, 13 male) resulted in 63 weighted average images in total. Figure 6.2 shows some examples.

	Type of Description		
	Full	Physical	Character
Female Character 1			
Female Character 2			
Male Character 1			
Male Character 2			

Figure 6.2. Examples of weighted averaged images from Experiment 11. Rows show images generated for two female and two male characters. Columns show variants based on Full, Physical, and Character descriptions.

Design

Each participant rated all 63 weighted average face images (21 *Full*, 21 *Physical*, and 21 *Character*) on three social dimensions (*trustworthiness*, *dominance*, and *attractiveness*), using a 7-point Likert scale. Participants rated each image on all three dimensions before moving on to the next image. Presentation of stimuli and recording of responses were controlled by an Apple Macintosh computer running PsychoPy2 (Peirce, 2007).

Procedure

In each trial, a single average image was displayed on the centre of a 21.5-inch computer screen, with a rating meter from 1 to 7 presented underneath the image, see Figure 6.3 for a task example. Participants were free to view each image for as many long as they wanted. They rated all three social inferences (*trustworthiness*, *dominance*, and *attractiveness*) of one average image by clicking on the rating meter. Each time, one social inference was rated. Social inferences were rated in the order of attractiveness, trustworthiness, and then dominance for all images. The task would only progress when participant gave a response on the current task.



How Attractive is this person?

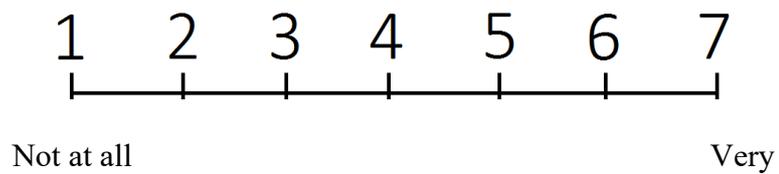


Figure 6.3. An example of the experiment task. Participants inspected the face image presented and then rated the rated on the scale underneath. They responded the attractive of the character above by clicking on the 1 – 7 scale with a mouse, with 1 representing not attractive at all, and 7 meaning very attractive.

Results and Discussion

Social inference ratings of weighted averaged faces

Mean social ratings for the average images are summarised in Table 6.1.

Table 6.1. Means and SDs of social inferences ratings for the weighted averaged face images of full, physical and character descriptions in Experiment 11.

	Social Inferences Rating of Weighted Averaged Face Images (Out of 7)					
	Trustworthiness		Dominance		Attractiveness	
Description Version	Mean	<i>S.D.</i>	Mean	<i>S.D.</i>	Mean	<i>S.D.</i>
Full	4.17	.76	3.78	.75	3.78	0.98
Physical	4.14	.81	3.95	.79	3.81	.96
Character	4.10	.70	3.90	.72	3.70	.86
All	4.14	.67	3.88	.63	3.76	.87

One-way ANOVA found no significant differences among the three information conditions (*Full, Physical, Character*) for *Trustworthiness* [$F(2,174) = .69, p > .05$], *Dominance* [$F(2,174) = .16, p > .05$] and *Attractiveness* [$F(2,174) = .21, p > .05$]. Thus, there was no evidence that the average images made divergent social impressions depending on the type of description on which they were based. Instead, combining photos that were chosen based on a *Full* description, a *Physical* description, or a *Character* description led to weighted averages that were statistically indistinguishable in this analysis. This null result is consistent with the different types of description evoking similar mental images.

Social inferences distances to the *full* descriptions of the characters

The main purpose of this study was to test whether weighted average images would be closer to their descriptions in social inference space, compared with the constituent photographs. Absolute distances from *Full* descriptions of the characters to other representations of the characters are summarised in Figure 6.4.

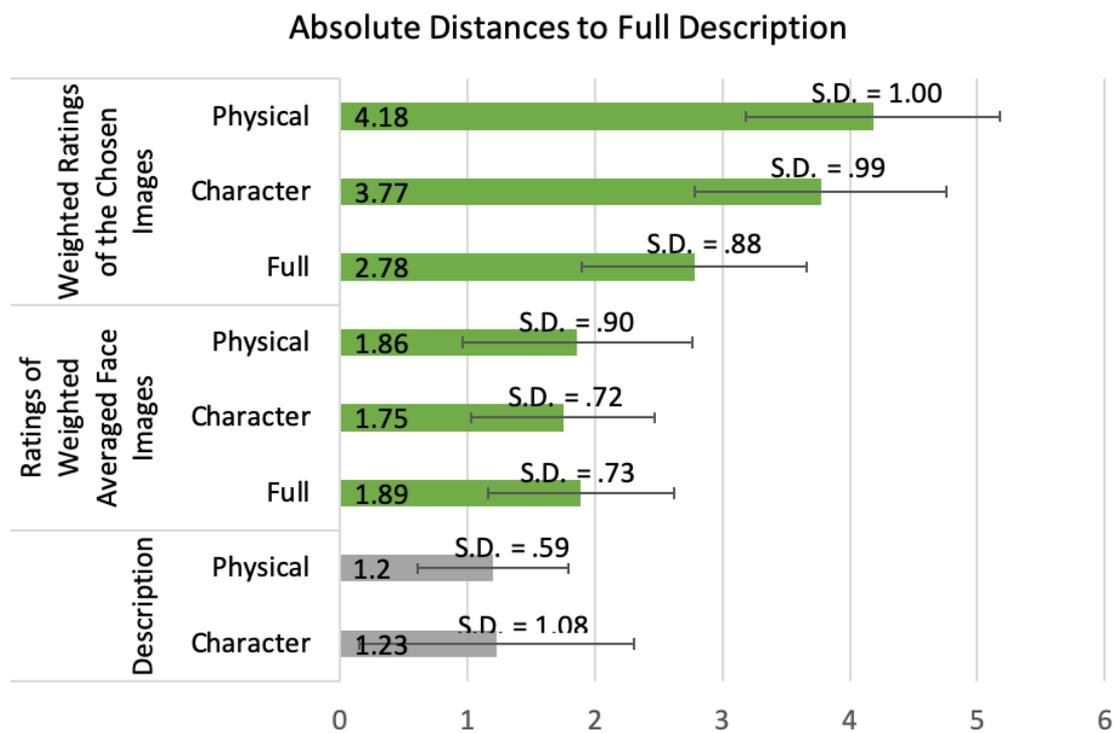


Figure 6.4. The average absolute distances of social inferences from full description to (i) character and physical descriptions, (ii) averaged rating of weighted averaged face images, and (iii) weighted ratings of the chosen images.

Distance data were submitted to a 2 x 3 repeated measures ANOVA with the factors of *Image Type* (*photos* or *average images*) and *Information Type* (*full*, *physical*, or *character*). This analysis found a main effect of *Image Type*, such that the distance from the descriptions was shorter overall for the *average images* ($M = 1.83$, $SD = .77$) than for the *photos* ($M = 3.58$, $SD = 1.11$) [$F(1, 20) = 130.05$, $p < .001$]. There was also a main effect of *Information Type*, the

shortest distances being in the *full* condition ($M = 2.34, SD = .92$), followed by the *character* condition ($M = 2.76, SD = 1.34$), followed by the *physical* condition ($M = 3.02, SD = 1.50$) [$F(2, 40) = 8.16, p < .005$]. The interaction between these factors was also significant [$F(2, 40) = 15.52, p < .001$]. Simple main effects showed that the effect of Information Type was significant in the *photo* condition [$F(2, 80) = 21.82, p < .001$], but not in the *average image* condition [$F(2, 80) = .24, p > .05$]. Effect of Image Type was significant in *full description condition* [$F(1, 60) = 16.53, p < .001$], *character description condition* [$F(1, 60) = 85.65, p < .001$], and *physical description condition* [$F(1, 60) = 112.34, p < .001$].

The main purpose of this social distance analysis was to compare the proximity of weighted average images, versus their constituent photographs, to the locations of the written descriptions upon which the images were based. The key question was whether the averaging process would bring the images closer to the originating descriptions. The analysis shows that this was indeed the case. Weighted average images, constructed from photos that were chosen as good matches to the *physical*, *character*, and *full* descriptions, were closer to the full descriptions than were the weighted ratings of the chosen photos.

