SOCIAL PERCEPTION OF FACES: NEUROPSYCHOLOGICAL AND BEHAVIOURAL INVESTIGATIONS

By

Constantin Rezlescu

March 2013

Thesis submitted to University College London
for the degree of Doctor of Philosophy in Cognitive Science

I, Constantin Rezlescu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:

Table of Contents

Abstract

This thesis concerns theoretical and empirical issues in face processing and facial trait perception. First, I present evidence that challenges two hypotheses proposed as alternatives to face specificity, namely the individuation and the expertise hypotheses. Inconsistent with the individuation hypothesis, an extensive investigation of a new case of acquired prosopagnosia (Herschel) revealed normal exemplar recognition memory for a wide variety of objects, and normal ability to discriminate between highly similar items within a novel object category. Inconsistent with the expertise hypothesis, Herschel and Florence, a second acquired prosopagnosic, showed normal learning profiles and response times putative of successful expertise acquisition in an eight-day training procedure with novel objects, demonstrating that faces are processed by specialised mechanisms not used for objects-of-expertise. Second, testing four patients with acquired prosopagnosia, I demonstrate that perceptual mechanisms underlying trait judgments are dissociable from those implicated in recognising identity. Furthermore, I show that perception of facial aggressiveness does not depend on mechanisms for facial sex recognition, and that normal facial trustworthiness judgments are likely to occur without intact recognition of facial expressions, therefore challenging the overgeneralisation theory in facial trait perception. Third, I present a series of experiments with healthy participants to characterise various properties of facial trait perception. Specifically, I examine: i) the role of facial width-to-height ratio in perceived trustworthiness; ii) the accuracy of facial trustworthiness judgments; iii) the interaction between facial trustworthiness and reputation; and iv) the interaction between face impressions and voice impressions. Overall, the findings of the present thesis have important implications for the nature of the mechanisms underlying facial identity processing, the organisation of facial trait perception and its relationship to other face perception abilities, as well as the physical, ecological, and multimodal aspects of facial trait perception.

List of figures

List of tables

Publications and presentations arising from thesis

Some of the work contained in this thesis has already been published or presented at conferences, or is in the course of preparation for publication.

Published papers:

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE*, 7(3), e34293.

Rezlescu, C., Pitcher, D., & Duchaine, B. (2012). Acquired prosopagnosia with spared within-class object recognition but impaired recognition of degraded basic-level objects. *Cognitive Neuropsychology*, 29(4), 325-347.

Manuscripts submitted for publication:

Rezlescu, C., Susilo, T., Barton, J. J. S., & Duchaine, B. (2013). Normal facial trait judgments without facial identity recognition.

Rezlescu, C., Barton, J., Pitcher, D., & Duchaine, B. (2013). Normal acquisition of expertise with a novel object class in two cases of acquired prosopagnosia.

Rezlescu, C., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. (2013). Dominant voices and attractive faces: the contribution of visual and auditory information to integrated person impressions.

Manuscripts in preparation:

Rezlescu, C., Susilo, T., Barton, J., & Duchaine, B. (2013). Dissociations of facial trait, facial expression, and facial sex processing in acquired prosopagnosia: Evidence against the overgeneralisation hypotheses.

Rezlescu, C., Heerey, E., Chater, N., & Olivola, C. (2013). Cheater detection based on faces in high-stake real-life situations.

Oral presentations:

A face you can trust? (2011). Talk presented at the Social Behavior Workshop, November 9th, Dartmouth College, Hanover, USA.

Facial physiognomy affects strategic behaviour in trust games (2011). Talk presented at the Subjective Probability, Utility and Decision Making conference (SPUDM 23), August 21-25, London, UK.

The deceptive lure of facial trustworthiness: Trustworthy faces are beneficial for owners but largely uninformative for observers (2010). Talk presented in the Social Cognition and Social Neuroscience Lab (Prof. Alex Todorov), Princeton University, May 26th, Princeton, USA.

Poster presentations:

Rezlescu, C., Barton, J. J. S., Pitcher, D., & Duchaine, B. (2013) Even you, greebles? Normal greeble performance in acquired prosopagnosia supports face specificity. Vision Science Society Conference, May 10-15, Naples, Florida.

Rezlescu, C., Susilo, T., Barton, J., & Duchaine, B. (2012) Dissociations of facial trait, facial expression, and facial sex processing in acquired prosopagnosia. UCL Neuroscience Symposium, June 29, London.

Rezlescu, C., Susilo, T., Barton, J., & Duchaine, B. (2012) Social perception of faces in acquired prosopagnosia. Vision Science Society Conference, May 11-16, Naples, Florida.

Duchaine, B., Rezlescu, C., Pitcher, D. & Witty, N. (2011) Acquired prosopagnosia with spared within-class object recognition but impaired recognition of degraded basic-level objects. Vision Science Society Conference, May 6-11, Naples, Florida.

Rezlescu, C., Duchaine, B., & Chater, N. (2010) A face you can trust? When faces affect behavior in trust games. Association for Psychological Science Conference, May 27-30, Boston.

Acknowledgments

First, I would like to thank my supervisors Nick Chater and Vincent Walsh for excellent support and guidance. Nick's brilliance, dedication and modesty were an inspiration. Vince's excellence in everything he does, his invaluable expertise and constant encouragement helped me tremendously. He may not know it, but many of his snap words of wisdom about science and life stayed with me. I am also grateful to Brad Duchaine, who introduced me to face perception, prosopagnosia and American football, and continued to provide close support and advice throughout my PhD. The three months I spent in his lab at Dartmouth College were a fantastic experience. I feel extremely lucky to have worked with all three of them.

Second, I would also like to extend a warm thanks to the people who generously shared some of their time and knowledge with me: Adam Harris, Chris Olivola, Chris Street, Crystal Goh, Daniel Richardson, David Lagnado, David Pitcher, Erin Heerey, Hua Yang, Ivo Vlaev, Joel Winston, Kirsten Dalrymple, Lili Tcheang, Lucia Garrido, Michael Banissy, Neil Muggleton, Sophie Scott, Stian Reimers, Thomas Dytie. Special thanks go to Tirta Susilo with whom I have worked closely on a number of projects and who provided helpful comments on earlier versions of this thesis.

Third, a huge embrace to all former members of the Chater Lab, we were a fantastic group! Thank you Christos Bechlivanidis, Konstantinos Tsetsos, Erica Yu, Irma Kurniawan, Petter Johansson, Ramsey Raafat and Tobias Gerstenberg, for the great time we had together.

I am deeply grateful to all prosopagnosic and healthy individuals for their participation and enthusiasm. I thank Lucia Garrido, Nicole Whitty, Kasia Borowska and Hikaru Tsujimura for help with running some of the tests. I thank Dave Perrett, Michael Stirrat and Alexander Todorov for sharing some of their stimuli. I thank Hannah Tickle for proofreading and Tommaso Romano for his Photoshop expertise and valuable feedback.

I was financially supported by a generous PhD fellowship from the Axa Research Fund.

This work is dedicated to my parents.

Chapter 1
INTRODUCTION

1.0. PREAMBLE

The human face is probably the richest social stimulus. From faces, people infer such diverse information as one's identity, race, sex, age, emotions or intentions. Furthermore, based on faces, people derive consistent evaluations of others' attractiveness, trustworthiness or aggressiveness, evaluations which influence their social interactions. It is thus not surprising that researchers from various disciplines have devoted considerable intellectual energy in their quest to understand the many aspects of face perception. Although significant progress has been made during the past decades, important questions remain open. The current work addresses several key issues.

First, are the mechanisms performing face recognition specific to faces? This concerns a larger question in cognitive psychology and neuroscience, of whether the brain has high-level mechanisms specialised in processing particular types of information. In his influential and controversial theory of modularity, Fodor (1983) suggested that many cognitive functions are carried out by domain-specific modules. In visual recognition, face-specific effects found in behavioural, imaging, patient and animal studies (McKone & Robbins, 2011) point towards distinct, highly specialised mechanisms not involved in recognition of other object types. This account is by no means generally accepted. In this thesis I will focus on two of the most prominent alternative hypotheses – one that claims faces are processed by general mechanisms implicated in within-class discrimination of all objects (Damasio, Damasio & Van Hoesen, 1982) and one arguing that apparent face-specific mechanisms are also involved in processing non-face objects for which one has acquired expertise (Diamond & Carey, 1986). An extensive investigation of a new case of acquired prosopagnosia and an expertise training exercise with two individuals with acquired prosopagnosia will test these two hypotheses.

Second, faces are not only important for identity or emotion recognition, but commonly lead to (not necessarily accurate) impressions about personality traits or behavioural propensities. However, face impressions have been largely overlooked by leading models of face perception (Bruce & Young, 1986; Haxby, Hoffman, & Gobbini, 2000) and their place within the face processing system remains relatively unknown. The two models cited here propose that face computations are performed along dimensions mirroring our semantic organisation of face-based judgments, for instance dissociating perception of invariant facial aspects (needed for identity recognition, but also other inferences) from perception of changeable aspects (such as gaze and expressions). An interesting question is whether these models can be modified to accommodate face impressions. For example, should face impressions be included among invariant or changeable facial aspects? Is trait perception related to identity or expression or sex perception? Or is it a category by itself? A more specific hypothesis about the functional bases of face evaluations has been recently proposed (Oosterhof & Todorov, 2008), but its predictions have not been critically tested. Investigations of four individuals with acquired prosopagnosia will reveal associations and dissociations between facial trait perception and perception of identity, expression and sex, allowing more precise inferences about the mechanisms behind trait perception.

Third, inferences about one's personality or behaviour based on their faces may be deemed unethical, but they were found to have a major impact on society, such as predicting electoral outcomes (Todorov, Mandisodza, Goren, & Hall, 2005) or influencing financial decisions (Ravina, 2008; Sydnor & Pope, 2008). Among the various traits inferred from faces, trustworthiness is one of the most important because of the crucial role that trust plays in social and economic interactions (Arrow, 1973; Fukuyama, 1995). Face trustworthiness impressions form fast (Willis & Todorov, 2006) and are consistent across observers (Todorov, Said, & Verosky, 2011), but little is known about facial aspects that are commonly associated with trustworthiness. Is width-to-height ratio one of them (Stirrat & Perrett, 2010)? Furthermore, there is a debate around the accuracy of face trustworthiness perceptions, i.e. whether they can predict deceptive (or untrustworthy) behaviour. Most of the previous studies featured laboratory stimuli and generated mixed results. Could it be that people are better detectors of deception from faces in real-world situations? In the laboratory, face trustworthiness was shown to influence investment decisions when there was no information about potential partners (van't Wout & Sanfey, 2008), but real economic interactions typically involve partners who know a great deal about each other. In this context, would perceived face trustworthiness still matter? Finally, social interactions usually involve verbal communication. Are trait impressions also formed from voices? How do face and voice impressions combine to reach an integrated person impression? A series of behavioural experiments will address the questions raised here.

Accordingly, this thesis is organised into five chapters. Chapter 1 provides an overview of research on face specificity and alternative hypotheses, a brief introduction to leading models of face perception with a focus on how they incorporate (or not) trait perception, and face impressions with a focus on trustworthiness and the issues of interest. The following three chapters present experimental work. Chapter 2 concerns face specificity. More specifically, I test the two alternative hypotheses to face specificity using acquired prosopagnosia. One extensive case study focuses on the individuation hypothesis, and an expertise training exercise with two acquired prosopagnosics provides key findings related to the expertise hypothesis. In Chapter 3 I use four cases of acquired prosopagnosia to examine if facial trait perception can dissociate from identity recognition and to test the emotion and sex overgeneralisation hypotheses in trait perception. Chapter 4 presents behavioural studies of healthy population on face impressions. I examine one facial aspect thought to influence trustworthiness perception, the accuracy of face trustworthiness judgments, and how face impressions interact with reputation and voice impressions. Chapter 5 summarises the findings and considers their implications.

1.1. FACE SPECIFICITY

The debate on face specificity is often presented as revolving around the question of whether or not faces are special to the brain. This is slightly misleading, because researchers generally agree on this point: the visual mechanisms involved in face recognition are thought to be different from those involved in most other objects recognition. The critical, unresolved question is *just how special* are faces? Are the 'face mechanisms' specific to faces, or would objects sharing critical properties with faces also engage them? The attempts to answer this question have produced considerable research, often with conflicting results and interpretations.

1.1.1. Evidence supporting face specificity

Advocates of face specificity cite evidence from developmental, behavioural, brain imaging and stimulation, single-cell recording and patient studies to support their view. Fantz (1963) and Goren, Sarty and Wu (1975) showed that new-born babies (many tested less than 10 minutes after birth) preferred looking at face patterns above non-face patterns. Because this preference toward face stimuli could not have been learned in new-borns who were seeing faces for the first time, the results suggest that humans possess an innate representation of faces (Morton & Johnson, 1991). If that were true, faces would distinguish themselves from most other objects for which evolution could not have determined the development of innate mechanisms. Twin studies (Polk, Park, Smith & Park, 2007) and studies of family members with developmental prosopagnosia, a lifelong inability to recognise faces (Duchaine, Germine & Nakayama, 2007; Grueter et al., 2007), support the evolutionary basis of face-specific mechanisms.

The behavioural evidence for face-specificity relates to qualitative differences noted between processing of faces and processing other objects in the adult normal population. These differences are attributed to a specific property of face mechanisms: holistic processing, which can be defined as the ability to integrate individual features into a perceptual whole (Tanaka & Farah, 1993). In contrast, non-face objects are thought to be processed in a more part-based manner. One of the most robust effects in face perception is the disproportionate inversion effect (Yin, 1969). When people are presented with inverted (i.e. upside-down) stimuli, recognition of faces is much more affected than recognition of other objects. This is because holistic processing is presumed to be preferentially implicated in processing upright faces and less so in inverted faces or upright and inverted objects. The robustness of the face inversion effect made it become an acid test that newly developed face perception tests have to pass to demonstrate the involvement of face-specific mechanisms. More direct evidence of holistic processing of faces comes from part-whole effects (Tanaka & Farah, 1993) and composite effects (Young, Hellawell, & Hay, 1987). In the part-whole paradigm, participants are better at identifying individual face parts in the context of a whole face than when presented alone. In the composite faces paradigm, participants find it more difficult to identify top halves of faces when perfectly aligned with bottom halves to form a seemingly new face, than when the two halves are misaligned. The new faces created by aligning top and bottom halves engage holistic processing, which makes irrelevant bottom halves interfere with recognition of top halves.

Cognitive neuroscience furnished good evidence for face-specific mechanisms in the brain by identifying a distributed network of brain areas that are preferentially activated by faces. The first area that was identified and that is most reliably found in imaging studies is the fusiform face area (FFA) (Kanwisher, McDermott, & Chun, 1997; McCarthy, Puce, Gore, & Allison, 1997). The activity in FFA has been shown not only to be higher when viewing faces of any kind (e.g. human, cat or cartoon faces; Tong, Nakayama, Moscovitch, Weinrib, & Kanwisher, 2000) than objects or other body parts, but also to correlate with face recognition abilities (Grill-Spector, Knouf, & Kanwisher, 2004; Yovel & Kanwisher, 2005) suggesting that at least part of the computations necessary to individuate faces take place in this area. Other areas implicated in face perception identified by imaging studies are the occipital face area (OFA) (Gauthier, Tarr, et al., 2000) and a face-selective part in the superior temporal sulcus (fSTS) (Allison, Puce, & McCarthy, 2000; Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011). Transcranial magnetic stimulation of the OFA revealed its causal role in face processing (Pitcher, Walsh, Yovel, & Duchaine, 2007) and that it is specifically activated by faces and not by objects or bodies (Pitcher, Charles, Devlin, Walsh, & Duchaine, 2009). Consistent with the imaging evidence from humans, single cell recordings from face-selective regions in monkeys' temporal cortex revealed an impressive number of face-responsive neurons in these areas (Tsao, Freiwald, Tootell, & Livingstone, 2006).