The convergence of the average images towards the written descriptions is especially interesting, given the independent sources of the social rating data. The written descriptions, the array photographs, and the average images were all rated by different groups of participants. Moreover, participants' photo choices in the physical and character conditions were based on non-overlapping descriptors. Nevertheless, some information about facial appearance is preserved along these different pathways from written description to social inference. The next experiment builds on this analysis of social inference ratings by testing behavioural predictions.

Experiment 12

The preceding study showed the weighted average images elicited similar social inference ratings to the descriptions upon which they were based. The purpose of the current experiment is to test whether this metric argument is borne out at the behavioural level. The rationale is similar to the casting task in Experiment 10. Once again, participants read written descriptions of fictional characters, and then chose which face image best captures the character's appearance. This time, however, the images were not the photographic images used previously, but the weighted average images that were generated in the preceding study. To reduce the cognitive load on participants, and to encourage full processing of all response options, the array of face images was reduced to just two alternatives. The *target* image was the weighted average image derived from the written description that was on screen. The *distractor* image was a weighted average image derived from one of the other descriptions. As well as reducing the cognitive load on participants, simplifying the task allows for a more straightforward analysis. If a weighted average captures the appearance of the character as described, then participants should choose the *target* image more frequently than one would expect by chance (>50%).

Methods

Participants

Sixty participants (35 female, 25 male; mean age 22.10; age range 18–27), who were naïve to the stimuli and to the original images, were recruited from the University of York in exchange for a small payment or course credit.

Stimuli and Design

All three versions of 21 descriptions from Experiment X were used to communicate the roles being cast. The face images used in this experiment were the 63 weighted averaged images

created in Experiment 11. Two weighted averaged images were presented in each trial: one was the image based on the presented description (the target image), and the other was an image based on the description of another character of the same sex, chosen at random.

Each weighted averaged face thus appeared twice in the experiment: once as the target and once time as the distractor. The target face was equally likely to appear on the left side or the right side of the display. Each participant completed 21 trials (7 descriptions of each version) in a different random order. The measure of interest was the frequency with which the target face was chosen as the best match for the written description.

Procedure

Participants were instructed to imagine that they were casting agents for a theatre company, whose task was to select actors and actresses for particular roles. As the written descriptions could mention superficial aspects of appearance that neither image matched, participants were told that makeup and hair artists could be recruited at a later stage to adjust makeup and hair.

On each trial, participants were asked to read the written description and to form a mental picture of the fictional character. Based on this mental picture, they should choose the actor or actress who fits the role better. To ensure that all participants had sufficient time for reading, there was no time limit for the task. The entire experiment took approximately 30 minutes to complete.

Results and Discussion

Participants chose the target image on 60% of trials, and the distractor image on 40% of trials. A one-sample *t*-test against chance (50%) confirmed that this was a statistically significant difference [$t(59) = 8.62, p < .001$]. The implication of this result is that the target

image captured the appearance of the fictional character better than the distractor face, even though both images depicted faces of the same sex.

Unlike Experiment 10, which compared (same sex) photos that were closest and furthest from the descriptions, this experiment provides a somewhat more stringent test that compared the ‘best’ (the target) against an alternative that was chosen at random (the distractor). The weighted average images presented here were combinations of the best match photographs selected by a separate group of participants. Experiment 11 showed that averaging together the best match photos reduced the distance to the corresponding description in social space. The new finding ties that analysis back to behaviour, by showing that the average images capture something of the facial appearance that is evoked by the written description. Given that the ratings of the descriptions, the ratings of the photos, the ratings of the average images, and the average image casting task were completed by entirely separate groups of participants, this finding implies a degree of convergence between different individuals. The next experiment focuses on convergence between different types of descriptive information.

Experiment 13

The analysis social inference space in Experiment 11 compared the distance of different visual representations to the *full* written description. Averaging together photos that participants selected after reading *physical*, *character*, or *full* descriptions of fictional characters reduced the distance to the *full* written description, relative to averaging the ratings of these selected photos. One interesting aspect of this finding was that the distances were reduced in all three information conditions (*physical*, *character*, and *full*). Whichever type of description the photo choice was based on, averaging together the chosen photos led to a closer approximation. This is perhaps surprising, given that the two types of images were made from images selected by different participants based on different types of information.

It suggests that physical description and character description may be separable channels that can each direct readers to the same region of the space. That idea emerges from the analysis of social inference ratings. However, it is reinforced by visually comparison of the average images across information conditions. Figure 6.5 shows some examples. Each row shows two average images for the same fictional character. The image on the left summarises photos that were chosen by participants who read the *physical* description. The image on the right summarises photos that were chosen by participants who read the *character* description.

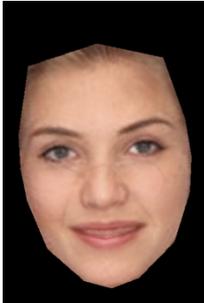
	Type of Description	
	Physical	Character
Female Character 1		
Female Character 2		
Male Character 1		
Male Character 2		

Figure 6.5. Examples of weighted averaged face images based on physical and character versions of descriptions of two female and two male characters.

The proposal is that not only do the paired images look *similar*, in some cases, they look like *the same person*. This experiment tests this proposal experimentally. If *physical* and *character* descriptions converge strongly (that is, on the same facial identity), then averages based on the same fictional character should be seen as the same person more often than averages based on different fictional characters. On the other hand, if *physical* and *character* descriptions converge only weakly, then averages based on the same fictional character should not be seen as the same person (even though they may look similar).

Methods

Participants

Twenty participants (10 female, 10 male; mean age 21.65; age range 19–24), who were naïve to the stimuli, were recruited from the University of York in exchange for a small payment or course credit.

Stimuli and design

Weighted averaged face image pairs The weighted average images created in Experiment 11 were used as stimuli. Two average images—one derived from a *physical* description and one derived from a *character* description—were presented side by side to create a set of face pairs. In match pairs, the two images were based on the same fictional character. In mismatch pairs, the two images were based on different fictional characters of the same sex, chosen at random. Each of the 21 fictional characters appeared in one match pair and one mismatch pair, resulting in 42 pairs in total. Each participant saw all 42 pairs in a random order, and made a “same” or “different” identity judgement for each pair via keypress. The main measure was the proportion of “same” responses in the match and mismatch trials. PsychoPy2 (Peirce, 2007) was used to control stimulus presentation and record participants’ responses.

Procedure

In each experimental trial, participants were instructed to decide if the two images in the pair show the same person or not, see Figure 6.6 for a task example. The 42 trials were presented in a random order. In each trial, a pair of images were displayed on the centre of a 21.5-inch computer screen, with a line of instruction underneath (i.e., are they the same person?).

Participants responded to the question by pressing one key for “same” and another key for “different”. Participants were allowed to spend as much time as they wanted comparing the images. The task would only progress when participant gave a response on the current task. The entire experiment took approximately 20 minutes to complete.



Are they the same person?

Press Z for YES

Press M for NO

Figure 6.6. An example of the experiment task. Participants inspected the two face images presented and responded to the question - ‘are they the same person’ by press the corresponding key on the keyboard – The key at the bottom left - Z for yes, and the key at the bottom right - M for no.

Results and Discussion

A within-subjects t-test revealed that participants made significantly more “same person” responses in the *match* condition ($M = 68\%$; $SD = 0.21$) than in the *mismatch* condition ($M = 38\%$; $SD = 0.19$) [$t(19) = 5.08, p < .001$].

This result implies that the images in match pairs were perceptually more alike than the images in mismatch pairs. In fact, the match images were often so alike that they were perceived as the same person. This is an intriguing finding, given that the *physical* and *character* descriptions of the same character were redacted to present entirely different information. Evidently, images projected from these two versions of descriptions could lead to the same perceived identity. Not only the metric comparison on social inferences distances (Experiment 11), but also the behavioural results of this study suggest that *physical* and *character* information might be two complementary pathways to mental pictures of faces. However, it is unclear whether *physical* or *character* information is more important in shaping imagined appearance. The next experiment compares their relative contributions.

Experiment 14

The metric analysis in Experiment 11 and the results of Experiment 13 suggest that *physical* description and *character* description both inform mental images of fictional faces. However, based on those findings alone, it is not clear whether one type of description dominates impression formation, or whether the two types of description contribute equally. The purpose of this experiment is to compare the contributions of *physical* and *character* information directly by putting them into competition. The participants’ task was to read the *full* description of a fictional character, and then decide which of two images best fits that description (similar to Experiment 12). This time however, the two images were both derived from the same fictional character. Specifically, the weighted average based solely on *physical*

description of that person and the weighted average based solely on *character* description of that person. The distribution of responses over these two options should indicate which image type is a better representation of the full description.

Methods

Participants

Forty participants (22 female, 18 male; mean age 21.13; age range 18–24), who were naïve to the stimuli, were recruited from the University of York in exchange for a small payment or course credit.

Stimuli and Design

Pre-rated Descriptions All 21 *full* descriptions from Experiment 8 were used to communicate roles being cast. Importantly, these descriptions contained all of the original physical and character information without redaction. Two weighted averaged images were also presented on each trial: one was the image based on the physical description, and the other was the image based on the character description. Each trial was single-sided colour-printed on a A4 size paper. Each participant completed all 21 trials in a different random order. The measure of interest was the percentage of trials on which the *physical* average was chosen.

Procedure

The procedure was the same as for Experiment 10, except that the different identity images were replaced with *physical* and *character* averages of the same identity, refer to Figure 6.7 for task example. The participant's task was to read the written description and to form a mental picture of the fictional character. They were then asked to decide which of the two images best captures that mental picture. To ensure that all participants had sufficient time for

reading, there was no time limit for the task. The entire experiment took approximately 30 minutes to complete.

She was very young - at the most nineteen, with a pale somewhat refined face, yellow hair, merry blue eyes, and shining teeth. Her beauty was of an ethereal type. She looked so white and light and fragile that she might have been the spirit of that storm-foam from out of which I plucked her. She had wreathed some of Madge's garments round her in a way which was quaint and not unbecoming.



Actress 1 Actress 2

Figure 6.7. An example of the experiment task. Participants read the paragraph showed at the left and select one actress from the two profile pictures at the right. Their casting decision was recorded by circle the number under the profile picture. For this task, Actress 1 is the average image of images selected based on the character description; Actress 2 is the average image of images selected based on the physical description.

Results and Discussion

Participants chose the *physical* average image on 55% of trials, and the *character* image on 45% of trials. A one-sample *t*-test against chance (50%) confirmed that this was a statistically significant difference [$t(39) = 2.53, p < .05$]. The implication of this result is that the *physical* image captured the appearance of the fictional person somewhat better than the *character* image, even though both images relate to the same identity.

Both *physical* and *character* information contribute to the formation of mental images of faces. However, *physical* information appears to be more dominant in this case. This could be explained by supposing that physical description is inherently more informative of facial appearance. Alternatively, it could be that the particular descriptions used in this experiment contained a greater quantity of physical information, compared with character information (as

the manipulation check in Experiment 8 suggests). Distinguishing between these possibilities would require comparison with a separate corpus of written descriptions that contain a different balance of character and physical information. What is already clear from the present experiment is that the methods developed here are capable of addressing such questions.

General Discussion

The experiments in this chapter were designed to achieve better visual representations of characters in written descriptions by applying an image average technique to selected photographs. The previous chapter examined agreement among participants as to which photographs could and could not match the description. By combining the selected photographs in a weighted average, I created for each description a new image that summarises the visual attributes of the entire set. In Experiment 11, photos selected in the casting task of Experiment 9 were morphed together in proportion to number of times they has been chosen (using InterFace software; Kramer et al., 2017a). The resulting weighted average images were rated for attractiveness, dominance, and trustworthiness by a new group of participants so that they could be projected into the same 3D social space as the rated descriptions and photographs. The key finding of this analysis was that the average images were generally closer to the *full* description of the characters than were the averaged ratings of their constituent photographs. In other words, the average images were better visual representations of the fictional characters—at least in terms of the social impression that they made. Interestingly, this improvement was seen across all three information conditions. Averaging photos that were chosen after reading *physical*, *character*, or *full* descriptions could all shorten the metric distance to the *full* description. This convergence suggests that *physical* and *character* information could provide two complementary pathways to the same visual appearance. In Experiment 12, I modified the two-alternative casting task from

Experiment 10 to compare two different weighted average images—one that was derived from the on-screen written description (the target), and one that was derived from the description of another character of the same sex (the distractor). Participants chose the target image as the better match more often than the distractor image. This behavioural finding builds on the preceding metric analysis, and suggests that the target images captured something of the described appearance that the distractor images did not. Behavioural data also bears out the metric convergence between *physical* and *character* cues to facial appearance. In Experiment 13, participants perceived average images based on *physical* and *character* descriptions of the same fictional character as the same identity. This finding implies remarkable preservation of information about facial appearance—across different types of representation, and across different participants—as the *physical* and *character* averages were created. Not only were *physical* and *character* descriptions in the casting task read by different participants, but also the descriptions and photos that comprise the casting task were rated by different groups of participants. Despite the fact that the *physical* and *character* averages were constructed through these independent routes, participants in Experiment 9 could still bind them by identity.

All of these studies suggest that *physical* description and *character* description both contribute to the reader's impression of facial appearance. This observation raises the question of whether the two types of information are equal partners in shaping mental imagery for faces, or whether one type of information carries more weight than the other. Experiment 14 showed that participants regard an average image based on physical information as better capturing the full written description, compared with an average image based on character information. This finding is consistent with the full descriptions containing more physical information than character information, or with physical

information having a greater impact on the reader's impression. The methods developed in this chapter allow us to explore such issues experimentally.

Chapter 7

General Discussion

In the General Introduction chapter, I argued that face representations for specific identities might be more elastic than previous studies have shown. I also argued that the social impression of a particular identity (quantified by social inference ratings) might provide a common currency by which various form of representations can be compared.

Chapter 2 used a card-sorting task to examine whether faces of fictional characters portrayed in different formats (film captures and comic book drawings) and by different actors and artists could be incorporated into unified identities, and whether standard familiarity effects could be observed for these fictional faces. Performance with both photos and drawings demonstrated a familiarity effect whereby same-identity cards were grouped together more frequently by high-familiarity observers, compared with the low-familiarity observers, while identity merge errors were rare for both subgroups. For the high-familiarity subgroup in particular, characters portrayed by different actors or artists were grouped according to the fictional identity they represented, suggesting that some idiosyncratic aspects of those characters were common to the various representations. These results demonstrate that faces can be learned and cohered by fictional identity, even when portrayed by different actual identities (the actors). In addition, faces can be learned through exposure to impoverished information (comic book drawings) that does not conform to naturalistic exposure. Chapter 3 used Linear Discriminant Analysis (LDA) to examine whether multi-format images (drawings and photos) of the two fictional faces could be grouped by identity based on social inference ratings alone (*Trustworthiness, Dominance, Attractiveness, and Age*). The LDA solution was as accurate as humans who sorted the visual images. Moreover, LDA performance was significantly correlated human performance. These similarities in performance are quite striking, given that the LDA analysis only had access to social inference ratings from humans, and had no direct information related to physical appearance.

For some analyses of face images, it is standard to distinguish between *shape* information and *texture* information. Various manipulations of these attributes have shown that *texture* information seems to be especially important in face identification. However, in natural stimuli, it is hard to make a clean distinction between *shape* and *texture*, particularly as shape affects shading. Comic book drawings of different types provide face stimuli in which *shape* and *texture* information are present in varying degrees. The findings from the card-sorting tasks in Chapter 2 are consistent with previous claims that *texture* is important. In the card-sorting task with comic book drawings, texture information was relatively impoverished. Familiarity effects could still be detected in that task, but they were not as strong as for photographic stimuli. This leads to the question: how much *texture* information is needed to support face learning?