Another indicator of face-selective mechanisms is the N170 component in human studies measuring event-related brain potentials. The N170 represents the higher electrical activity elicited by faces around 170 ms after stimulus presentation at occipitotemporal electrodes (Bentin, Allison, Puce, Perez, & McCarthy, 1996; Eimer, 1998). Although there are competing hypotheses regarding the source of this enhanced negative amplitude, being linked with increased activity in the OFA (Rossion, Joyce, Cottrell, & Tarr, 2003), FFA (Henson et al., 2007) or STS (Itier & Taylor, 2004), the N170 is reliably found in most electrophysiology studies. Its high sensitivity to face inversion (Bentin et al., 1996; Eimer, 2000) is additional support that it is related to face-selective mechanisms.

More evidence for face-specific mechanisms comes from individuals with acquired prosopagnosia (AP) (Bodamer, 1947). Individuals with AP experience severe faces recognition deficits following brain lesion. Although most acquired prosopagnosics (APs) have problems with object identification as well (which is to be expected given that, usually, brain lesions due to strokes or accidents have a low specificity) (Barton, 2008; Boutsen & Humphreys, 2002; Delvenne, Seron, Coyette, & Rossion, 2004; Gauthier, Behrmann, & Tarr, 1999; Levine & Calvanio, 1989; Steeves et al., 2006), there are also APs who appear to have retained their ability to correctly recognise non-face objects. Busigny, Joubert, Felician, Ceccaldi, and Rossion (2010) summarised 14 such cases. Unfortunately, object recognition has been rigorously tested in only a few cases. One such case is PS who, despite her severe prosopagnosia, could identify objects at a basic level (Rossion, Caldara, et al., 2003) and was normal at discriminating between highly similar novel shapes and between exemplars of common object classes - such as cars, dogs, cups, shoes - parametrically manipulated for similarity (Busigny, Graf, Mayer, & Rossion, 2010). Similarly, prosopagnosic patient GG performed in the normal range for within-category discrimination of birds, boats, cars, and chairs (Busigny, Joubert, et al., 2010) and of cars parametrically manipulated for similarity (Busigny & Rossion, 2010). Another case, WJ, became a farmer after the stroke that left him prosopagnosic, and could accurately identify his own sheep and learn unfamiliar sheep from face photographs (McNeil & Warrington, 1993). The reverse pattern was shown by CK: he could accurately identify faces even under the most challenging conditions while being object agnosic (Moscovitch, Winocur, & Behrmann, 1997). His results together with the normal performance at basic object recognition and fine-grained discrimination of non-face objects or nonhuman faces displayed by PS, GG and WJ show a double dissociation between mechanisms implicated in human face processing and those involved in other objects processing.

In one of the most comprehensive studies addressing the issue of face-specificity in a single individual, Duchaine, Yovel, Butterworth and Nakayama (2006) tested Edward, a developmental prosopagnosic, for all major alternative hypotheses to face specificity. The results did not fit any of these alternative hypotheses (discussed below), favouring instead the existence of face-specific mechanisms.

1.1.2. Alternative hypotheses to face specificity

Despite these suggestive results, some researchers question the existence of face-specific mechanisms, claiming that face effects are due to particular critical properties of faces rather than to faces per se. Importantly, these properties are presumed not to be exclusive to faces; effects may appear face-specific solely because performance with faces has been compared to performance with objects that did not have these properties. A proper comparison between faces and objects sharing the critical properties would reveal no difference.

The main alternative hypotheses to face-specificity are based on two observations: i) faces are a highly homogeneous class of objects, for which exemplar discrimination is more challenging than most other objects recognition; and ii) due to faces' pervasiveness and importance in daily life, humans have become experts at discriminating faces.

1.1.2.1. Individuation hypothesis

It has been argued that, because face recognition usually occurs at an individual level (Jean's face versus Pat's face), it should be compared with exemplar recognition within non-face object classes (Jean's car versus Pat's car) if we are to make any statements about the special treatment of faces (Damasio et al., 1982). Basic level object recognition (a car versus a bicycle) is generally considered to be easier and thus, a potential difference in recognition abilities with faces versus basic objects may simply reflect a difference in cognitive demands rather than face specificity. In one of the first papers to note this point, Damasio et al. (1982) observed that three prosopagnosic patients were also impaired at different food and car recognition and concluded that the visual recognition inability in prosopagnosia is not specific to faces, but extends to other objects when the proper within-class discrimination task is used.

If the mechanisms involved in face recognition and recognition of non-face objects at the individual level are the same, one would expect to observe similar behavioural and neural effects. However, the behavioural effects thought to accompany holistic/configural processing – inversion, part-in-whole and composite effects – appear to be disproportionately large for faces compared to other objects (see McKone & Robbins, 2011, for a review). For example, the inversion effect for faces is usually 15-20%, while for several objects it was found to vary between 0-10% in tests of memory and perceptual matching of within-class exemplars. In theory, it is still possible that certain objects elicit inversion effects similar to faces, although the wide range of objects tested so far make this possibility somehow unlikely. The tested objects include airplanes (Yin, 1969), shoes (De Gelder, Bachoud-Lévi, & Degos, 1998), houses (Boutet & Faubert, 2006; Leder & Carbon, 2006), chairs (Boutet & Faubert, 2006), animals such as cats, dogs and birds (Minnebusch, Suchan, & Daum, 2009), and also dog faces (Scapinello & Yarmey, 1970). The two exceptions – non-face objects where a face-like inversion effect was noted – came from an artificial class of stimuli called greebles (Ashworth, Vuong, Rossion, & Tarr, 2008) and from bodies (Reed, Stone, Bozova, & Tanaka, 2003). Greebles were designed to place similar demands on the recognition mechanisms as faces and their configuration resembled that of faces, with two identifying features aligned horizontally (like eyes in faces) and two identifying features aligned vertically (like nose and mouth in faces). Because of this close resemblance, the larger inversion effect with greebles might have been driven by the faulty engagement of face mechanisms (McKone & Kanwisher, 2005). Indeed, Ge, Wang, McCleery and Lee (2006) showed that the inversion effect depends on whether an ambiguous stimulus is perceived as a face or not. Similarly, some authors (Minnebusch et al., 2009; Yovel, Pelc, & Lubetzky, 2010) claimed that the face-like inversion effects with bodies are triggered by face mechanisms. The inversion effects with bodies will be discussed in more details in the next section, as they pertain more to the expertise hypothesis.

Most of the imaging, brain stimulation and ERP studies presented in the face-specificity section speak against the within-level discrimination hypothesis. In these studies, face individuation was shown to dissociate from individuation of other objects from within highly homogeneous categories (e.g. Kanwisher et al., 1997; Pitcher et al., 2009; Rossion et al., 2000). A controversial study (Thierry, Martin, Downing, & Pegna, 2007) claimed that the apparent face-sensitivity of the N170 component was due to an uncontrolled difference in the interstimulus perceptual variance between face stimuli (low variance) and other objects (high variance). However, several researchers (Bentin et al., 2007; Rossion & Jacques, 2008) demonstrated this claim to be false, strengthening the original support given by the N170 component to a face specific account.

Individuals with prosopagnosia showing normal recognition of exemplars within highly homogeneous object classes have the potential to furnish decisive evidence against the within-level discrimination hypothesis. However, as we have seen in the section on the face-specificity hypothesis, there are only a few cases rigorously and extensively tested for within-class object individuation and among them, as Busigny, Joubert, et al., (2010) note, none that can be called a “pure” prosopagnosic, with visual recognition impairments that can be unquestionably classified as face-specific. Impaired face recognition with spared individuation of various non-face objects may arise from different processing strategies (e.g. holistic versus part-based) rather than face-specific mechanisms; this dissociation in performance may not replicate with other objects processed in the same manner as faces. Therefore, a rigorous investigation of each new case of acquired prosopagnosia with within-level discrimination of a large set of objects sharing as many critical aspects with faces as possible (e.g. within-class similarity, consistent first-order configurations between parts) is important. Although, practically, one may never ensure an individual is normal with all objects individuation, a rigorous approach minimises the chance of wrong inferences. Section 2.1 presents an extensive investigation (21 experiments) of a new case of acquired prosopagnosia, with the main focus on recognition and individuation of non-face objects.

1.1.2.2. Expertise hypothesis

The other alternative to face-specificity that I present here is the expertise hypothesis (Bukach, Gauthier, & Tarr, 2006; Diamond & Carey, 1986; Gauthier & Tarr, 1997). This is probably the most widely considered alternative and it states that the mechanisms involved in face processing are not specific to faces but contribute to individuation of all objects with which observers have had extensive experience. It can be considered a refinement of the within-class discrimination hypothesis, restricting the range of objects on which mechanisms implicated in face processing operate to objects for which people have developed expertise. Two lines of research can be distinguished inside the expertise account, one that considers real-world expertise developed over years of experience (Diamond & Carey, 1986) and another one that examines ‘rapid expertise’ acquired in the laboratory over a few hours of practice (Gauthier & Tarr, 1997).

The first proponents of the expertise hypothesis, Diamond and Carey (1986) compared inversion effects for faces with inversion effects for dogs shown by dog experts. Their prediction was that dog expertise encourages configural processing of dogs and thus dog experts (but not novices) would show inversion effects with dogs, and these effects would be similar to those typically found for faces. Indeed, with dog pictures, dog experts did show significantly larger inversion effects (22%) than novices (2%), and these effects were comparable with those found with faces (23%). However, Robbins & McKone (2007) failed to replicate these findings. In one dog memory and one dog matching test, dog experts did not show significantly larger inversion effects (7% and 2% respectively) compared to novices (3% and 1% respectively). Furthermore, the dog inversion effects were substantially smaller than those obtained for faces (23% and 11%, respectively). Similar results, inconsistent with the expertise hypothesis predictions, were obtained in studies examining inversion effects in experts of other object categories: birds (Gauthier, Skudlarski, Gore, & Anderson, 2000), fingerprints (Busey & Vanderkolk, 2005), cars (Gauthier, Skudlarski, et al., 2000; Xu, Liu, & Kanwisher, 2005), houses (Husk, Bennett, & Sekuler, 2007) and handwriting (Bruyer & Crispeels, 1992). Reviewing studies investigating behavioural markers of face-like mechanisms (the inversion effect, the whole-over-part advantage, the composite effect), McKone, Kanwisher and Duchaine (2007) concluded that the early study of Diamond and Carey (1986) is the only one to show unequivocally the hallmarks of face processing in objects of expertise.

Recently, another category of objects has emerged to elicit face-like inversion effects. Bodies are in many respects similar to faces; they are biological stimuli with a consistent first-order configuration of their parts (torso, arms and legs). The level of exposure and the need for individuation make it reasonable to assume that humans developed perceptual expertise with bodies approaching that acquired with faces. The face-like inversion effects with bodies (Reed et al., 2003; Yovel et al., 2010; Susilo, Yovel, Barton, & Duchaine, 2013) may be interpreted as evidence that bodies are processed by the same expertise mechanisms involved also in processing faces. Some researchers suggested that the inversion effects with bodies are triggered by erroneous recruitment of face-specific mechanisms, because they observed that when headless bodies are presented, the inversion effect is significantly reduced (Minnebusch et al., 2009; Yovel et al., 2010). While this claim is weakened by findings from Susilo et al. (2013) who showed that face-specific mechanisms are not required for a body inversion effect (four APs showed face-like inversion effects with full bodies, including faceless heads), a large body inversion effect does not necessarily imply shared mechanisms between body and face processing. In fact, acquired prosopagnosics who show body inversion effects (Susilo et al., 2013) are strong evidence for distinct mechanisms.

Comparable behavioural markers in processing faces and other objects of expertise are not conclusive evidence for shared perceptual mechanisms. For example, configural processing may be as important for bodies as it is for faces. If one assumes that configural processing is what is disrupted by stimuli inversion, the sizes of the inversions effects for bodies and faces will be similar, even when the two stimuli are processed by distinct mechanisms. Additionally, challenging the idea that faces and bodies share the same mechanisms, possibly specialised for objects of expertise, functional neuroimaging studies found that bodies preferentially activate areas that may be adjacent to, but do not overlap face-selective areas. These are the fusiform body area (Peelen & Downing, 2005; Schwarzlose, Baker, & Kanwisher, 2005) and the extrastriate body area (Downing, Jiang, Shuman, & Kanwisher, 2001).

Functional imaging also produced evidence claimed to support the expertise hypothesis. Gauthier, Skudlarski, et al. (2000) found increased activation in the fusiform face area (FFA) for cars and birds in cars and birds experts but not in control participants. The percent signal change (PSC) correlated with the behavioural expertise of the subjects, but it was smaller than the PSC obtained when faces were presented. Additionally, the effect was not specific to FFA but extended beyond it. McKone and Kanwisher (2005) suggested the increased activation might be due to an attentional confound (experts tend to pay more attention to their objects of expertise). More recently, behavioural expertise was shown to correlate with FFA activation to cars in a high-resolution fMRI study (McGugin, Gatenby, Gore, & Gauthier, 2012), although in the right FFA the effect was not specific to face-selective voxels.

Testing the expertise hypothesis is hampered by the difficulty to find real objects of expertise that are properly matched to faces in terms of critical aspects. To alleviate this problem, expertise researchers turned to expertise developed in the laboratory ('rapid expertise') and trained participants to become experts with greebles, an artificial class of objects designed to place face-like demands on recognition mechanisms (Gauthier, Williams, Tarr, & Tanaka, 1998; Gauthier & Tarr, 1997). Greebles have a body (which can vary in shape) and four distinguishing parts varying in size and shape, in a stable first-order configuration (i.e. fixed spatial relations between the parts) designed to replicate the facial parts configuration. Body shape signals family membership, while the four parts are used to identify greebles at the individual level. The training procedure involves seven to ten one-hour sessions of learning to identify individual greebles as well as greeble families, after which participants should reach the criterion claimed to signal expertise: response times for recognizing individual greebles become comparable to response times for recognizing greeble families (Gauthier & Tarr, 1997). The main advantages of studying greeble experts instead of real-world experts are that expertise is developed fast and that it allows testing of participants before and after becoming experts.

Much of the evidence cited in support of the expertise hypothesis comes from experiments involving greebles. The basic claim in greeble studies is that the training procedure determines a shift in how participants process greebles, from feature-based to more configural processing as they become experts (Gauthier & Tarr, 1997; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999). Several studies claimed that greeble processing elicited face-like perceptual (Gauthier & Tarr, 1997; Gauthier, Tarr, et al., 1999; Tarr, 2003) and neural (Gauthier, Skudlarski, et al., 2000) effects after but not before training. This implies the effects are related to expertise and not face-specific.

Some researchers (e.g. McKone & Kanwisher, 2005) expressed doubts with respect to using greebles for examining the expertise hypothesis. Their main concerns were that; i) greebles are too similar to faces and thus may activate the face-specific mechanisms only because the brain spuriously interprets them as faces; ii) the ten-hour laboratory practice cannot produce expertise that is comparable with real-world expertise developed over years of practice. Indeed, a recent study showed that FFA activation to greebles is correlated with how much face-like subjects think greebles are, not with the amount of greeble training involved (Brants, Wagemans, & Op de Beeck, 2011). Nevertheless, greebles constitute an important part of the research used to support the expertise hypothesis and thus a careful examination of the claims based on this experimental paradigm is needed.