Kramer et al. (2017b) examined the importance of natural variation for learning new faces. Actors' faces could be learned by watching television (natural variation), but learning did not occur when faces were inverted or contrast-reversed (non-natural variation), even though the amount of visual information was equivalent across these conditions. Chapter 2 and 3 raise the possibility that faces could be learned from some forms of non-natural variation, such as drawings, and thereafter show the cognitive hallmarks of familiar face recognition (e.g. image invariance). Demonstrating this conclusively would require a slightly different experimental setup in which the different identities were learned from drawings only, rather than drawings and films (as was the case in Chapters 2 & 3). However, millions of readers apparently have no difficulty tracking dozens of characters through complex story arcs (e.g. Japanese Manga sagas), based on drawings alone. This is a relevant cultural observation, as the readers' ability in this situation bears a strong resemblance to familiar face recognition. Given the abundance of test materials (and test subjects), comics offer an alternative approach to understanding the role of shape and texture information in face processing. They

may be particularly well suited to investigating the quantity and quality of texture information that is required to support face learning.

Chapter 4 investigated further impoverishment of texture information with photograph-based face drawings. Participants were required to compare a target face (either a drawing or a photo) to a line-up of five face drawings below. Experiments 2 and 3 found that performance was better for drawing-to-drawing trials than for photo-to-drawing trials, meaning that within-format identification was easier, even though photos provided more information than drawings. Non-target drawings were more frequently selected in photo-to-drawing trials as a photo could plausibly match several of the available drawings, so that it was hard to be certain that the target was not there. Experiment 6 showed that photo-to-drawing identification accuracy was above chance performance, confirming that some recognizability between the drawing and photo was preserved in the very sparse visual representations. Experiment 11 used a simple paired matching task to eliminate the effect of the foils. For this task, performance in the drawing-to-drawing condition was almost perfect for same person trials, unlike performance in traditional paired matching tasks with photographs, in which errors are common. One possible explanation for the high level of accuracy in this condition is that different drawings of the same face were visually very similar, because they were all derived from the same reference photograph. For different-person trials, accuracy was equivalent in the drawing-to-drawing and photo-to-drawing conditions. In the photo-to-drawing condition, performance on same person trials was not significantly different from chance level, presumably due to the low image similarity between the two images in this condition.

To sum up, drawing-to-drawing comparisons in this task may be easy because they are more like image matching. Although photographic faces vary in their surface properties, environmental lighting conditions, and the camera characteristics (Kramer et al., 2017a), these sources of variation do not apply to simple line drawings, such as the *Beano* drawings in Chapter 4. By the same token, visual comparison between drawings and photos is difficult because the sources of variation in *texture* information are fundamentally different. On this account, viewers resort to using whatever is left (the *shape* or *texture* variation in the face) or extracting something abstract (impression/characteristic) or both to complete the task.

The *Beano* line drawings used in Chapter 4 are arguably a different category of drawing to caricatures and veridical line drawings used in previous studies, in that some *texture* information has been added into the *Beano* drawings (e.g., overall skin tone; red cheeks to indicate blushing). A caricature is ‘a symbol that exaggerates measurements relative to any measure which varies from one person to another’ (Perkins, 1975), and captures (or even exaggerates) the essential characteristics or shape of a face, and therefore, is ‘super-fidelity’ (Rhodes et al., 1987). In the *Beano* line drawings, *shape* was simplified, but a certain amount of *texture* information was added, providing additional cues to match the identities. These drawings were identifiable to viewers to some extent. However, viewers found it hard to distinguish them from drawings of other faces that fit the same general appearance were hard to distinguish. This is perhaps not surprising, as unlike caricatures, these drawings were not designed to systematically exaggerate idiosyncratic features of the individual face.

Drawings styles can be arranged on a scale from highly representational to highly abstract (see Figure 7.1). One possible reason that drawing-to-photo comparisons were difficult in this task is that the drawing style for the *Beano* cartoons was somewhat less representational, compared with the Batman and Spider-Man drawings used in Chapter 2. In general, the

proposal is that the more representational the drawing style (Figure 7.1.A), the more matching will rely on visual comparison. The more abstract the drawing style (Figure 7.1.B), the more matching will rely on impression comparison. The *Beano* drawings in Chapter 4 take a big step in the abstract direction. They do not seek to capture a pattern of light in the world, or the surface properties of the skin. Instead, the drawing ‘stands for’ the person who is depicted in the photograph. In this sense, the drawing may function more like a label. Personal names are another type of label. There is no visual information related to appearance in a name. In matching personal names of an unfamiliar person to photographs, learned association is required (see Figure 7.1.C). If sufficiently simple drawings function more like labels, it may not be surprising that accuracy is low in a perceptual matching task. But drawings differ from nominal labels in that they can preserve visual cues (e.g. distinguishing features) or character cues (e.g. expressions that drive social inferences) (see Figure 7.1.B). These simple drawings are supposed to make viewers think of the right person, but not in the same way as a photographic image. In terms of cognition, are these representations more like face images or more like labels? Healey, Swoboda, Umata and King (2007) found that in graphical communication, graphical complexity across time increases without communication, but decreases with verbal or lexical interaction. That is, drawings become more symbol-like if high-interaction is involved (Garrod, Fay, Lee & MacLeod, 2007).

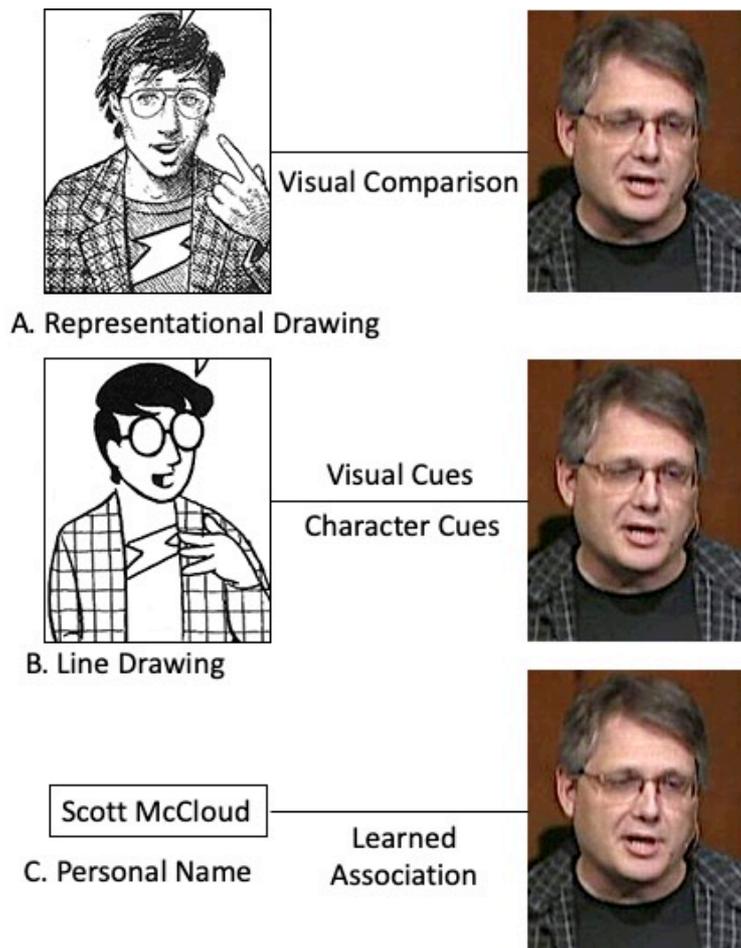


Figure 7.1. The possible connections between photograph and (A) representational drawing, (B) line drawing, and (C) personal name. Images from McCloud (1993).

In communicating identities by drawings, familiarity at both the viewer and the artist levels should be taken into consideration. The drawings from *Beano* were drawn by an artist who was unfamiliar with the subject's face, and the drawings were based on one single reference photograph. The drawings were then viewed by readers who are unfamiliar with the subject's face, and can only compare the drawings to that particular photo that the drawings were based on. If the viewers are familiar with the subjects, would these drawings, created by an unfamiliar artist, look less representative (especially when the reference photo is not representative)? Future studies could answer this question by obtaining likeness ratings between a photograph and a drawing based on the photograph with two dependent variables:

familiarity (familiar/unfamiliar to the target), and viewpoint (artist/viewer). If the likeness rated higher for artists and viewers from same than different familiarity level, it may imply that the artists and viewers from different familiarity level captured different ‘representations’ of the target, which is possibly not inferred from the presented photograph but other previously familiarized representations of the target.