A fundamental prediction of the rapid expertise hypothesis is that individuals with severe face recognition deficits should also be impaired at acquiring expertise with greebles. This prediction was tested in Section 2.2 in two cases of acquired prosopagnosia, Herschel and Florence, who were asked to complete the greeble training procedure and a similar training procedure with faces. It should be noted that Herschel shows abnormal functioning of the fusiform face area (FFA), an area selectively responsive to faces but which has been claimed to also mediate greeble expertise (Gauthier, Tarr, et al., 1999; Gauthier & Tarr, 2002), and so his results are doubly relevant to the expertise account. Two individuals with acquired prosopagnosia were previously reported to fail at acquiring greeble expertise (Behrmann, Marotta, Gauthier, Tarr, & McKeeff, 2005; Bukach et al., 2012), consistent with the expertise hypothesis. Another individual with developmental prosopagnosia performed normally in greeble training (Duchaine, Dingle, Butterworth, & Nakayama, 2004), but was not tested with faces on a parallel face training task to show a clear dissociation between greeble and face performance.

1.2. TRAIT PERCEPTION IN MODELS OF FACE PROCESSING

Faces are rapidly judged on multiple traits critical for human interactions, such as attractiveness, trustworthiness and aggressiveness (Bar, Neta, & Linz, 2006; Willis & Todorov, 2006). Although these judgments can predict important real-world outcomes (Hamermesh, 2012; Olivola & Todorov, 2010a; Zebrowitz & McDonald, 1991), little is known about their underlying perceptual, cognitive and neural bases. Furthermore, even though face evaluations are obviously derived from facial cues, leading models of face perception (e.g. Haxby et al., 2000) are not very specific about mechanisms underlying trait perception. My aim is to investigate the (dis)connections between facial trait perception and perception of other facial aspects, such as identity, expressions and sex, through the study of people with acquired prosopagnosia. The hypothesis is that the mechanisms involved in facial trait perception are (at least partly) distinct and may be spared when visual recognition deficits affect other aspects of face perception.

Before discussing some of the most influential studies related to this topic, a brief note about terminology. In what follows, terms such as 'facial trait perception', 'face impression', 'face-based judgment' or 'face evaluation' will be used interchangeably, to refer to the spontaneous inferences about an individual's perceived trustworthiness, attractiveness, aggressiveness, etc. that take place when seeing a face. Although the terms are not strictly speaking equivalent, they will be treated in this way to avoid repetitions.

1.2.1. Models of face perception

Divisions between mechanisms implicated in processing different aspects of faces are featured in most face perception models, although the nature and extent of these divisions may differ. Probably the most influential model that has stimulated much of the research in this area for the past 25 years is that of Bruce and Young (1986) (see Figure 1.1). They proposed a functional model of face perception that begins with a structural encoding phase, in which view-centred descriptions and expression-independent descriptions are generated. Two parallel routes then follow this initial stage. One route is dedicated to processing the abstract, expression-independent information leading to person identification through face recognition units. The other route is responsible for processing all other aspects of face perception that are not identity-specific, including separate functional components for facial speech analysis, expression analysis and 'directed visual processing'. The last component is thought to be responsible for information about age, sex, gaze direction and person impressions such as honesty and intelligence. The model has a highly modular nature; according to Bruce and Young (1986), the functional components used to implement this architecture can be independently affected either by experimental manipulation or brain lesions. As specified by this model, face impressions dissociate from identity recognition after the very first stage of structural encoding, and may also dissociate from expression recognition at a later stage. Although Bruce and Young (1986) do not reject the possibility of further divisions (i.e. functionally-independent components) within the 'directed visual processing' module, face impressions are implied as being carried out by the same mechanisms implicated also in age and sex perception.

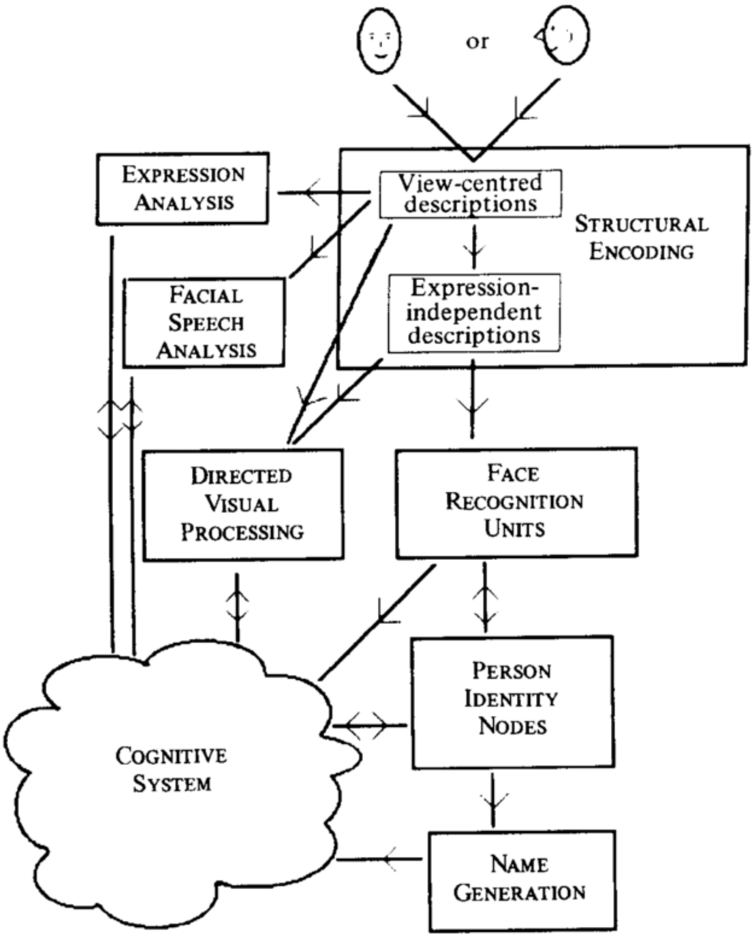


Figure 1.1. The functional model of face processing - Bruce and Young (1986).

The dual-process model proposed by Bruce and Young (1986) was extended and refined by Haxby et al. (2000), who were in addition concerned with specifying a neural implementation for their model. Instead of the distinction between identity recognition and all other aspects of face processing proposed by Bruce and Young (1986), the division in Haxby model is between invariant aspects of faces (mainly related to identity, but which potentially include also sex and age perception) and changeable aspects of faces, such as expression, eye gaze and lip movement. The model has a core system responsible for the visual analysis of faces, implemented in face-selective regions in the occipitotemporal visual extrastriate cortex, and an extended system, where information is further processed by functional components involved in other cognitive functions (Figure 1.2). The core face system has an initial stage responsible for the early perception of facial features (similar to the structural encoding phase from Bruce and Young model) from where information is split into a route dedicated to processing invariant facial aspects (implemented in the lateral fusiform gyrus - FG) and another one for processing variable facial aspects (implemented in the superior temporal sulcus - STS). There is no indication of computing face impressions in the core system. Information from STS is relayed to the extended system for analysing spatial attention (in the intraparietal sulcus), speech perception (in the auditory cortex) and emotion (in the amygdala, insula and limbic lobe). Information from FG reaches the anterior temporal lobe, part of the extended system, where identity and semantic information about a person are activated. There is no explicit mentioning of face impressions in the extended system either, but by elimination one may infer that these are performed in the extended system component responsible for emotion (based on the information about changeable facial aspects received from STS).

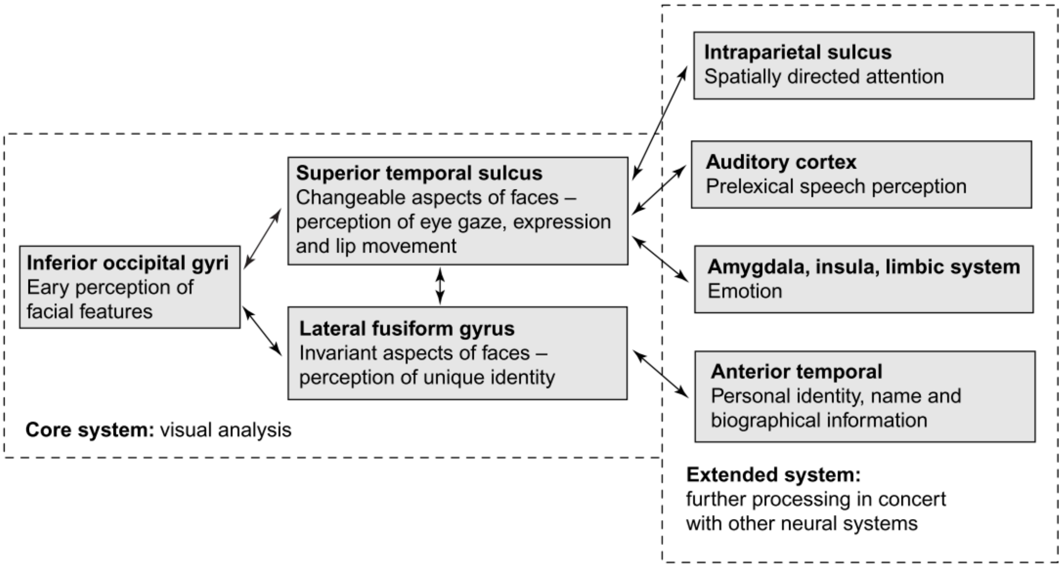


Figure 1.2. The distributed model of face processing proposed by Haxby et al. (2000). The model includes neural implementation of the functional components.

The Haxby model has been recently updated by its authors (Haxby & Gobbini, 2011) to reflect new findings accumulated during the past ten years. The updated model no longer sees the inferior occipital gyri as the "entry gate" for all face-related information, but continues to make a distinction between the visual mechanisms involved in processing dynamic facial features (in the posterior STS) and invariant facial features (the fusiform plus the inferior occipital gyri) (Figure 1.3). Furthermore, the bottom-up and top-down exchange of information between the core face system and the extended system is less constricted (there are no particular connections specified between individual components within the core system and individual components within the extended system). A brief reference to trait inferences in Haxby and Gobbini (2011) confirmed the initial hint that the authors see them as relying on mechanisms involved in processing changeable facial aspects. This conjecture was based on the emotion overgeneralisation hypothesis put forward by Todorov and colleagues (Todorov, Said, Engell & Oosterhof, 2008; discussed at length in the following sections), according to which trustworthiness inferences are driven by subtle emotional cues detected in faces by mechanisms typically responsible for facial expressions.

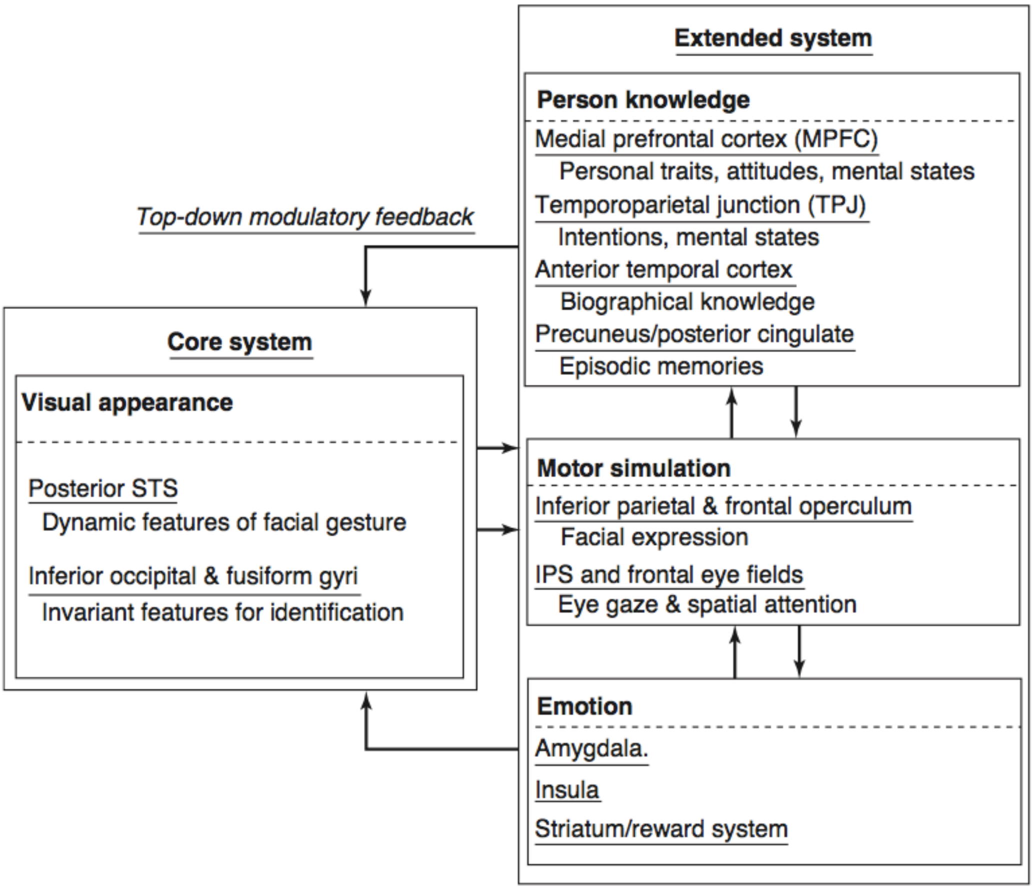


Figure 1.3. The updated distributed model of face processing (Haxby & Gobbini, 2011).

1.2.2. Dissociations between facial identity and trait perception

As we have seen, existing models of face processing were not particularly concerned with trait perception, focusing instead on the division between identity versus non-identity related aspects of face perception. Therefore, a first step in attempting to place face evaluations within face perception models is to clarify their position in relation to identity recognition.

Bruce and Young's model suggested face evaluations are performed by the same mechanisms involved in all other aspects of face perception that are not related to identity recognition. The dissociation between identity and trait perception is also implied in the Haxby model, with trait judgments likely to be carried out in the extended system based on information about changeable aspects of faces (distinct from information about invariant aspects needed for identity recognition). This dissociation was supported by findings from amygdala patients, who showed abnormal facial trait evaluations with spared facial identity recognition (Adolphs, Tranel & Damasio, 1988). Subsequent neuroimaging studies, discussed in more detail in the following section, confirmed the critical role of amygdala in making face evaluations (e.g. Winston, Strange, O'Doherty, & Dolan, 2002; Engell, Haxby, & Todorov, 2007; Todorov, Said, Oosterhof, & Engell, 2011). Furthermore, transcranial magnetic stimulation of the STS induced longer response times for face trustworthiness judgments (Dzhelyova, Ellison, & Atkinson, 2011), suggesting the STS, which is associated with processing changeable facial aspects, is also involved in trait perception. At the same time, face recognition is thought to depend on face-selective areas in the occipitotemporal cortex (Haxby et al., 2000). The fact that trait perception and identity recognition seem to rely on different neural areas is consistent with a dissociation between identity and trait perception. More convergent evidence for independent mechanisms in perception of identity and facial traits came from studies of individuals with developmental and acquired prosopagnosia (DP and AP, respectively). DP is a neurodevelopmental condition characterised by severe face recognition deficits, in the absence of intellectual or low-level vision problems. The main difference from AP is that in DP the face recognition deficits are not the result of known damage to a previously normal brain.

Todorov and Duchaine (2008) tested trait perception in four individuals with DP, who were asked to rate three sets of faces for perceived trustworthiness. Their ratings of the first two sets, which featured full face photographs, were in line with controls. When hair and facial marks were removed (third set), two DPs provided ratings which differed significantly from controls. The other two, however, gave ratings that were highly correlated with consensus judgments (and comparable to the ratings of a control group). Furthermore, Todorov and Duchaine (2008) found no relation between the severity of face recognition problems and performance with trustworthiness judgments, indicating that the two abilities rely on distinct mechanisms. These findings are consistent with face perception models presented above. Unfortunately, the four DPs were not tested for other aspects of face perception (e.g. emotional expressions) to reveal potential associations or dissociations with face impressions.

The fact that perceptions of identity and at least one trait (i.e. trustworthiness) are independent in some individuals with DP is not necessarily indicative of the organisation principles in normal brains. A recent study (Quadflieg, Todorov, Laguesse, & Rossion, 2012) examined perception of several facial traits in PS, a thoroughly studied case of acquired prosopagnosia (Rossion, Caldara et al., 2003). PS was asked to rate faces on two occasions (separated by a few months) and she showed judgments comparable to controls for aggression, attractiveness, confidence, intelligence, sociability, trustworthiness and typicality. The only abnormal ratings were for dominance. Because PS showed impaired holistic processing of faces, the results suggest that many traits (except dominance) do not rely on holistic processing.