Chapters 5 and 6 focused on a still more visually impoverished form of face representation: verbal descriptions of faces. Face descriptions were presented in *Full*, *Physical*, and *Character* versions, with *Physical* descriptions containing only information about anatomical appearance (e.g. bushy eyebrows), and *Character* descriptions containing only information about non-anatomical attributes (e.g. shifty manner). Social inference ratings (*Trustworthiness*, *Dominance*, and *Attractiveness*) were gathered for these face descriptions (Experiment 8), which allowed metric comparisons with pre-rated face photos from Sutherland et al. (2013). In this way, the image with closest absolute social distance to each description could be located. Participants were then asked to select the best fit actor/actress for each description from an array of face photos in the theatrical casting task (Experiment 9). There was high agreement on the image selected for each description. Moreover, selected images were generally closer to the description in social inference space, compared with non-selected images. In direct comparison, images with closest social distances were considered a better representation than images with furthest social distances (Experiment 10). Weighted average images for each description were created by morphing chosen images with weighting in accordance to the number of times each image was selected in Experiment 9. These weighted averaged images were closer to the full description than weighted ratings of the chosen images (Experiment 11). Even though images were selected by different participants who read non-overlapping information about facial appearance (*physical* or *character* descriptions), by morphing the selected face images, the absolute social distances between

the character being described and the face image could be significantly shortened. Moreover, the weighted averaged images were behaviourally considered better representations of their corresponding description, compared to weighted averaged images for a different description, chosen at random (Experiment 12). The perceptual likeness of weighted averaged images from reading *Physical* and *Character* descriptions of the same character were high (Experiment 13), suggesting that these two different types of description provide convergent information. However, average images derived from *Physical* information were more frequently selected as a better representation of the character being described in the *Full* description (Experiment 14). This pattern suggests that *Physical* and *Character* information provide two complementary pathways to the same appearance, and that *Physical* information is more important than *Character* information in the translation between verbal and visual representation of faces.

Because photographic images and written descriptions of faces contain fundamentally different types of information (pictorial and verbal respectively), translation between them must be mediated by an amodal representation. This level of representation could correspond to the *Person Identity Nodes (PINs)* level in the *Interactive Activation and Competition (IAC) Model* of face recognition (Burton et al., 1990; Burton et al., 1991; Burton & Bruce, 1993; Bruce et al., 1992a). An adapted IAC model with faces of fictional identities in different formats as input is shown Figure 7.2. Faces of fictional characters (either from comic books or film adaptations) from the *Input* are connected to the corresponding *Face Recognition Units (FRUs)*. In the adapted model, the FRUs are format-independent as well as being view-independent, so that recognizable face representations in both photos and drawings will activate the appropriate FRU.

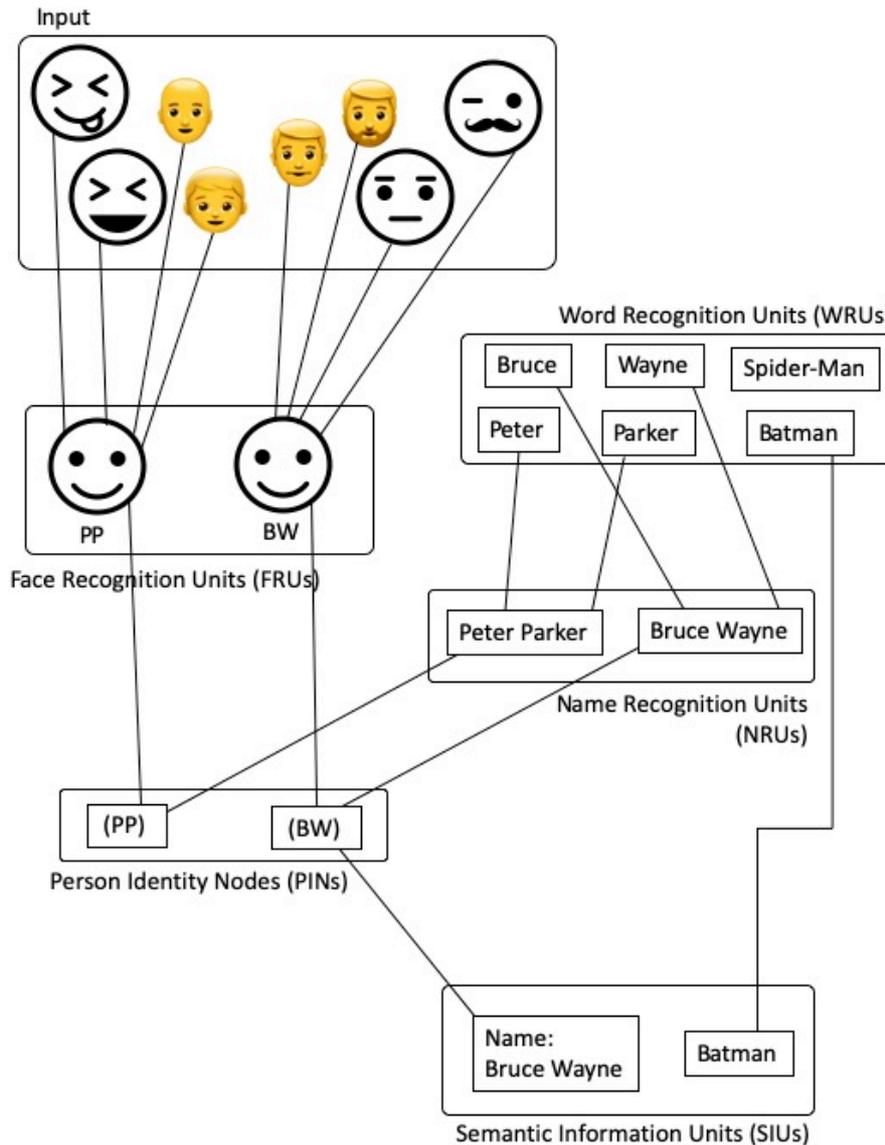


Figure 7.2. Adapted Interactive Activation and Competition (IAC) Model of Face Recognition (Burton et al., 1990; Burton et al., 1991; Burton & Bruce, 1993; Bruce et al., 1992a) with Faces of Fictional Identities in Different Formats as Input.

Input of the IAC model can also be verbal descriptions of faces, and directly connect to the WRUs instead the FRUs (see Figure 7.3). In this situation, there are verbal descriptions that directly describe physical facial appearance in WRUs (e.g. brown hair, and dark face) and can be connected to the face images in FRUs. SIUs contain descriptions which are neither names nor physical appearance (e.g. handsome face, manly jaw). For verbal descriptions used in

Chapter 5 & 6, these descriptions are mainly character descriptions of the faces. One major difference to the IAC model with image input (*Figure 7.2*) is that the semantic information from verbal description can be inferred from unknown faces. In this scheme, the *PINs* (which are amodal) mediate the connection of words and images. From Chapter 6, we know that different participants show some agreement on which face is best described by a given description, and which faces definitely do not fit the description. This would imply that there are some common activation patterns for the same input among different readers.