These are suggestive results for a dissociation between identity recognition and (at least some) face impressions, but they do not answer all questions. First, traits seem to be an enduring and invariant feature of the face (repeated face evaluations are highly correlated, Oosterhof & Todorov, 2008). From this point of view, it might be more sensible to be processed by mechanisms involved in processing invariant facial aspects (such as identity), if the division between invariant and changeable aspects proposed by the Haxby model exists. Haxby and Gobbini (2011) dismiss this possibility, stating that although people may “naively” (p. 97) assume trait inferences are based on invariant facial features, this is not true. Unfortunately, they do not elaborate beyond a brief reference to the emotion overgeneralisation hypothesis. Instead, perception of facial traits depending on invariant features seems a reasonable possibility. Note that relying on the same invariant facial features does not preclude a dissociation between identity and trait perception; it would instead signal that Haxby model would need to be further updated to allow for finer dissociations within components of the face core system.

Second, and supporting the possibility that trait judgments and identity recognition rely on the same invariant facial features, most neuroimaging studies of trait perception find that face-selective areas, including the fusiform gyrus, are also sensitive to changes in perceived facial traits (Winston et al., 2002; Engell et al. 2007). Coupled with the fact that lesion of right amygdala may lead to prosopagnosia (see the patient Florence), this questions the existence of completely distinct neural mechanisms for identity and trait perception. Behavioural studies showed that facial identity can modulate trustworthiness inferences: observers perceived faces resembling themselves more favourably (DeBruine, 2002; DeBruine, 2005).

Third, in all previous studies, trait perception was measured with only one task – ratings – which was not comparable with any of the tasks used to measure face identification. Furthermore, some face stimuli (e.g. Quadflieg et al., 2012) were in colour and included hair; normal trait judgments of face stimuli including hair do not necessarily replicate with face stimuli excluding hair (Todorov & Duchaine, 2008), and skin colour can substantially influence trait judgments (Stephen, Coetzee, & Perrett, 2011; Swami, Furnham, & Joshi, 2008).

The current thesis aims to confirm the dissociation between facial identity and trait perception by: i) controlling many of the potentially confounding factors in previous studies, and ii) using multiple tests for each ability to increase the reliability of the findings. Because no suitable tests were available for trait perception, a new battery of tests was developed. The tests were designed to allow direct comparisons across face perception abilities and were based on two formats (sorting and categorisation) to measure perception of three traits (trustworthiness, attractiveness, aggressiveness) and two expressions (happy and angry, needed for testing the overgeneralisation theory, see next section). **Section 3.1** details the steps taken to develop and validate these tests. **Section 3.2** presents the examination of trait perception in four individuals with AP. The stimuli judged were grayscale faces cropped so that only internal features were visible. Perception of three traits (trustworthiness, attractiveness, aggressiveness) was evaluated with ratings, the sorting task and the categorisation task. The sorting tasks were developed based on a widely used test of face perception, facilitating direct comparison between trait and identity perception.

1.2.3. Overgeneralisation hypotheses in facial trait perception

The next step in trying to understand the mechanisms behind facial trait perception is to explore their relation to other judgments derived from faces, such as expression and sex recognition. My specific aim was to test the predictions of the overgeneralisation hypotheses derived from the leading model of facial trait perception (Oosterhof & Todorov, 2008).

Oosterhof and Todorov (2008) collected judgments of 15 commonly-inferred traits on a set of faces and, through principal component analysis (PCA), identified two principal components (PC) that accounted for the majority of variance (81%) in ratings. The first PC was a valence evaluation, with positive loadings for positive traits (e.g. trustworthiness, emotional stability, responsibility) and negative loadings for negative traits (e.g. meanness, weirdness). It was named the trustworthiness dimension because trustworthiness had the highest loading (.94) on it. The second PC had highest loadings from judgments of dominance, aggressiveness and confidence, and so it was named the dominance dimension. Similar two-dimensional models of social perception were proposed previously by Fiske, Cuddy and Glick (2007) and Wiggins, Philips and Trapnell (1989). According to these models, most person impressions could be reduced to warmth and competence (Fiske et al., 2007), or affiliation and dominance (Wiggins et al., 1989). The similar structure across all these models is convergent evidence for the reliability of the Todorov model.

Oosterhof and Todorov (2008) then used computer-generated faces to build an empirically validated model for how faces vary on trustworthiness and dominance. The model was adjusted such that the two dimensions were orthogonal, and it allowed new faces to be parametrically manipulated for trustworthiness or dominance. Oosterhof and Todorov (2008) observed that exaggerating trustworthiness led to faces exhibiting smiles, while exaggerating untrustworthiness made faces appear angrier. These results led them to speculate that observers judge trustworthiness from facial features resembling very subtle happy and angry expressions. Similarly, manipulating dominance in their computer model led to more masculine appearance of faces, which suggests dominance is inferred from features indicative of the sex of a face. Oosterhof and Todorov (2008) proposed that facial trustworthiness and dominance are byproducts of face mechanisms designed to make expression and sex judgments - the emotion and sex overgeneralisation theories. The emotion overgeneralisation theory was supported by later studies.

Said, Sebe and Todorov (2009) showed that a Bayesian network classifier trained to detect subtle cues to emotional expressions in neutral faces could predict how these faces were rated for impressions. Positive evaluations (e.g. trustworthiness) were associated with faces that were more likely to be categorised as happy, while negative evaluations were linked to faces categorised as disgusted and fearful. Dominant faces tended to be categorised as angry. Furthermore, Oosterhof and Todorov (2009) showed that face trustworthiness moderates perception of happy and angry expressions. Their participants saw two-second clips of faces expressing happy or angry emotion. When stimuli did not change identity and displayed changes in happiness (i.e. transition from neutral to expressive faces), participants rated trustworthy faces as happier than untrustworthy faces. The reverse was true for stimuli expressing anger. In addition, when face identity changed through morphing, the perceived intensity of an emotion varied according to whether it was congruent with the change in perceived trustworthiness between the start and end face in the clip. For example, perceptions of anger intensity increased for high-to-low trustworthiness transitions and decreased for low-to-high trustworthiness transitions.

These studies demonstrate perceptual similarities between trustworthy/ untrustworthy and happy/angry faces, but do not directly address the overgeneralisation hypothesis, which claims common functional and neural mechanisms between these aspects of face perception. In Engell, Todorov and Haxby (2010), behavioural adaptation to happy/angry faces (but not fearful faces) shifted trustworthiness perceptions of neutral faces upwards/downwards. The results were interpreted as evidence for a common neural population. However, note that the effect may simply arise because adaptation to angry faces makes subsequent neutral faces appear happy and happy faces were shown to positively influence perceptions of trustworthiness (Scharlemann, Eckel, Kacelnik, & Wilson, 2001). This does not necessarily imply common neural mechanisms.

Only a few studies investigated the neural bases of person impressions. In a landmark study, Adolphs et al. (1998) demonstrated the critical role of amygdala in perception of fear, but also trustworthiness and approachability. Patients with complete lesions of bilateral amygdala showed abnormal judgments related to face trustworthiness, i.e. they perceived untrustworthy faces as trustworthy. The findings fit well with the Haxby model, where amygdala is part of the extended system for face perception, important for emotion recognition (and possibly person impressions).

Amygdala has been found to be involved in facial trustworthiness judgments by several fMRI studies. In Winston et al. (2002), bilateral amygdala and right insula showed increase activation to untrustworthy faces, independent of the task (in contrast, the right superior temporal sulcus was activated only in explicit judgments of trustworthiness). Untrustworthy faces activated a few other areas, among which were the right and left fusiform gyri. The authors speculated the increased brain activity in the fusiform gyri, known for their role in identity recognition, was a result of modulatory influences from amygdala (possibly by anatomical back-projections) to enhance facial identification of threat stimuli. However, one cannot rule out the possibility that the fusiform gyrus is directly involved in facial trustworthiness perception, a finding that would be inconsistent with the Haxby model.

Engell et al. (2007) replicated the findings of Winston et al. (2002) by showing that amygdala was sensitive to changes in face trustworthiness when participants completed an unrelated face memory task. Furthermore, they showed that the amygdala response depended on consensus ratings of trustworthiness (average ratings from a pool of observers) and not on individual ratings, suggesting amygdala is sensitive to facial cues generally associated with trustworthiness by most observers, and not to particular cues which vary in their trustworthiness significance from one observer to the other.

A further replication came from Todorov, Baron and Oosterhof (2008). Using the same face memory task from Engell et al. (2007), this time with computer-generated faces instead of natural faces, they observed increased amygdala activation to untrustworthy faces. However, right and left amygdala displayed different response profiles. While right amygdala showed linear effects, mirroring previous findings, the authors noted a quadratic response in left amygdala, with both untrustworthy and trustworthy faces eliciting higher activation. A quadratic effect for face trustworthiness in amygdala was also found by Said, Baron and Todorov (2009), Todorov et al. (2011).

Todorov and Engell (2008) reanalysed the data from Engell et al. (2007) and found that amygdala activated to faces which were judged negative on 12 other traits (e.g. caring, sociable, mean etc), with the exception of dominance. They suggested that amygdala represents the valence dimension from the dual-component model of facial trait perception (Oosterhof & Todorov, 2008). The fact that amygdala was also shown to respond more strongly to happy than to neutral faces (Breiter et al., 1996; Winston, O'Doherty, & Dolan, 2003; Yang et al., 2002) is consistent with the emotion overgeneralisation hypothesis according to which face trustworthiness judgments are by-products of mechanisms involved in recognition of emotional expressions.

To summarise, Todorov and colleagues presented considerable evidence in support of the emotion overgeneralisation hypothesis. The sex overgeneralisation hypothesis received less attention, but previous studies showed that perceptions of facial masculinity or femininity influence facial dominance perception (Perrett et al., 1998). However, none of the above studies were designed to decisively test two fundamental predictions of the overgeneralisation hypotheses: i) impaired mechanisms for recognition of happy and angry facial expressions should lead to abnormal perception of face trustworthiness, and ii) impaired mechanisms for facial sex recognition should determine abnormal perception of face dominance. In **Section 3.3** I test these two predictions in one individual with AP who showed impairments with facial expression and sex perception in preliminary tests.

1.3. FACE IMPRESSIONS

The final experimental sections of the present thesis are dedicated to four topics in face impressions. First, I look at facial cues influencing trait perception. More specifically, I examine whether facial width-to-height ratio predicts trust and explicit evaluations of trustworthiness as it has been recently suggested (Stirrat & Perrett, 2010). Second, I investigate whether people can detect trustworthy behaviour in a real-life setting and whether the accuracy of their judgments depends on the stimuli source (real versus laboratory-based) and type (static versus dynamic clips). Third, I test if faces continue to influence economic interactions in an information-rich environment, a more realistic setting than those used in prior studies (e.g. van't Wout & Sanfey, 2008) in which participants did not have access to any information about their potential partners. Finally, because social exchanges usually involve both visual and auditory interactions, I examine how face impressions combine (or not) with voice impressions to form an integrated person perception. Of the various traits inferred from faces, the focus here is on trustworthiness, with the exception of the final study in which impressions of attractiveness and aggressiveness are also investigated.

1.3.1. Cues to face trustworthiness

Face evaluations are surprisingly consistent across observers (Todorov et al., 2011; Penton-Voak, Pound, Little, & Perrett, 2006; Todorov et al., 2005), suggesting they are driven more by stimulus characteristics than observers' idiosyncrasies. In other words, facial traits are less in the 'eye of the beholder' (i.e. individual and cultural background), and more universal preferences (Cunningham, Roberts, Barbee, Druen, & Wu, 1995; Perrett, May, & Yoshikawa, 1994; Rhodes et. al., 2001). This observation leads one to wonder which facial characteristics (or complex of characteristics) are typically associated with various traits. For example, attractiveness has been linked to facial averageness, symmetry and sexual dimorphism (feminine traits are found attractive in women, and masculine traits are found attractive in men) (Rhodes, 2006), while perceived honesty has been linked to large eyes (Zebrowitz, Voinescu, & Collins, 1996).

A recent study (Kleisner, Priplatova, Frost, & Flegr, 2013) found that brown-eyed faces tended to be perceived as being more trustworthy than blue-eyed faces, even after correcting for perceived attractiveness and dominance. However, the effect was not due to eye colour itself, but to the correlation between eye colour and face shape. Brown-eyed faces were generally rounder and broader, with big eyes, small distance between eyebrows and mouth corners pointing up (blue-eyed faces tended to have the opposite features). These features may be associated with trustworthiness (Oosterhof & Todorov, 2008; Zebrowitz & Montepare, 2008). Interestingly, a previous study from the same lab (Kleisner, Kocnar, Rubesova, & Flegr, 2010) found that brown-eyed faces were also perceived to be more dominant, which is difficult to reconcile with the negative correlation usually found between face trustworthiness and face dominance (Oosterhof & Todorov, 2008). Kleisner et al. (2013) suggested that eye colour may transmit two independent signals, one related to trustworthiness and another one related to dominance.

Because trustworthiness is positively correlated with attractiveness and negatively correlated with dominance/aggressiveness (Oosterhof & Todorov, 2008), it is expected that facial characteristics affecting perceptions of attractiveness or aggressiveness will also influence perceptions of trustworthiness. One reliable cue to aggressiveness has been found to be the facial width-to-height ratio (WHR), defined as the ratio between the bizygomatic breadth (the distance between the left and the right zygion) and the distance between the upper lip and the brow (Figure 1.4). According to Weston, Friday and Liò (2007), facial WHR is a sexually dimorphic aspect of the face, with greater ratios (i.e. wider faces) in men (but see Kramer, Jones, Ward, 2012; Özener, 2011; for challenges to this claim). Greater WHRs in men are independent of body size and may be related to increased testosterone level (Verdonck, Gaethofs, Carels, Zegher, 1999). At the same time, Carré and McCormick (2008) found that facial WHR in men correlated with self-reported dominance and aggressive behaviour in hockey games (expressed as penalty minutes). In a subsequent study, Carré, McCormick and Mondloch (2009) showed that observers' perceptions of facial aggressiveness were strongly correlated with the facial WHR and with the actual aggressiveness of the targets, even under brief presentation times (39 ms). Putting these findings together, one can infer that facial WHR in men is a valid cue that helps observers make accurate predictions regarding targets' aggressive behaviour. Carré, Morrissey, Mondloch and McCormick (2010) confirmed this conjecture in a series of experiments. They found that aggressiveness judgments remained accurate for stimulus manipulation that did not interfere with the facial WHR (blurring and cropping) but not when faces were scrambled. Furthermore, they showed that facial WHR was the main predictor (from a larger set of facial metrics) of aggressiveness judgments.

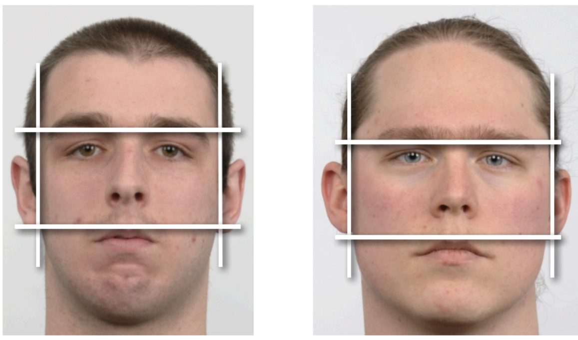


Figure 1.4. Computation of facial width-to-height ratio (WHR). The metric is calculated by dividing the distance between the left and right zygion and the distance between upper lip and brows. The man on the right has a higher WHR than the man on the left.