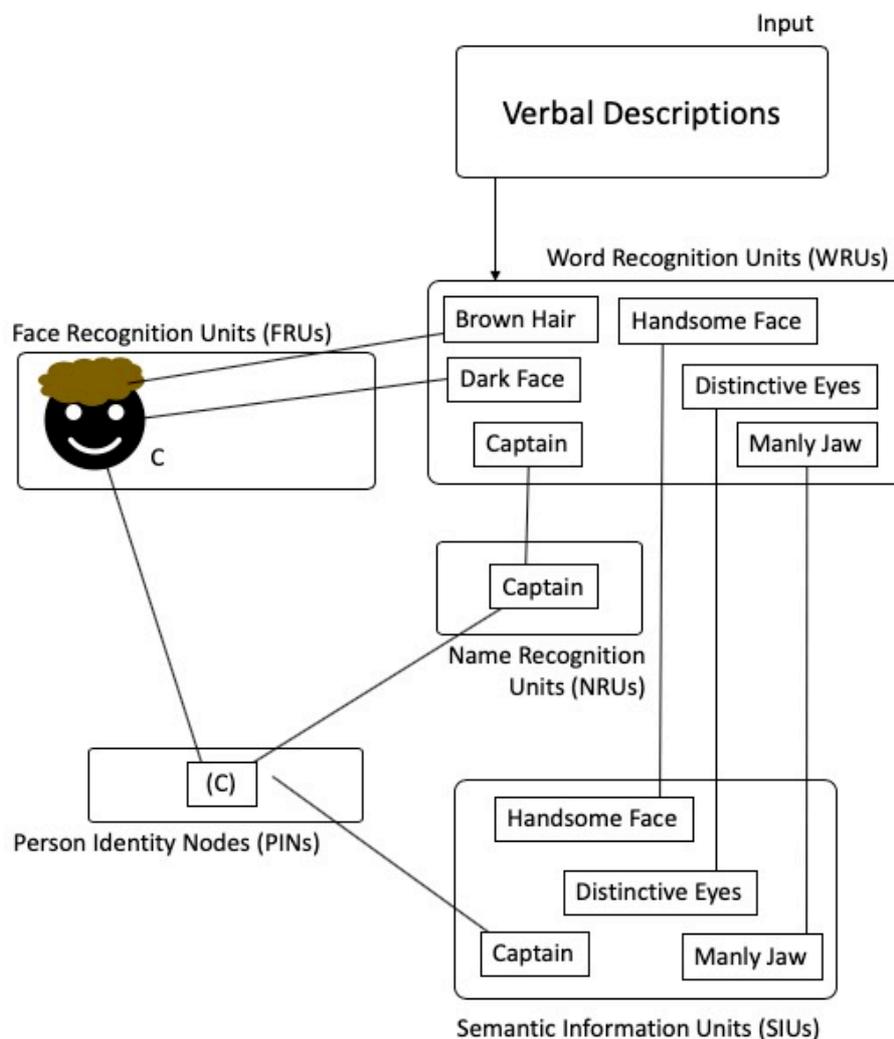


Figure 7.3. Adapted Interactive Activation and Competition (IAC) Model of Face Recognition (Burton et al., 1990; Burton et al., 1991; Burton & Bruce, 1993; Bruce et al., 1992a) with Verbal Descriptions as Input.

The PIN is an abstract level of representation that is hard to measure directly. Social inference ratings are useful because they allow measurement at the same level of abstraction, thus providing a connection between words and images (Figure 7.4). The framework of Todorov et al. (2008) offers a means to quantify and compare mental representations of faces, by putting them into a common dimensional space constructed from social inferences. This thesis extends Todorov’s framework to written descriptions of faces. Face images that were selected to represent a description were generally closer to the description in this social inference space.

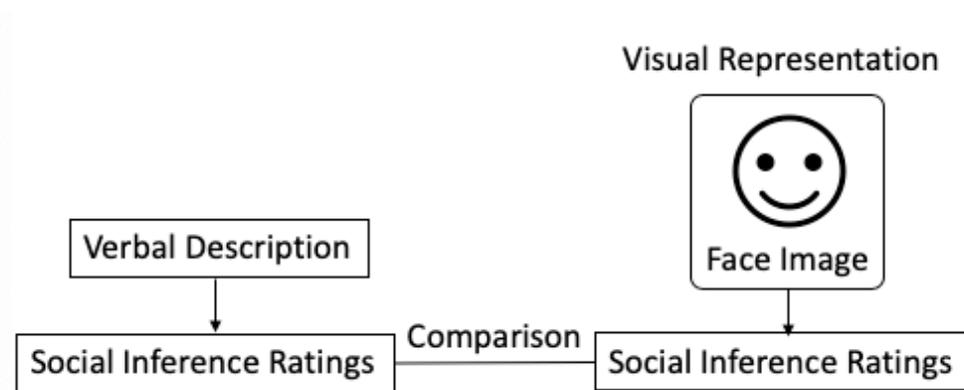


Figure 7.4. Comparison between Verbal Description and Face Image are made at a Measurable Social Inference Ratings Level.

Whether comparisons between written descriptions and visual images of faces are made at the level of mental pictures inferred from verbal description (Figure 7.5.A) or at level of abstracted impressions (Figure 7.5.B) is not yet clear. Either mechanism could give rise to the observed performance. One possible way to distinguish between these possibilities would be to conduct a version of the casting experiment with aphantasic participants (i.e. people who do not experience mental imagery; Keogh & Pearson, 2018; Zeman, Dewar & Della Sala, 2015; Zeman, Dewar & Della Sala, 2016). If aphantasic participants show a similar pattern of behaviour to other participants, this would imply that the task can be completed without

comparing the physical image to a mental image, consistent with comparison at a more abstracted level (Figure 7.5.B). Alternatively, if aphantasic participants show a very different pattern of behaviour, this could suggest that mental imagery is critical to the task, consistent with pictorial comparison (Figure 7.5.A).

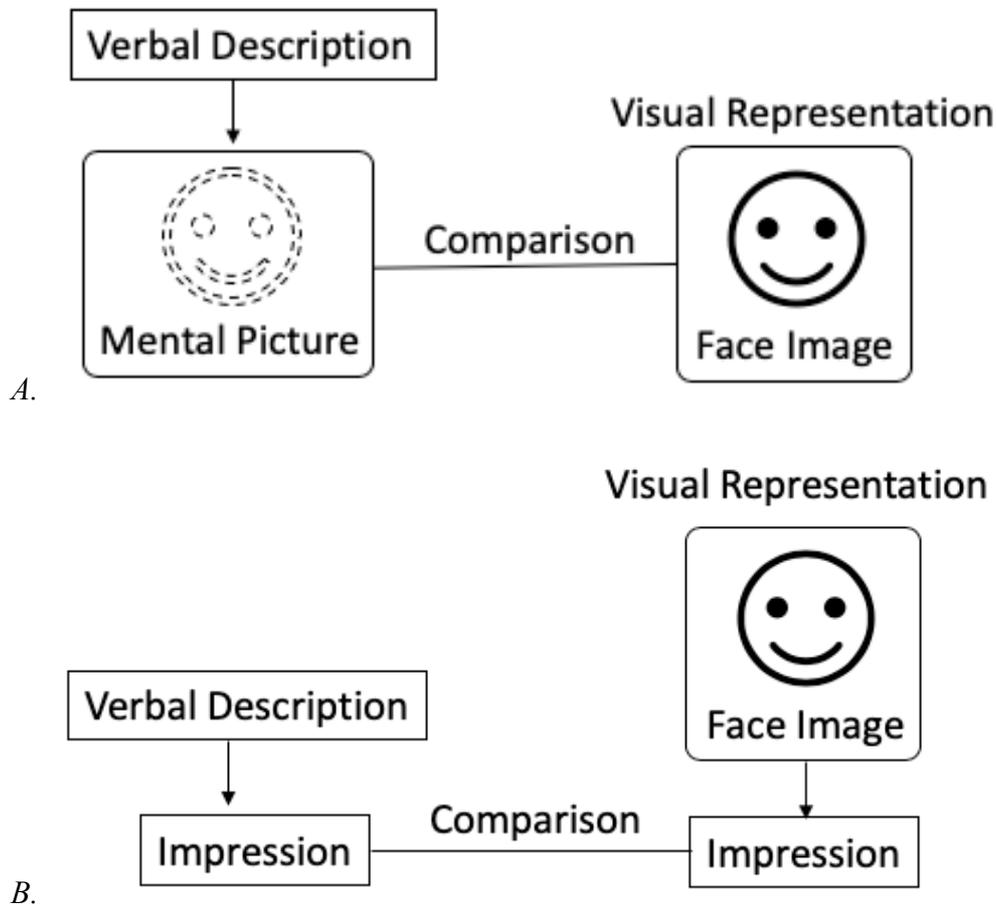


Figure 7.5. Comparison between Verbal Description and Face Image are made at a Level of (A) Mental Picture Inferred from Verbal Description, and (B) Impression.

The findings from Chapters 5–7 suggest that verbal descriptions of both *physical* and *character* aspects of faces could contribute to identifying a face. Morphing faces selected by different viewers who read either *physical* or *character* description lead to similar looking average face images, and these images gave rise to social inference ratings (*Trustworthiness*, *Dominance*, and *Attractiveness*) that were similar to ratings from the *full* description of that

face (Chapter 7). These findings suggest that social impressions from faces could complement physical information in helping identify an individual. The need to translate verbal descriptions of faces into images is a sometimes arises in forensic settings (e.g. in eyewitness testimony). Drawings of criminal suspects are usually informed by eyewitness descriptions, which are heavily based on physical appearance. Similarly, facial composite techniques (e.g. Photofits) are usually based on facial features (e.g., eyes, noses, mouths) that to the operator selects to piece together the suspect's appearance (Brandl, 2014). The findings of this thesis suggest that *character* aspects of face descriptions could be usefully introduced as a complementary pathway towards creating a composite. Alongside photo-to-photo matching, description-to-photo confidence ratings could be used in parallel. If this combined method can improve overall identification accuracy, relative to a purely pictorial approach, this would imply some independent (non-pictorial) factor behind the description-to-photo route. In future studies, it would be interesting to explore the effects of these different modes of comparison for faces that look different but generate similar social impressions; and conversely faces that look similar but generate different social impressions. Those are the cases in which complementary information is likely to have the biggest impact.

All of the findings in this thesis may be explained by assuming that the face representation for a particular identity might be more elastic than was previously thought. Radically different visual depictions of a face can be tied to a single identity, even when the visual information is sparse (as with line drawings). Specifically, movie stills and comic book drawings of fictional characters could be sorted together by identity (Chapter 3). Highly simplified face drawings, based on a reference photograph, could be associated with the original photograph despite the low visual similarity between these image types (Chapter 4). Moreover, different types of face representation can be associated even in the absence of pictorial information. Facial appearance could be constructed from reading a verbal

description of a face, and different readers of the same text showed substantial agreement about the inferred appearance (Chapter 5).

In sum, this thesis brings new insight into the possible identifiable representations of one identity, and emphasises that face representations can be more elastic than previous studies have revealed, encompassing photographic images, different types of drawings, and verbal descriptions of both physical and character information. These different forms of face representation can be quantified, compared, and connected via social inferences. Social inference space provides a simple framework for measuring different face representations in common terms. Moreover, both physical and character descriptions of faces could contribute to identification, as these two types of information tend to converge. Visual information of physical appearance is not the only pathway to face recognition. As well as enriching our psychological understanding of face processing, the current thesis also forms a bridge between the scientific study of faces, and the portrayal of faces in the arts.

References

- Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*(6), 1043-1047.
- Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *The Quarterly Journal of Experimental Psychology*, *68*(10), 2041-2050.
- Andrews, T. J., Baseler, H., Jenkins, R., Burton, A. M., & Young, A. W. (2016). Contributions of feature shapes and surface cues to the recognition and neural representation of facial identity. *Cortex*, *83*, 289-291.
- Arad, A., Tolmach, M. (Producers), & Webb, M. (Director). (2014). *The Amazing Spider-Man 2* [Motion Picture]. United States: Sony Pictures Releasing.
- Arad, A., Tolmach, M., Ziskin, L. (Producers), & Webb, M. (Director). (2012). *The Amazing Spider-Man* [Motion Picture]. United States: Sony Pictures Releasing.
- Beano Studio. (2018). Make Me A Menace - Beano. Retrieved from <https://www.beano.com/categories/make-me-a-menace>.
- Becker, D. V., Kenrick, D. T., Neuberg, S. L., Blackwell, K. C., & Smith, D. M. (2007). The confounded nature of angry men and happy women. *Journal of personality and social psychology*, *92*(2), 179.
- Bindemann, M., Burton, A. M., & Jenkins, R. (2005). Capacity limits for face processing. *Cognition*, *98*(2), 177-197.
- Bindemann, M., Jenkins, R., & Burton, A. M. (2007). A bottleneck in face identification: Repetition priming from flanker images. *Experimental Psychology*, *54*(3), 192-201.
- Bonner, L., & Burton, A. M. (2004). 7–11-year-old children show an advantage for matching and recognizing the internal features of familiar faces: Evidence against a

- developmental shift. *The Quarterly Journal of Experimental Psychology Section A*, 57(6), 1019-1029.
- Brandl, S. G. (2014). *Criminal Investigation 3*. Thousand Oaks, CA: Sage Publications.
- Brown, C., & Lloyd-Jones, T. J. (2005). Verbal facilitation of face recognition. *Memory & Cognition*, 33(8), 1442-1456.
- Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, 73(1), 105-116.
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77(3), 305-327.
- Bruce, V., Burton, A. M., & Craw, I. (1992a). Modelling face recognition. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273), 121-128.
- Bruce, V., Hanna, E., Dench, N., Healey, P., & Burton, M. (1992b). The importance of 'mass' in line drawings of faces. *Applied Cognitive Psychology*, 6(7), 619-628.
- Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, 5(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., Ness, H., Hancock, P. J., Newman, C., & Rarity, J. (2002). Four heads are better than one: Combining face composites yields improvements in face likeness. *Journal of Applied Psychology*, 87(5), 894.
- Burton, A. M., & Bruce, V. (1993). Naming faces and naming names: Exploring an interactive activation model of person recognition. *Memory*, 1(4), 457-480.

- Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*(3), 361-380.
- Burton, A. M., Jenkins, R., Hancock, P. J., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*(3), 256-284.
- Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223.
- Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243-248.
- Burton, A. M., Young, A. W., Bruce, V., Johnston, R. A., & Ellis, A. W. (1991). Understanding covert recognition. *Cognition*, *39*(2), 129-166.
- Burton, A., Bruce, V., & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception*, *22*, 153-176.
- Conway, G., Kane, G., & Romita Sr., J. (2002). *Spider-Man: Death of the Stacy's*. New York, NY: Marvel Comics.
- Davies, G., Ellis, H. D., & Shepherd, J. (1978). Face recognition accuracy as a function of mode of representation. *Journal of Applied Psychology*, *61*, 180-187.
- Di Novi, D., Burton, T. (Producers), & Burton, T. (Director). (1992). *Batman Returns* [Motion Picture]. United States: Warner Bros. Pictures.
- Dodson, C. S., Johnson, M. K., & Schooler, J. W. (1997). The verbal overshadowing effect: Why descriptions impair face recognition. *Memory & Cognition*, *25*(2), 129-139.
- Doyle, A. C. (1890). *The captain of the Polestar, and other tales*. London: Longmans, Green & Co.

- Doyle, A. C. (1895). *The Strank Munro letters: Being a series of twelve letters written by J. Stark Munro, M. B., to his friend and former fellow-student, Herbert Swanborough ... During the years 1881-1884*. London: Longmans, Green & Co.
- Doyle, A. C. (1998). *The hound of the Baskervilles*. Oxford: Oxford University Press.
- Doyle, A. C. (1999). *The return of Sherlock Holmes*. Oxford: Oxford University Press.
- Doyle, A. C. (2001a). *The lost world and other thrilling tales*. London: Penguin.
- Doyle, A. C. (2001b). *The sign of the four*. London: Penguin.
- Doyle, A. C. (2004). *The parasite*. BookSurge.
- Doyle, A. C. (2007). *Sir Nigel and the white company*. Tucson: Fireship Press.
- Doyle, A. C. (2009). *The case-book of Sherlock Holmes*. Oxford: Oxford University Press.
- Etemad, K., & Chellappa, R. (1997). Discriminant analysis for recognition of human face images. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, 14, 1724–1733.
- Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7), 450-451.
- Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research*, 20(2), 226-241.
- Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31(6), 961-987.
- Hancock, P. J., Bruce, V., & Burton, M. A. (1998). A comparison of two computer-based face identification systems with human perceptions of faces. *Vision Research*, 38(15-16), 2277-2288.
- Hancock, P. J., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition*, 24(1), 26-40.