Inspired by these studies, Stirrat and Perrett (2010) examined whether the relations between facial WHR and perceived and actual aggressiveness extend to trustworthiness. They measured the facial WHR of male and female students and found that, for male students, it correlated with reciprocation rates in trust games (i.e. actual trustworthiness). In a second study, male facial WHR predicted the amount invested by first movers in the trust game (i.e. perceived trustworthiness). Finally, manipulation of male facial WHR was shown to influence the answers in a task asking to select the more trustworthy face from a pair, with narrower faces selected more often. The effect was more pronounced for female participants with lower scores on a self-rated social dominance questionnaire. Stirrat and Perrett concluded that, as for aggressiveness, facial WHR is a valid predictor of male actual and perceived trustworthiness.

In Section 4.1 I present an attempt to replicate these findings and answer a few outstanding questions. First, a replication is important because the effects associated with WHR are relatively small and context-dependent. For example, the initial correlation between facial WHR and aggressive behaviour found by Carré and McCormick (2008) was not replicated in a subsequent study. After measuring facial WHR in 470 Turkish university students, Özener (2011) failed to find a correlation with self-reported aggressiveness. Haselhuhn and Wong (2011) replicated the initial findings that facial WHR predicts actual trustworthiness; in their study, men with wider faces were more likely to cheat. But Stirrat and Perrett (2012) showed that large facial WHR is not always linked to antisocial behaviour; in a public goods game, men with wider faces were willing to sacrifice more for the benefit of other in-group members. My replication attempts include the original faces used by S&P and a new set of more controlled faces generated by Facegen Modeller 3.3 software (Singular Inversions, 2008) according to the metrics from S&P. Although facial WHR is thought to be a reliable cue to aggressiveness/ trustworthiness only for men, our computer-generated faces included both female and male faces. A second outstanding issue is that WHR may not be a direct predictor of perceptions of trustworthiness, but one of the cues important for other variables, such as perceptions of attractiveness and femininity, that lead to variances in trustworthiness (e.g. Perrett et al., 1998). I test whether any of these variables is a better predictor of face trustworthiness than facial WHR. Third, I examine if observers' self-reported traits like propensity to trust, risk-aversion, dominance and attractiveness mediate the relationship between facial WHR and trustworthiness evaluations. S&P found individual differences in raters' judgments, with self-reported submissive female participants more sensitive to the male facial WHR.

1.3.2. Accuracy of face impressions

Face impressions are pervasive (Willis & Todorov, 2006) and typically show high inter-rater agreement (Oosterhof & Todorov, 2008). The confidence that people generally have on their impression of others' faces (Hassin & Trope, 2000) might suggest that personality traits and/or behavioural propensities can be predicted from faces. Consistent with this idea, evaluations of strangers' faces were found to correlate with certain self-reported traits such as extraversion and conscientiousness (Albright, Kenny, & Malloy, 1988; Little & Perrett, 2007), agreeableness and emotional stability (Penton-Voak et al., 2006), power and warmth (Berry, 1991), honesty (Bond, Berry, & Omar, 1994) and intelligence (Zebrowitz, Hall, Murphy, & Rhodes, 2002). A higher-than-chance correlation between face evaluations and personality attributions was found across cultures (Albright et al., 1997). Other authors claimed that, based on faces, people can predict membership to religious (Rule, Garrett, & Ambady, 2010), political (Rule & Ambady, 2010) or sexual minority groups (Rule, Ambady, & Hallett, 2009; Rule & Ambady, 2008; Rule, Macrae, & Ambady, 2009). Together, these studies support the "kernel of truth" hypothesis (Berry, 1990), according to which there is some truth (i.e. small, but significant effect) in face-based judgments. However, the evidence about the accuracy of face impressions is mixed, with many studies failing to find correlations with actual traits (Hassin and Trope, 2000; Pound, Penton-Voak, & Brown, 2007; Zebrowitz et al., 1996; Zebrowitz, Andreoletti, Collins, Lee, & Blumenthal, 1998). Furthermore, as Penton-Voak et al. (2006) point out, very few studies showing significant correlations used stimuli properly controlled to present only facial cues, so personality impressions might have been helped by non-facial cues such as clothing and hairstyle.

Trust is central to most social exchanges and so perceptions of trustworthiness have the potential to yield large payoffs when accurate and damaging penalties when false. However, despite its prominent role in social interactions, relatively few studies have explicitly examined our ability to infer trustworthiness from faces. While there is a large body of research on lie-detection[1], these studies typically presented participants audiovisual material recorded in the laboratory, encouraging reliance on both verbal and nonverbal information to reach hopefully accurate lie-truth judgments about artificial behaviour (e.g., lies according to a script). Only a fraction of these studies presented face videos exclusively, and in this case the performance was not better than 50%. Closer to face trustworthiness detection are the studies derived from the evolutionary psychologists' hypothesis that people have evolved a specialised module for recognising abusers of trust - the 'cheater detection module' (Cosmides, 1989; Cosmides & Tooby, 1992). Faces of cheaters were found to be better remembered than faces of non-cheaters (Chiappe & Brown, 2004; Mealy, Daood, & Krage, 1996; Oda, 1997; but see also Barclay & Lalumiere, 2006; Mehl & Buchner, 2008; for failed replications of this finding). However, better memory for cheaters does not mean cheaters can be recognised *before* interacting with them or in the absence of any information about their actual behaviour.

More recent studies examined whether people can reliably predict cooperativeness in economic games such as the trust game, prisoner's dilemma game, ultimatum game or dictator game, from 5-20s video clips of the players - what Ambady and Rosenthal (1992) refer to as "thin slices" of behaviour. Results were mixed, with some studies observing chance-level performance (Gollwitzer, Rothmund, Alt, & Jekel, 2012) and others finding significant correlations between face-based trustworthiness impressions and cooperative behaviour (Fetchenhauer, Groothuis, & Pradel, 2010).

Information available in still images of faces, under certain conditions, seems to be sufficient to predict trustworthy/cooperative behaviour in laboratory games. Verplaetse, Vanneste and Braeckman (2007) showed that participants could discriminate cooperators from non-cooperators in prisoner's dilemma games from still photos taken at the moment of making the decision of whether or not to cooperate. Discrimination from still images taken before the experiment or during practice trials failed. In Bonnefon, Hopfensitz and De Neys (2012), successful

[1] Bond and DePaulo (2006) carried out a comprehensive meta-analysis of 206 lie-detection studies and found that people correctly discriminated between truthful and deceptive messages about 54% of the time, on average (where 50% would be chance-level performance).

prediction of reciprocation rates from trustees in trust games depended on the face images being black-and-white and cropped to exclude external cues. When pictures were in colour and included non-facial features, performance was at chance level, even though participants preferred these stimuli.

Stiratt and Perrett (2010) suggested that a particular facial metric is informative about people's propensity to reciprocate trust. In their experiment, people with lower facial width to height ratios were judged more trustworthy and this metric correlated with their cooperative behaviour in the trust game. Furthermore, they showed that varying facial width modified perceptions of face trustworthiness and concluded that facial width is a valid cue used by observers to detect trustworthiness in strangers. However, a later study (Stirrat & Perrett, 2012) showed that people with wider faces may switch between antisocial and prosocial behaviour according to the context.

Other studies looked at *generalised* trustworthiness (a long-lasting character trait) rather than *state* trustworthiness (referring to a particular action in a given context). Generalised trustworthiness is related to state trustworthiness (can be seen as a collection of state trustworthiness actions), but is not a deterministic predictor of it. An individual high on generalised trustworthiness may display state untrustworthiness (e.g., a Nobel Peace Prize recipient stealing a pen at a conference). Porter, England, Juodis, Ten Brinke and Wilson (2008) investigated face-based trustworthiness perceptions elicited by two groups of people supposed to vary dramatically in their generalised trustworthiness (Nobel Peace Prize recipients and wanted criminals) and found that judgment accuracy was above chance only for trustworthy targets (i.e., Nobel prize winners). Another study (Valla, Ceci, & Williams, 2011) claimed that people can accurately discriminate between criminals and non-criminals from their photos. However, the criminals' photos were police mugshots and the differences in pose and lighting to the non-criminals' photos made it relatively easy to infer which faces presented criminals. When participants were asked to discriminate between violent and non-violent offenders, both groups now presented in mugshots, they failed. Similar confounds in the stimuli might have also affected the results of Porter et al. (2008) showing above chance performance at distinguishing wanted criminals from Nobel laureates. Even though photos were cropped to include only the faces, criminals' photos taken from FBI's "most wanted" list might still look different from photographs of Nobel laureates (no example photos were presented in the paper to verify this possibility).

To summarise, there is limited evidence that state or generalised trustworthiness can be inferred from faces. Furthermore, all studies investigating state trustworthiness detection from faces relied on analysing the behaviour exhibited by targets in relatively low-stakes laboratory games. It is debatable whether such behaviour is a good proxy for actual trustworthiness. By not reciprocating trust or by not cooperating, targets follow acceptable rules of the games. In the absence of an explicit commitment to cooperate, it is difficult to classify their behaviour as deceptive or untrustworthy. Would the evidence for the accuracy of face trustworthiness impressions be stronger if people have to judge targets for clear deceptive behaviour in real-world settings? Data from a television program featuring a high-stake prisoner's dilemma game allowed me to investigate this question. In **Section 4.2** I present three experiments in which I examine if face impressions based on static and dynamic stimuli can predict trustworthy behaviour in the real-world (using the television data) and also in the laboratory (where targets were asked to play trust games). In addition to responses about targets' behaviour, I also measure participants' confidence in their responses to assess whether it correlates with accuracy. Expecting participants to reliably classify all targets may be unrealistic - some targets may be easier to "read" than others. Similarly, some participants may be better than others at "reading" faces. However, if "reading" faces is a skill, people should have insight relative to its variance and a reliable correlation between cheater detection and confidence levels should be found.

1.3.3. Face impressions and reputation

The temptation to judge strangers by their faces is hard to resist, with many people convinced they can accurately tell the virtuous from the wicked simply by their faces (Hassin & Trope, 2000). Research suggests there is limited value in face-based judgments (Hassin & Trope, 2000; Olivola & Todorov, 2010b; although see Berry, 1990; Bond, Berry, & Omar, 1994) and, overtly, we label the use of this self-perceived ability as unethical. Nevertheless, the fast and spontaneous process (Willis & Todorov, 2006) of inferring traits from faces appears to influence a wide range of consequential decisions: face-based impressions of competence were found to predict election results (Olivola & Todorov, 2010; Todorov et al., 2005), facial dominance predicted achieved military rank of West Point cadets (Mueller & Mazur, 1996) and facial baby-facedness and attractiveness influenced court decisions (Zebrowitz & McDonald, 1991). In the laboratory, van't Wout and Sanfey (2008) showed that participants in trust games invested more in anonymous partners with faces perceived to be more trustworthy. Rezlescu, Duchaine, Olivola and Chater (2012) (experiment 1) replicated these results with face stimuli that had no hair, facial marks, glasses etc., such that the effect of perceived trustworthiness was unambiguously linked to the facial structure.

However, an important limitation in prior experimental studies examining the role of facial trustworthiness on social or economic interactions is that participants had no information about their potential partners beyond their faces (e.g. van't Wout & Sanfey, 2008). In the absence of information about past behaviour, it may be expected that people use any cues available about a partner's trustworthiness, including facial cues assumed to be unreliable. But in real life people usually have access to rich information about prospective partners. Considering that face judgments were found to be updated quickly in line with reputation (Todorov & Olson, 2008), a legitimate question is whether the effects of trustworthy facial configurations survive when reliable background information is available.

Recent studies (Duarte, Siegel, & Young, 2010; Ravina, 2008) have found that appearance-based perceptions of borrower trustworthiness predict lending tendencies in online peer-to-peer lending, even when lenders have demographic and financial information about borrowers. But these studies were correlational, and borrower photos often included more than faces, so it is uncertain what aspects of appearances influenced investment choices. In a more controlled laboratory environment, Chang, Doll, van't Wout, Frank, & Sanfey (2010) found that, in trust games involving multiple interactions with the same trustee, participants dynamically tuned their investment strategies to favour partners who reciprocated their trust. Across 15 repeated interactions, the main effect of facial trustworthiness was not significant, but trustworthy-looking partners who reciprocated trust still received more money than reciprocating partners with untrustworthy looks.

The experiment presented in Section 4.3 has a similar goal: to study the interaction between face trustworthiness and reputational information. In Rezlescu, Duchaine et al. (2012) (experiment 1), stable facial configurations varying in perceived trustworthiness affected investment decisions. I examine if the effect survives in a richer, more ecologically valid environment. Keeping the trust game setup (Berg, Dickhaut, & McCabe, 1995), participants playing the part of the investors will now see not only faces but also information about their counterparts' past behaviour. This information is presented in the form of visual unambiguous histories, aimed to suggest high or low reciprocity in previous trust games.

Unlike Chang et al. (2010), in the current study I use highly controlled facial stimuli to focus on unfakeable facial features. Furthermore, rather than gradually discovering trustee reputations from first-hand interactions, participants see visual summaries of their partners' past reciprocations (just as one might receive third-party reports about potential business partners). Thus, participants have simultaneous access to faces and reputational information, so they can integrate both immediately. Finally, participants interact with each trustee only once, eliminating the potential confound, associated with repeated games, that investment decisions might be used to punish or reward trustees, or to otherwise communicate (dis)satisfaction with a partner's choices (Xiao & Houser, 2005).

Rationally, people should consider only objective cues about a trustee's behaviour and ignore facial cues. However, the prediction is that, rather than being obliterated, the effect of face trustworthiness will be substantially reduced (but still significant).

1.3.4 Voice impressions

Voices are a rich source of socially relevant information. In this respect, they closely resemble faces, despite being markedly less researched. Recognition of a person's identity, age, sex and emotions is commonly achieved based on a combination of facial and vocal information, but it can also succeed based solely on either faces or voices (e.g. Banissy et al., 2010; Meyer, Baumann, Wildgruber, & Alter, 2007; Sauter, Eisner, Ekman, & Scott, 2010; Scott, 2008). These similarities between faces and voices led some authors to suggest that voice perception relies on functional and neural architectures similar to face perception (Belin, Fecteau, & Bédard, 2004; Burton, Bruce, & Johnston, 1990; Ellis, 1989). However, this does not imply the mechanisms behind face and voice processing are the same. Indeed, in addition to the obvious differences in low-level perception, arising from the different physical attributes of voices and faces, there is conclusive evidence for separate mechanisms at higher levels. For instance, most individuals with prosopagnosia can recognise people from their voices (e.g. Hoover, Démonet, & Steeves, 2010), and individuals with phonoagnosia (the inability to recognise vocal identity) can have normal face recognition (Garrido, Eisner et al., 2009). A dissociation was also documented in perception of affective information: amygdala patient SM could recognise fear from voice prosody but not from faces (Adolphs & Tranel, 1999).

While face-based trait impressions have been studied extensively (for a review see Todorov et al., 2011), most research on voice perception has revolved around identity and emotion recognition, with only a few studies examining trait impressions. These studies were mainly interested in vocal attractiveness and its influence on electoral outcomes. Voice pitch was found to be a strong cue to male vocal attractiveness, with deeper voices preferred by both men and women (Collins, 2000; Feinberg, DeBruine, Jones, & Little, 2008; Feinberg, Jones, Little, Burt, & Perrett, 2005; Jones, Feinberg, DeBruine, Little, & Vukovic, 2010). Low voice pitch is also associated with perceived dominance in men (Jones et al., 2010; Puts, Gaulin, & Verdolini, 2006; Puts, Hodges, Cardenas, & Gaulin, 2007). Just like faces, voices were shown to bias political preferences and voting behaviour (Gregory & Gallagher, 2002; Surawski & Ossoff, 2006; Tigue, Borak, O'Connor, Schandl, & Feinberg, 2012) and, also similar to many studies on the accuracy of face impressions, researchers failed to find a relationship between body characteristics and vocal characteristics or voice impressions (Collins, 2000). Nevertheless, deep voices may be honest signals to dominance in men because they correlate with testosterone level (Dabbs & Mallinger, 1999).

Trustworthiness is one of the main components in the two-dimensional model of facial trait perception (the other being dominance; Oosterhof & Todorov, 2008). Despite its likely important role in person evaluation, voice trustworthiness has not been studied so far. With faces, perceived trustworthiness was shown to have a positive correlation with attractiveness and a negative correlation with dominance (Oosterhof & Todorov, 2008). Would these correlations also emerge with voices? Would voice-based trustworthiness, attractiveness and dominance judgments show the high inter-rater agreement commonly found for face-based judgments? These are two of the questions I address in Section 4.1, Experiment 1, where I validate a novel set of voices and faces for perceived trustworthiness, attractiveness and dominance. In the same experiment, I examine whether voice impressions correlate with face impressions. Feinberg (2008) suggested that faces and voices might signal the same qualities related to hormone levels and thus evaluations of those qualities from faces and voices should correlate. Indeed, women with masculine faces were found to also have lower-pitched voices (Feinberg et al., 2005) and voice and face attractiveness were related in women (Collins & Missing, 2003; Feinberg et al., 2005; Lander, 2008) and men (Saxton, Caryl, & Roberts, 2006).

Because most real-world interactions involve simultaneous access to facial and vocal information, another interesting question is how face and voice impressions combine (or not) to form an integrated person impression. However, research on this topic is extremely scarce, with only one study showing that physical and vocal attractiveness produced a supra-additive effect on perception of political candidates (Surawski & Ossoff, 2006). In Experiment 2 I examine integrated person impressions of attractiveness, trustworthiness and dominance based on faces and voices previously rated low or high on these dimensions. The prediction is that both faces and voices will influence person perception, but that the balance between visual and auditory cues and the presence or absence of an interaction will vary according to the judged trait.

Chapter 2

FACE SPECIFICITY IN ACQUIRED PROSOPAGNOSIA

The aim of this chapter is to test two alternative theories to face specificity: the individuation hypothesis and the expertise hypothesis. To this end, I report a series of studies involving two individuals with acquired prosopagnosia (AP). The individuation hypothesis implies that, in addition to their definitional face recognition deficits, individuals with AP will also be impaired at fine within-level discrimination of non-face objects. The expertise hypothesis predicts that individuals with AP will not be able to acquire object expertise.

2.1. SPARED WITHIN-CLASS OBJECT RECOGNITION IN ACQUIRED PROSOPAGNOSIA

INTRODUCTION

Acquired prosopagnosics (APs) are individuals who, following brain damage, experience severe face recognition deficits in the absence of intellectual or low-level visual impairments. Deficits in acquired prosopagnosia (AP) are varied and different forms seem likely to have particular unique neural correlates, so an understanding of AP has the potential to dramatically advance our understanding of face processing in normal brains. Here we report a new case of acquired prosopagnosia, Herschel, a 56-year old British man with a degree in astronomy, who contacted us in 2009 because of difficulties with face identification following two strokes. The current study aimed to investigate his prosopagnosia and contribute to the debate on face specificity, a central issue in visual recognition.

Are the mechanisms involved in face recognition different from those used in other types of visual recognition? If they are different, we should expect to find APs who are normal with object recognition ("pure" prosopagnosics). If face and other sorts of visual recognition depend on the same mechanisms, the brain damage responsible for prosopagnosia should always impair nonface object recognition. While most APs present severe deficits with faces *and* objects (e.g. Barton, 2008; Boutsen & Humphreys, 2002; Delvenne et al., 2004; Gauthier, Behrmann, & Tarr, 1999; Levine & Calvanio, 1989; Steeves et al., 2006), the neuropsychological literature also describes APs who seem to be able to correctly identify nonface objects of different types (basic level recognition) and same type (within-level recognition).

A recent paper, Busigny, Joubert et al. (2010), included a detailed description of one such case along with a summary table of 13 other potentially "pure" cases, spanning 25 years of research on acquired prosopagnosia. But as Busigny, Joubert et al. (2010) note, there is yet no irrefutable evidence for "pure" prosopagnosia. Damasio et al. (1982) pointed out that within-class discrimination (e.g. distinguish car A from car B) is more comparable than basic object recognition (e.g. distinguish car from chair) to face recognition, and a careful examination of the 13 cases summarized in the table from Busigny, Joubert et al. (2010) revealed that the ability to identify items within various highly homogeneous object categories was not rigorously tested in many of them. For example, Case 3 from Takahashi, Kawamura, Hirayama, Shiota and Isono (1995) was tested for naming real basic objects (normal performance), but not for identification of any within-category items. The same can be said about patient 009 (Barton, Cherkasova, Press, Intriligator, & O'Connor, 2004) and the unnamed patient tested by Wada and Yamamoto (2001), both of whom were able to name fruits, vegetables, and animals normally. Although some researchers (Barton et al., 2004; Schweinberger, Klos, & Sommer, 1995) argued for a parallel between these tests and face recognition, the basic level recognition tests place demands on the cognitive system that are likely to be different from the individuation demands of face recognition. In addition, these basic level tests typically produce ceiling effects in control participants, making the discovery of subtle impairments difficult.

Another problem is that some within-class recognition tests used in previous papers had issues that made the interpretation of results problematic. For example, De Renzi and colleagues assessed within-category discrimination in Patient 4 (De Renzi, 1986a) and VA (De Renzi, Faglioni, Grossi, & Nichelli, 1991) with two tests. One was a coin discrimination test, in which patients were required to sort local from foreign coins, and in the second test, the patients were successful at recognizing personal items from a set of similar items (e.g. their necktie from other neckties). However it is not certain they would succeed with less familiar items and it is impossible to gauge whether the tasks were as difficult as the face tests these patients scored poorly on. Even when sensitive tests for within-level recognition were used, APs often were tested with only one category of objects. Such was the case for patients Anna (De Renzi & di Pellegrino, 1998), WB (Buxbaum, Glosser, & Coslett, 1996), LR (Bukach, Bud, Gauthier, & Tarr, 2006), FB (Riddoch, Johnston, Bracewell, Boutsen, & Humphreys, 2008), DC (Rivest, Moscovitch, & Black, 2009), who recognized exemplars normally from within one of the following categories: glasses, doors, novel objects and dog breeds.

To summarise, many of the potentially face-specific cases of AP do not convincingly demonstrate accurate perception of visually similar exemplars within non-face categories. A rigorous documentation of cases with normal exemplar discrimination for a wide array of objects is crucial to support claims that acquired prosopagnosia can result from deficits to face-specific mechanisms as the face specificity hypothesis entails. Below we present results from Herschel on perceptual and/or memory tests for cars, houses, tools, sunglasses, guns, horses, novel objects, human bodies and hairstyles, making him one of the most thoroughly tested prosopagnosics for within-class object discrimination, along with PS (Busigny, Graf et al., 2010) and GG (Busigny, Joubert et al., 2010).

In addition to investigating non-face recognition, another way to address whether prosopagnosia can be face-specific is to test predictions of accounts of prosopagnosia that do not involve face-specific mechanisms. One such account is Levine and Calvanio's (1989) proposal that prosopagnosia arises as a result of a general impairment in visual configural processing, defined as "the ability to obtain an overview of an item as a whole in a single glance". If individuals have problems forming unified representations of objects from individual parts, they would be impaired with objects for which recognition by parts matching is difficult and Levine and Calvanio (1989) suggest faces are such a category. A testable prediction of this view is that prosopagnosics will be impaired with objects for which recognition is normally done by parts, but for which critical chunks have been occluded such as Mooney-type objects (Mooney, 1957). Indeed, patient LH performed poorly in tests of visual closure, in which he was required to identify objects or words presented under challenging viewing conditions (in incomplete form or with visual noise added) (Levine & Calvanio, 1989). LH, however, had a wider range of cognitive deficits, including basic level object naming difficulties (Levine, Calvanio, & Wolf, 1980; Levine & Calvanio, 1989), which makes the interpretation of his results problematic. In contrast, several other APs with unimpaired object recognition showed normal performance in visual closure tasks (Whitely & Warrington, 1977; De Renzi, 1986a; De Renzi et al., 1991; De Renzi & di Pellegrino, 1998; Henke, Schweinberger, Grigo, Klos, & Sommer, 1998). Here we investigate whether Herschel's prosopagnosia is associated with a deficit in general configural processing that may pass undetected in usual object recognition, but become apparent in tests of visual closure. We also test Busigny, Joubert et al.'s (2010) related but more specific claim that acquired prosopagnosia is necessarily linked to a deficit in face-specific (rather than general) configural processing.

CASE REPORT: HERSCHEL

Herschel is a 56-year old (born 1956) right-handed British man. He holds a degree in astronomy (hence his patient name) and currently manages a science and technology team. Herschel contacted us through the Prosopagnosia Research Center website (www.faceblind.org) in October 2009 because he suffered from face recognition problems. In February 2008 he suffered a stroke that produced prosopagnosia, face-related visual anomalies ("mouths had tiger-like snarls"), severe navigation problems ("I could not navigate around the streets where I live") and an upper left quandrantanopia. In June 2008 a second stroke produced a temporary loss of color perception and upper right quandrantanopia. In August

2008 he suffered two transient ischemic attacks that produced temporary loss of control of the left leg and temporary speech problems. Currently, he reports only face recognition difficulties and an upper visual field loss (complete left and two thirds right). Navigation abilities and color perception largely returned, although Herschel says that they remain different from how they were before his strokes. Nevertheless, he seemed to effortlessly find his way around London and the inside of the building where he was tested, and he performed normally in our color perception tests (see below). He is intellectually normal (see tests below) and continues to run his lab.

NEUROIMAGING FINDINGS[†]

Whole-brain imaging was performed on a Siemens 1.5 Tesla MR scanner at the Birkbeck-UCL Neuroimaging Centre (BUCNI) in January 2010. Functional data were acquired over four blocked-design functional runs each lasting 234 s using a gradient-echo echo planar imaging (EPI) sequence (23 slices, repetition time, TR = 2 s, echo time, TE = 50 ms, voxel size = 3x3x3 mm). In addition, a high-resolution anatomical scan (T1-weighted FLASH, TR = 12 ms; TE = 5.6 ms; 1-mm^3 resolution) was acquired at the start of each scanning session for anatomically localizing functional activations. In addition to Herschel, we also scanned four male control subjects (age range 38 to 48 years).

Structural data (Figure 2.1) showed hydrocephalus with enlargement of the lateral ventricles, the third ventricle, the fourth ventricle, and the interpeduncular fossa. An extensive cyst located above the right tentorium cerebelli suppressed the right ventral and medial occipital lobe including occipitotemporal gyrus, occipital gyrus, and lingual gyrus. The cyst extended to the right hippocampal formation, but did not reach or affect the right amygdala. There was a very minor midline shift in the cerebellar vermis.

[†] Structural and functional imaging were performed by David Pitcher. This section ("Neuroimaging findings") was written by him for Rezlescu, Pitcher and Duchaine (2012).

To identify category-selective regions in the visual cortex, we used a dynamic functional localiser (Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011). Stimuli were 3-second movie clips of faces, bodies, scenes, objects, and scrambled objects. Each functional run contained two sets of five consecutive dynamic stimulus blocks (faces, bodies, scenes, objects, or scrambled objects) sandwiched between rest blocks, making two blocks per stimulus category per run. Each block lasted 18 seconds and contained stimuli from one of the five stimulus categories. The order of stimulus category blocks in each run was palindromic (e.g., fixation, faces, objects, scenes, bodies, scrambled objects, fixation, scrambled objects, bodies, scenes, objects, faces, fixation) and was randomized across runs. To focus attention, participants were instructed to press a key whenever the subject of the movie clip was repeated twice in a row (1-back task).

Functional imaging data were analysed using FSL (Smith et al., 2004). After deleting the first four volumes of each run to allow for T1 equilibrium, the functional images were realigned to correct for small head movements (Jenkinson, Bannister, Brady, & Smith, 2002). The images were then smoothed with a 5-mm FWHM (full width at half maximum) Gaussian filter and pre-whitened to remove temporal autocorrelation (Woolrich, Ripley, Brady, & Smith, 2001). The resulting images were entered into a subject-specific general linear model with five conditions of interest corresponding to the five categories of visual stimuli (faces, bodies, scenes, objects, and scrambled objects). Blocks were convolved with a double-gamma canonical hemodynamic response function (Glover, 1999) to generate the main regressors. In addition, the estimated motion parameters were entered in as covariates of no interest, to reduce structured noise due to minor head motion. Functional images were then registered to each participant's individual structural scan using a 12-degrees-of-freedom affine transformation (Woolrich et al., 2001).

The last two functional runs were used to define category-selective regions of interest (ROIs) within each participant. We identified face regions by contrasting faces greater than objects, body regions by contrasting bodies greater than objects, scene regions by contrasting scenes greater than objects, and object regions by contrasting objects greater than scrambled objects. The same statistical threshold ($p = 10^{-3}$, uncorrected) was used for all participants. Within each functionally defined ROI, we then calculated the magnitude of response (percentage signal change, or PSC, from a fixation baseline) to the conditions of the four stimulus categories (faces, bodies, scenes, objects, and scrambled objects), using the data collected from the first two runs. All of the data used to calculate PSC were independent of the data used to define the ROIs.

Although we could identify all core face areas in Herschel's right hemisphere, a comparison between Herschel's and controls' activation levels to different stimulus categories revealed an atypical pattern for Herschel (Figure 2.2). In all face areas, his absolute PSCs to faces were below those of controls, as were the relative increases in PSC from objects to faces. We note that the differences are not statistically significant (all ps > .13), although we also point out the limited statistical power provided by such a small control group. The levels of activation in the left fusiform face area (FFA) and left superior temporal sulcus (STS) were similar to those in the right hemisphere. We could not identify a left occipital face area (OFA) in Herschel, but one control did not show a left OFA as well. Herschel also failed to show a right parahippocampal place area (rPPA) as this region of cortex was damaged by his stroke. The other category-selective regions targeted by our functional localizer (Pitcher et al., 2011) were present, including the lateral occipital complex (LOC). Herschel's activation to objects in the LOC was comparable to that seen in controls, suggesting spared functional mechanisms for object recognition.