- Healey, P. G., Swoboda, N., Umata, I., & King, J. (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science*, *31*(2), 285-309.
- Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *15*(4), 445-464.
- Hess, U., Adams, R. B., Grammer, K., & Kleck, R. E. (2009). Face gender and emotion expression: Are angry women more like men?. *Journal of Vision*, *9*(12), 19-19.
- Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception*, *31*(10), 1221-1240.
- Hugenberg, K., Young, S. G., Sacco, D. F., & Bernstein, M. J. (2011). Social categorization influences face perception and face memory. *Oxford Handbook of Face Perception*, 245.
- Ishai, A., Ungerleider, L. G., & Haxby, J. V. (2000). Distributed neural systems for the generation of visual images. *Neuron*, *28*(3), 979-990.
- Jenkins, R., & Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science*, *319*(5862), 435-435.
- Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323.
- Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *11*(3), 211-222.
- Keogh, R., & Pearson, J. (2018). The blind mind: No sensory visual imagery in aphantasia. *Cortex*, *105*, 53-60.

- Kleinberg, K. F., Vanezis, P., & Burton, A. M. (2007). Failure of anthropometry as a facial identification technique using high-quality photographs. *Journal of Forensic Sciences*, 52(4), 779-783.
- Kosslyn, S. M., Alpert, N. M., Thompson, W. L., Maljkovic, V., Weise, C. F., Chabris, C. F., . . . Buonnano, F. S. (1993). Visual mental imagery activates topographically organized visual cortex: PET investigations. *Journal of Cognitive Neuroscience*, 5, 263-287.
- Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, 2(9), 635-642.
- Kosslyn, S. M., Sukel, K. E., & Bly, B. M. (1999). Squinting with the mind's eye: Effects of stimulus resolution on imaginal and perceptual comparisons. *Memory & Cognition*, 27(2), 276-287.
- Kosslyn, S. M., Thompson, W. L., & Alpert, N. M. (1997). Neural systems shared by visual imagery and visual perception: A positron emission tomography study. *Neuroimage*, 6(4), 320-334.
- Kramer, R. S., Jenkins, R., & Burton, A. M. (2017a). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior research methods*, 49(6), 2002-2011.
- Kramer, R. S., Jenkins, R., Young, A. W., & Burton, A. M. (2017b). Natural variability is essential to learning new faces. *Visual Cognition*, 25(4-6), 470-476.
- Kramer, R. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*, 172, 46-58.
- Kramer, R. S., Young, A. W., Day, M. G., & Burton, A. M. (2017c). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review*, 124(2), 115-129.

- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.
- Laub, C., & Bornstein, B. H. (2008). Juries and eyewitnesses. In Cutler B. L. (Ed.), *Encyclopedia of Psychology and Law (Vol. 1)*. California: USA: Sage Publications, Inc.
- Lee, S., Kane, G., & Romita, J. (2007). *Spider-Man: Death of the Stacy's*. New York, NY: Marvel Comics.
- Martinez, A. M., & Zhu, M. (2005). Where are linear feature extraction methods applicable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1934–1944.
- McCloud, S. (1993). *Understanding comics: The invisible art*. New York: Harper Collins Publishing.
- McCloud, S. (2011). *Making comics*. New York: Harper Collins Publishing.
- Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, 34(4), 865-876.
- Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, 69(7), 1175-1184.
- Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, 15, 603-616.
- Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, 41(3), 577-581.
- Neil, L., Cappagli, G., Karaminis, T., Jenkins, R., & Pellicano, E. (2016). Recognizing the same face in different contexts: Testing within-person face recognition in typical development and in autism. *Journal of Experimental Child Psychology*, 143, 139-153.

- O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience*, *12*(6), 1013-1023.
- O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329-337.
- O'Neil, D., García-López, J. L., & Braun, R. (2012). *Batman: Venom*. Burbank, CA: DC Comics.
- O'Toole, A. J., Price, T., Vetter, T., Bartlett, J. C., & Blanz, V. (1999). 3D shape and 2D surface textures of human faces: the role of “averages” in attractiveness and age. *Image and Vision Computing*, *18*(1), 9-19.
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092.
- Osborne, C. D., & Stevenage, S. V. (2008). Internal feature saliency as a marker of familiarity and configural processing. *Visual Cognition*, *16*(1), 23-43.
- Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8-13.
- Perkins, D. (1975). A definition of caricature and caricature and recognition. *Studies in Visual Communication*, *2*(1), 1-24.
- Peters, J., Guber, P. (Producers), & Burton, T. (Director). (1989). *Batman* [Motion Picture]. United States: Warner Bros. Pictures.
- Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, *19*(4), 473-497.

- Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, 70(5), 897-905.
- Robertson, D. J., Kramer, R. S., & Burton, A. M. (2015). Face averages enhance user recognition for smartphone security. *Plos One*, 10(3), e0119460.
- Rossion, B., & Michel, C. (2011). An experience-based holistic account of the other-race face effect. *Oxford Handbook of Face Perception*, 215-244.
- Roven, C., Thomas, E., Franco, L. (Producers), & Nolan, C. (Director). (2005). *Batman Begins* [Motion Picture]. United States & United Kingdom: Warner Bros. Pictures.
- Santos, I. M., & Young, A. W. (2008). Effects of inversion and negation on social inferences from faces. *Perception*, 37(7), 1061-1078.
- Santos, I. M., & Young, A. W. (2011). Inferring social attributes from different face regions: Evidence for holistic processing. *The Quarterly Journal of Experimental Psychology*, 64(4), 751-766.
- Santos, I., & Young, A. (2005). Exploring the perception of social characteristics in faces using the isolation effect. *Visual Cognition*, 12(1), 213-247.
- Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1), 36-71.
- Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257-268.
- Snyder, S., Capullo, G., & Albuquerque, R. (2013). *Batman: City of Owls*. Burbank, CA: DC Comics.
- Stafford, O. (2016, 11 22). *Jim Jarmusch: 'The spirit of punk is even more valuable now than ever'*. Retrieved from Little White Lies: <https://lwlies.com/interviews/jim-jarmusch-paterson-adam-driver/>

- Sutherland, C. A., Oldmeadow, J. A., Santo, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*, 105-108.
- Sutherland, C. A., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology*, *108*(2), 397-415.
- Thomas, E., Nolan, C., Roven, C. (Producers), & Nolan, C. (Director). (2012). *The Dark Knight Rises* [Motion Picture]. United States & United Kingdom: Warner Bros. Pictures.
- Thomas, E., Roven, C., Nolan, C. (Producers), & Nolan, C. (Director). (2008). *The Dark Knight* [Motion Picture]. United States & United Kingdom: Warner Bros. Pictures.
- Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences*, *1124*(1), 208-224.
- 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.
- Turow, J. (1978). Casting for TV Parts: The Anatomy of Social Typing. *Journal of Communication*, *28*(4), 18-24.
- Wein, L., & Andru, R. (2008). *Spider-Man: A New Goblin*. New York, NY: Marvel Comics.
- Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, *54*(1), 277-295.
- Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592-598.

- White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, *21*(1), 100-106.
- Yan, X., Andrews, T. J., Jenkins, R., & Young, A. W. (2016). Cross-cultural differences and similarities underlying other-race effects for facial identity and expression. *The Quarterly Journal of Experimental Psychology*, *69*(7), 1247-1254.
- Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14*(6), 737-746.
- Zebrowitz, L. A., & Montepare, J. M. (2008a). First impressions from facial appearance cues. *First Impressions*, 171-204.
- Zebrowitz, L. A., & Montepare, J. M. (2008b). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*(3), 1497-1517.
- Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin*, *28*(2), 238-249.
- Zeman, A. Z., Dewar, M., & Della Sala, S. (2015). Lives without imagery-Congenital aphantasia. *Cortex*, *73*, 378-380.
- Zeman, A., Dewar, M., & Della Sala, S. (2016). Reflections on aphantasia. *Cortex*, *74*, 336-337.
- Ziskin, L., Arad, A. (Producers), & Raimi, S. (Director). (2004). *Spider-Man 2* [Motion Picture]. United States: Sony Pictures Releasing.
- Ziskin, L., Arad, A., Curtis, G. (Producers), & Raimi, S. (Director). (2007). *Spider-Man 3* [Motion Picture]. United States: Sony Pictures Releasing.
- Ziskin, L., Bryce, I. (Producers), & Raimi, S. (Director). (2002). *Spider-Man* [Motion Picture]. United States: Sony Pictures Releasing.