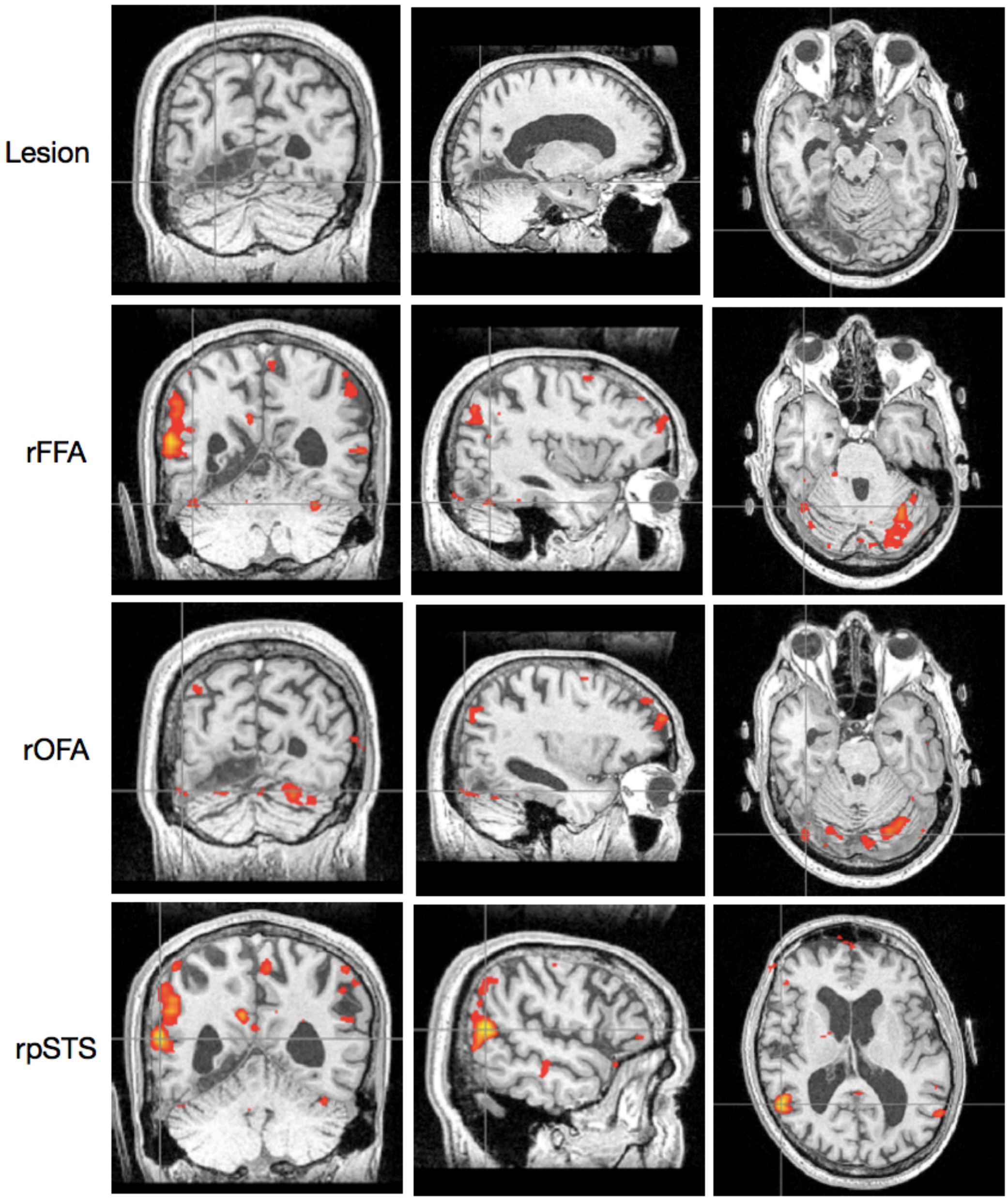
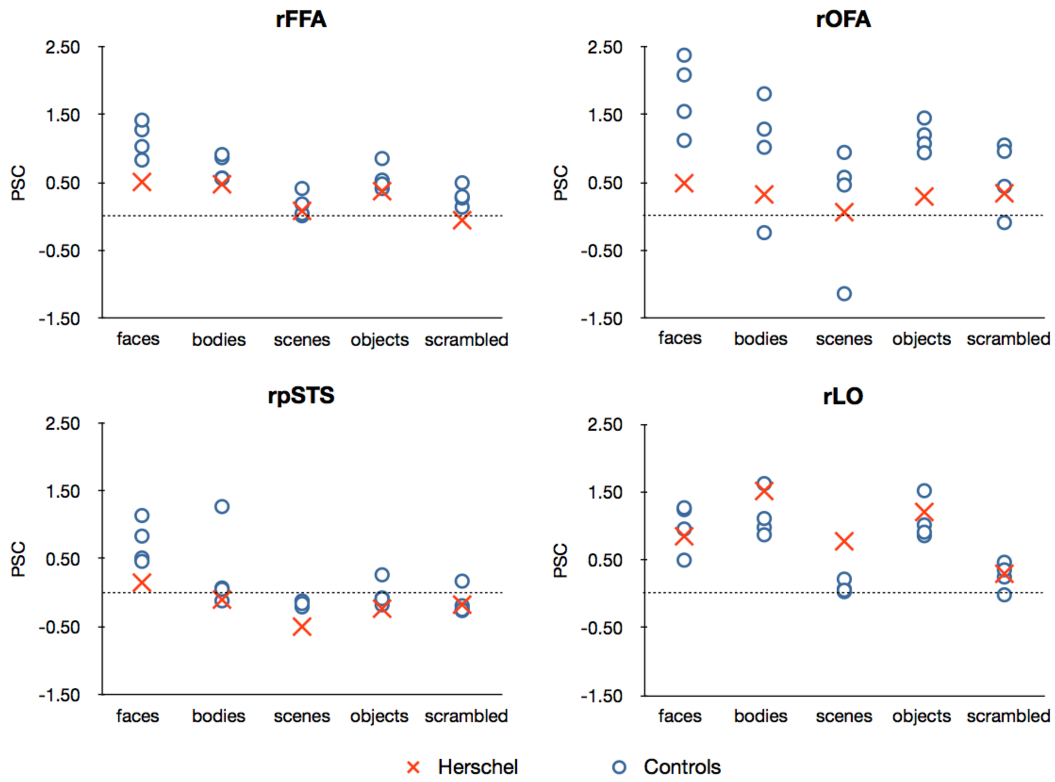


Figure 2.1. Structural and functional imaging of Herschel. Lesion location and activation of right fusiform face area (rFFA), right occipital face area (rOFA), and right posterior superior temporal sulcus (rpSTS). Images are shown in radiological orientation (right hemisphere on the left).



	Herschel	Control 1	Control 2	Control 3	Control 4
rFFA	42, -42, -28	42, -34, -25	42, -42, -28	38, -46, -20	40, -44, -18
rOFA	44, -80, -16	43, -82, -16	40, -80, -10	38, -88, -14	44, -82, -14
rpSTS	58, -52, 16	59, -41, 8	52, -48, -2	60, -36, 4	50, -42, 2

Figure 2.2. Magnitude of response (percentage signal change, PSC, from a fixation baseline) in the right hemisphere face and object areas to five stimulus categories (faces, bodies, scenes, objects, and scrambled objects). Below, MNI (Montreal Neurological Institute) coordinates for peak activations (during the face localizer) in the face areas. rFFA = right fusiform face area. rOFA = right occipital face area. rpSTS = right posterior superior temporal sulcus. rLO = right lateral occipital cortex.

GENERAL NEUROPSYCHOLOGICAL ASSESSMENT

We tested Herschel on his general intellectual and low-level visual abilities, to exclude them as possible causes of his reported difficulties with faces. Consistent with his reports, he scored within the normal range for all these tests (see Table 2.1 for detailed results). Herschel was in the 98th percentile for his age group at the Abbreviated Raven's Matrices, a test designed to measure abstract reasoning, and was in the high range for Digit Span Memory (scaled score 11), demonstrating his normal working memory. His language skills were also intact: he correctly recognized 57 out of the 61 words presented in the National Adult Reading Test (he did not know how to pronounce: *epergne*, *vivace*, *talipes*, *synecdoche*).

Herschel suffers from a full upper left and partial right quadrantanopia, but his low level vision is otherwise normal. His visual acuity, contrast sensitivity and color perception are all in the upper ranges. His visuospatial perceptual skills were assessed using the Birmingham Object Recognition Battery of tests (Riddoch & Humphreys, 1993), and he discriminated line length, size, orientation and position of gap normally.

A. General cognitive skills	Herschel	Max	
Abbreviated Ravens	11	12	98th percentile
National Adult Reading	57	61	high upper range

B. Digit Span Memory	Herschel	Controls *M*	Controls *SD*
Scaled score	11	10	3.0

C. Low-level vision	Herschel	Max/normal	
Visual acuity	20/19	20/20	normal
Contrast sensitivity (Michelson)	0.95%	0%	normal
Ishihara Color Perception Test	8	8	normal
Munsell Hue Text	23	0	upper range

D. Birmingham Object Recognition Battery	Herschel	Controls *M*	Controls *SD*
Length Match	28	26.9	1.6
Size Match	28	27.3	2.4
Orientation Match	27	24.8	2.6
Position of Gap	37	35.1	4.0

Table 2.1. General neuropsychological assessment of Herschel. He shows normal general cognitive skills (A and B) and normal low-level vision (C and D)

FACE PROCESSING ABILITIES

Next we performed a series of tests to experimentally confirm his deficits in recognizing facial identity and to determine whether he also had deficits with evaluations of facial expressions and facial gender. In all following experiments, significance of test scores differences between Herschel and controls was assessed using Crawford's modified t-test for single case studies (Crawford & Howell, 1998). Results are summarized in Figure 2.3.

Experiment 1: Famous face recognition

In a *Famous Faces* test (Duchaine & Nakayama, 2005), participants were presented with 60 photographs of people familiar to most Britons, cropped so that only faces are visible. Each face was displayed for 3s, and participants named the individual or provided uniquely identifying information. We compared Herschel to a middle-aged control group of 16 UK adults (*M* = 44.1 years).

Herschel identified only 3 out of the 60 faces presented, more than 7 standard deviations below controls (*M* = 47.3, *SD* = 6.2; t = 6.93, p < .001). To verify that the personalities presented were familiar to Herschel, we asked him afterwards which individuals were sufficiently known to him to allow recognition; he confirmed that he knew 48 of the famous people presented.

Experiment 2: face matching

The *Face Matching* task is designed to test the ability to match unfamiliar faces for identity across different viewpoints. In contrast to the Famous Faces test, performance does not rely on long-term memory. Participants were presented with a target face (frontal view) for 400 ms, followed immediately by three faces presented simultaneously as half-profiles for 2000 ms. Participants chose which one of the three test faces was the same individual as the target face. The stimuli were all male faces, with their hair completely covered by a standard black cap, so that judgments were based on facial features. Sixty trials with upright faces and 60 trials with inverted faces were randomly interleaved. Ten age-matched participants (6 female, age range: 47 – 61 years old, *M* = 53.6 years) provided control data.

Herschel's accuracy at matching upright faces was 41.8% (chance level 33.3%), or 2.8 standard deviations below controls (*M* = 78.7%, *SD* = 13.2%; t = 2.67, p = .026). He scored slightly below chance for inverted faces: 30.0% compared to the average 48.0% (*SD* = 13.1%; t = 1.31, p = .223) achieved by control participants. Herschel's inversion effect of 11.8% was the lowest from all participants and substantially lower than controls' average of 30.7% (SD = 14.0%), although the difference was not statistically significant (t = 1.29, p = .230).

Experiment 3: face perception

The results above demonstrate Herschel's problems with recognizing the identity of previously seen faces. It is possible that his pronounced prosopagnosia is due to an inability to remember faces, while his perception of faces is still normal. To test this possibility, we used the *Cambridge Face Perception Test* - CFPT (Duchaine, Germine, & Nakayama, 2007; Duchaine, Yovel, & Nakayama, 2007). In it, participants have one minute to sort six test faces according to their similarity to a target face. Test faces are morphed images containing different proportions of the target face - 28%, 40%, 52%, 64%, 76%, and 88%. There are 8 trials with upright faces and 8 trials with inverted faces, presented in a random order. The final accuracy score represents the percentage of correct sorts using the formula: 100 * [1 – (sum of deviations from the correct ordering for each trial / maximum deviations possible)]. Chance performance is 35.6%. The control data came from 21 age-matched participants (mean age = 46.5 years).

We tested Herschel twice with the CFPT, once in October 2009 and once in September 2011. As can be seen in Figure 2.3, his scores for upright faces were outside the normal range on both occasions: 55.6% and 33.3% respectively, whereas controls' mean score was 74.5% (*SD* = 8.5%; $t = 2.17$, $p = .042$; and $t = 4.74$, $p < .001$, respectively). These results indicate a perceptual component in Herschel's prosopagnosia. The lower score in the repeated test may be due to Herschel's acknowledgment of the difficulty of this test for him after the first session; as a result he might have not tried as hard the second time. Indeed, a comparison of the average time taken per trial (each trial had a time limit of one minute, but could be ended sooner) revealed that Herschel spent more time sorting the faces in the first session ($M = 44.9$s) compared to the second session ($M = 37.9$s; paired-samples t-test: $t(7) = 2.36$, $p = .051$). When asked to sort inverted faces, Herschel performed comparably to controls in the first session: 47.2% compared to 54.9% (*SD* = 6.8%; $t = 1.11$, $p = .282$), and significantly worse than controls in the second session: 37.5% ($t = 2.50$, $p = .021$). This drop in performance from session one to session two is consistent with a general decrease in attention/ effort during the second session.

Experiment 4 & 5: facial expression recognition

Many APs also have difficulty extracting other information from faces, such as expressions and age (Fox, Hanif, Iaria, Duchaine, & Barton, 2011; Sergent &

Signoret, 1992). We assessed Herschel's ability to recognize expressions with two tests. Control data were provided by nine age-matched male participants (age range: 41 – 55, mean age 46.8).

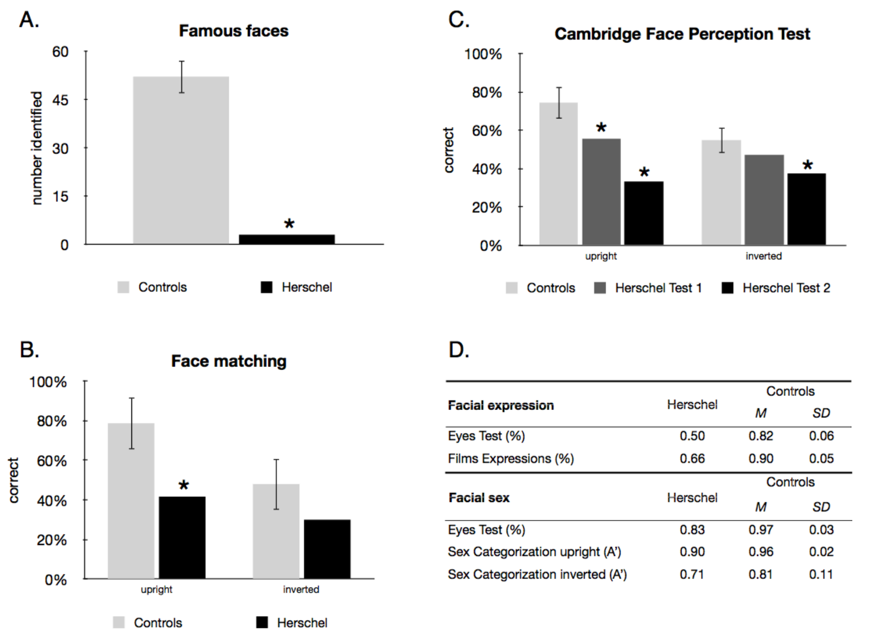
In the *Eyes Test* (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), participants are presented with the eye region of a face along with four emotion state words. In 36 trials, they choose the word that best describes what the person is thinking or feeling. Herschel proved to be impaired at this task, scoring more than 2 standard deviations below control data. He matched the eye region with the correct adjective in 50% of trials, whereas controls' average was 82.2% (*SD* = 6.4%; t = 4.77, p = .001).

The second task was the *Films Facial Expression Task*. In this task, participants are asked to identify subtle facial expressions captured from 18 obscure foreign films. In each of the 58 trials, participants see a word describing an emotional state, followed by three static images of the same actor/actress portraying different facial expressions. Images are presented for 500ms. Participants indicate which of the three images best matches the target emotional adjective (see Garrido, Furl et al., 2009, for more details). Herschel was severely impaired at this task too; he could recognize the expression in only 65.5% of the clips, while controls' average performance was 89.8% (*SD* = 5.4%; t = 4.27, p = .003).

Experiments 6 & 7: facial gender recognition

We assessed Herschel's ability to judge the sex of the faces with two tests. The first was the *Eyes test* described in the previous section (same control participants); after choosing the expression of each pair of eyes, participants also indicated the sex of the eyes. Herschel was correct on 83.3% of trials, which was significantly below the controls' average of 96.9% (*SD* = 2.8%; t = 4.61, p = .002).

The second was the *Sex Categorisation* task. In this task, 60 upright faces and 60 inverted faces are presented for 500 ms in a fixed, randomised order, and subjects have to categorise them as male or female. Faces were cropped below the eyebrows so participants had to rely on other information. Control data were provided by 9 age-matched participants (age range 52 – 59, mean age = 56.1 years; 5 female). Performance was measured with *A´* (Macmillan & Creelman, 1991), a bias-free measure that varies between 0.5 and 1.0 with higher scores indicating better discrimination ability between male and female faces. Herschel was mildly impaired with the upright faces (A' = 0.90, controls: A' = 0.96, SD = 0.02; t = 2.85, p = .022). His results were not significantly different from controls with inverted faces (A' = 0.71; controls: A' = 0.81, SD = 0.11; t = 0.86, p = .414).



Facial expression	Herschel	Controls *M*	Controls *SD*
Eyes Test (%)	0.50	0.82	0.06
Films Expressions (%)	0.66	0.90	0.05
Facial sex	Herschel	Controls *M*	Controls *SD*
Eyes Test (%)	0.83	0.97	0.03
Sex Categorization upright (A')	0.90	0.96	0.02
Sex Categorization inverted (A')	0.71	0.81	0.11

Figure 2.3. Herschel's face processing assessment. Herschel was severely impaired with: A) face recognition; B) face matching; C) face perception; D) inferring emotional and sex information from faces. Error bars show ±1 SD. Stars show significant differences (using Crawford's modified t-test) between performance of Herschel and controls.

SPECIFICITY OF HERSCHEL'S PROSOPAGNOSIA: FACES VS. OBJECTS

The results of the experiments above demonstrate that Herschel has severe face processing problems and suggest they result from deficits to high-level visual mechanisms. We next examine whether these deficits affect only face processing or extend to other objects.

According to an influential view of object recognition (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), objects are first categorized at a basic level (e.g. a car, a chair) and then identified as a specific exemplar from within that category (e.g. my car). We assessed Herschel's basic level object recognition with a reduced set of the Snodgrass and Vanderwart (1980) image set and his within-level object recognition with three other tests.

Experiment 8: Basic-level object recognition

For most people, identifying objects at a basic level happens instantaneously and effortlessly. Thus, it is perhaps not surprising that previous tests of basic object recognition suffered from ceiling effects, with normal participants consistently achieving maximum scores. These tests may be useful for detecting pathological performance, but unfortunately they are not suitable for uncovering finer deficits, potentially affecting individuals that do not report object recognition problems in everyday life, such as Herschel. We therefore created a more sensitive test for basic object identification to more subtly probe the functioning and organization of Herschel's visual recognition system.

Basic object recognition is often assessed using the Snodgrass and Vanderwart (1980) picture set. The set includes 260 black and white line drawings of diverse common objects. Based on the norms published in Snodgrass and Vanderwart (1980), we selected a subset of 82 pictures, half of which were living (e.g. zebra, arm, leaf) and half of which were nonliving (e.g. umbrella, glass, chair) objects. Familiarity and complexity were matched across the two groups. To increase the difficulty of this simple recognition task, we presented each stimulus for only 50ms, followed by a pattern mask. After the presentation, participants had unlimited time to name the object. Seven age-matched participants (4 female, age range: 51 - 66 years, M = 56.7 years old) provided control data.

Herschel correctly identified 72.0% of the objects presented, which was slightly below the average score for controls (M = 77.0%, SD = 12.7%, range: 52.4% - 93.9%; the difference was not significant: t = 0.37, p = .725). A closer examination of Herschel's results revealed that a disproportionate number of errors were made for the 15 images depicting mammals, of which he identified only 7 (his 46.7% performance was below normal range: controls' M = 81.9%, SD = 14.3%, range: 53.3% - 93.3%; t = 2.30, p = .061), while for the rest of the images his performance was in line with that of controls (77.6% compared to 75.9%, *SD* = 12.7%, range 52.2% - 94.0%). Furthermore, Herschel was the only one to score lower for mammals (46.7%) compared to non-mammals (77.6%), while controls scored on average 6.0% better with mammals than non-mammals. Interestingly, for most missed items, he seemed to correctly identify the stimuli at a superordinate level (as mammals), but mistook, for example, a bear for a cat, a deer for a cow, a donkey for a dog, and a zebra for a horse. These results and Herschel's low score on the horse old-new test (but still within normal range, see next experiment) raise the possibility that Herschel has a deficit with a particular type of object – mammals, even though he is normal with a wide range of other objects.

Basic level object recognition (e.g. fork vs. spoon) is generally considered to be easier than face recognition and thus, a difference in recognition abilities for faces versus basic objects in acquired prosopagnosia may simply reflect a difference in cognitive demands rather than face specificity. It has been argued (e.g. Damasio et al., 1982) that because face recognition involves discriminating highly similar exemplars within the same category (face A vs. face B), it must be compared with discrimination of other objects within the same category (e.g. car A vs. car B).

Experiment 9: old-new discrimination of faces vs. non-face objects

In the *Old-New Recognition Memory Test* (Duchaine & Nakayama, 2005), participants first see 10 target items from within the same object category, with each item presented twice. They are then presented with 50 items, 20 targets (10 x 2) and 30 nontargets and must discriminate between targets (old) and nontargets (new). We tested Herschel with nine different old-new tests: cars, horses, houses, tools, natural landscapes, sunglasses, guns, and two separate face tests (for details about the stimuli see Duchaine & Nakayama, 2005). Control data were provided by nine age-matched participants (age range 52 – 59, mean age = 56.1 years; 5 female).

As can be seen in Figure 2.4, Herschel was severely impaired with Faces 1 [*A'* = .85 compared to average *A'* for controls: *M* = .95, *SD* = .03; t = 3.16, p = .013) and Faces 2 (*A'* = .68 compared to average *A'* for controls: *M* = .95, *SD* = .04; t = 6.40, p < .001). With horses, he performed at the lower end of the normal range (*A'* = .86 compared to average *A'* for controls = .94, *SD* = .04; t = 1.66, p = .135). For all the other categories, Herschel's performance was normal (all *p*s > .31). His normal performance with objects could not be attributed to speed/accuracy trade-offs (see Figure 2.4).

Experiment 10, 11 & 12: memory for faces vs. hairstyles vs. cars

The Cambridge Face Memory Test (CFMT) was developed to detect prosopagnosia (Duchaine & Nakayama, 2006) and has been shown to have good psychometric properties (Bowles et al., 2009; Wilmer et al., 2010). In this test, participants study six target faces and then must recognize them in 72 three-option forced-choice trials. The trials vary in difficulty, with the three images presenting faces in views and lighting conditions that are the same or different to those studied. Some trials also had images with visual noise added. A score of 24 represents chance level performance. Our control participants were 20 age-matched individuals (average age = 45.1 years). Herschel had severe difficulties with this task, scoring well outside the normal range: 31 correct responses compared to controls' average score of 59.6 (*SD* = 7.6, see Figure 2.4; t = 3.67, p = .002).

Using the same procedure as the CFMT, we tested Herschel with two other non-face stimulus classes: hairstyles and cars (Figure 2.4). The *Cambridge Hair Memory Test* (CHMT) presented male hairstyles cut out from head shots from the same image set which the faces used in the CFMT were drawn from. Like the CFMT, the 72 test items presented the hairstyles in views and lighting that were the same and different to those studied and with and without added noise. Control participants were 20 undergraduate students from Dartmouth College (15 female, age range 18 - 27 years old, *M* = 19.5 years, *SD* = 2.1). Herschel's score was comparable to those of the young controls: 44 versus a control average of 50.85 (*SD* = 6.05; t = 1.10, p = .283).

The *Cambridge Car Memory Test* (CCMT) replaced faces with cars (Dennett et al., 2012). Because males were found to score higher than females on this task, we compared Herschel's performance to that of the 60 males from a larger mixed pool of young adults (age range 18 - 35 years, $M = 20.63$, $SD = 2.88$). As with hairstyles, Herschel's score with cars was well within the normal range; he scored 54, only slightly below the control mean ($M = 57.43$, $SD = 8.31$; $t = 0.41$, $p = .684$).

Experiment 13, 14 & 15: matching for faces vs. matching for bodies and objects

The previous experiments indicate that Herschel has a selective impairment in remembering previously learned faces. We next tested Herschel on three matching tasks with more limited memory demands (Pitcher et al., 2009). In these tasks participants see a target image for 500ms, followed by a mosaic mask for 250ms and a test image for 500ms. Participants decide whether the test and the target images are the same or different (50% chance level). To avoid matching based on low-level visual cues, the test image was displayed in a slightly different position on the screen than the target.

Herschel completed one matching test for bodies, one for objects and one for faces (in that order). FantaMorph software (Abrosoft, 2002) was used to make ten morph series between 20 pairs of stimuli for each category. Pairs of images were then drawn from the morph series to create 80 unique experimental trials for each category (40 same, 40 different). Within each block, the trial order was randomized. The original faces were created using FaceGen Modeller software (Singular Inversions, 2008; http://www.facegen.com); the bodies were created using Poser software (Smith Micro, 2009; http://poser.smithmicro.com); the objects were selected from a novel objects set available on Michael Tarr's website (http://stims.cnbc.cmu.edu/Image%20Databases/TarrLab/Novel%20Objects/). More details on the stimuli are provided in Pitcher et al. (2009). Control data came from 8 age-matched participants (5 female, age range: 52 – 59, mean age = 56.0).

As can be seen in Figure 2.4, Herschel scored close to chance level for faces (53.8% compared to controls $M = 74.8\%$, $SD = 5.2\%$; $t = 3.81$; $p = .007$), but within normal range for bodies (73.8% compared to controls *M* = 73.1%, *SD* = 7.9%; t = 0.08) and objects (81.3% compared to controls *M* = 78.6%, *SD* = 7.3%; t = 0.35).

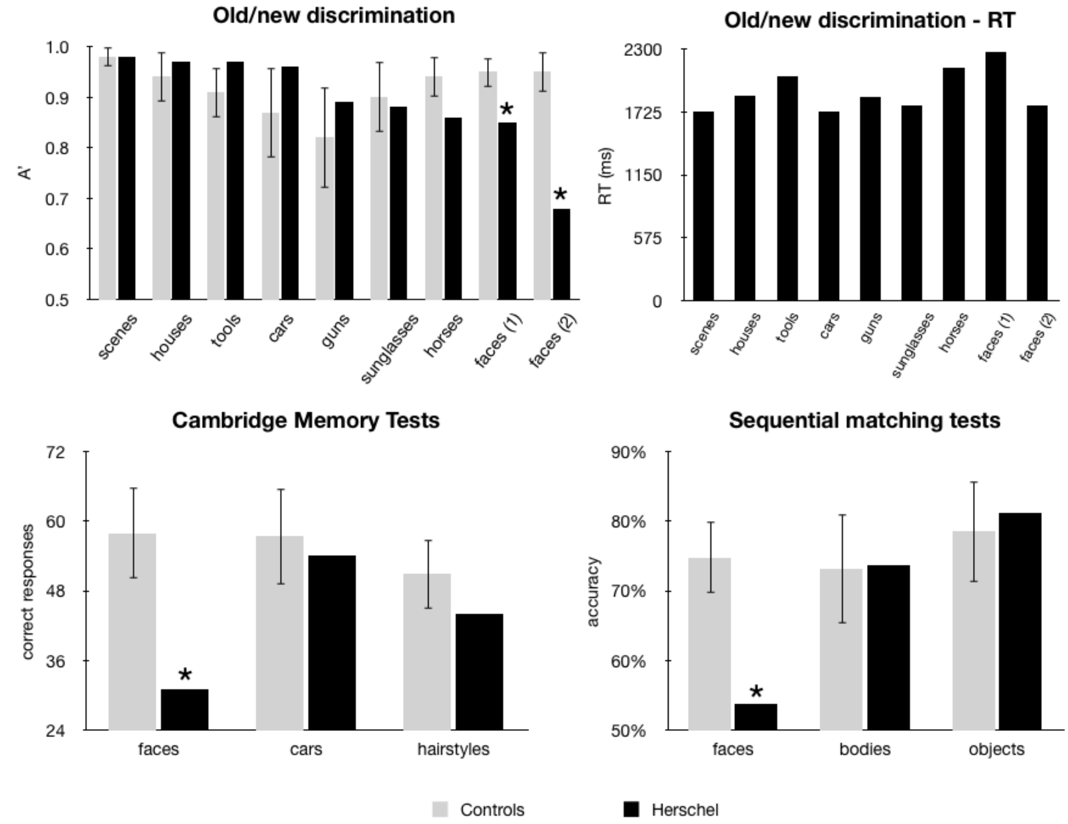


Figure 2.4. Specificity of Herschel's prosopagnosia. He was markedly impaired with faces but normal with non-face stimuli in tests of learning, memory and sequential matching. Error bars represent ± 1 SD. Stars show significant differences using Crawford's modified t-test.

The results in this section indicate that Herschel's high-level visual problems are largely restricted to faces. He performed normally with other object categories in recognition memory tests and sequential matching tasks. He seemed to have difficulties only with basic level recognition of four-legged animals.

GENERAL AND FACE CONFIGURAL PROCESSING

Abnormal visual general configural processing has been advanced as one of the possible causes of prosopagnosia (Levine & Calvanio, 1989). Indeed, many APs failed tests of visual closure (e.g. Levine & Calvanio, 1989; Bauer & Trobe, 1984; Lê et al., 2002; De Renzi, 1986a; De Renzi, 1986b; De Renzi et al., 1991) or other tests used to assess configural processing, such as the Navon hierarchical letters (e.g. Behrmann & Kimchi, 2003). However, most of these patients were also severely impaired with object recognition, making it difficult to know whether their results were due to a specific deficit in configural processing or to general object agnosia. Other researchers (Busigny, Joubert et al., 2010) suggested that, while configural processing of non-face items may be spared, configural processing of faces is impaired in all cases of AP.

Visual closure tests are recognition tests that require the ability to reconstruct the whole from incomplete parts or shape information. This ability was considered by Levine and Calvanio (1989) to rely on general configural processing. All tests of visual closure present degraded images of objects or words, either by adding visual noise or deleting essential parts. Herschel was tested on four such tests, with a modified presentation procedure involving brief presentations so that the tasks depended more on normal recognition processes and less on the slow, visual problem-solving processes that can be used when stimuli are presented for long durations (Farah, 2004). Basic level recognition (Rosch et al., 1976) was required in all tests, and given Herschel's good performance with most within-class object tests we expected he would perform normally in the tests of visual closure. In addition, we tested Herschel's configural processing of faces with the composite test.

Experiment 16, 17 & 18: visual closure

The following three tasks are adapted from the Kit of Factor Referenced Cognitive Tests (Ekstrom, French, & Harman, 1976), previously used by Levine and Calvanio (1989) and Duchaine (2000) to assess visual closure in prosopagnosics. Pilot studies were conducted to remove the stimuli that were either always or never correctly identified. For the modified tests, the control participants were 9 age-matched participants (6 female, age range 47 - 61 years, *M* = 53.0 years old). Two participants who were not native English-speakers (both male, age 47 and 48) were not run on the words test.

Modified Gestalt Completion Task (MGCT). The Gestalt Completion Task involves identifying a common object from a group of black blotches created by erasing parts of the object (see Figure 2.5A for examples of stimuli from all visual closure tests; images from MGCT are very similar to those from the Street test; Street, 1931). Each object was shown for 500ms followed by a pattern mask for 250ms, after which participants were required to say the name of the object. The original test had 20 items, but after the pilot test we kept only 16 for the main experiment. Herschel was severely impaired at this task; he was not able to identify even one object, while controls averaged 6.2 (*SD* = 2.5; range: 4 – 12; t = 2.35, p = .046).

Modified Concealed Words Task (MCWT). The Concealed Words Task is similar to the above test, except objects are replaced with words: participants must identify common words from their fragments. The occluded words were shown for 1500ms, followed by a pattern mask for 250ms, and after the pilot study we kept 28 of the original 50 stimuli. Again, Herschel was severely impaired, recognizing only three words; his score was more than 3 standard deviations below controls (*M* = 14.9, *SD* = 3.9; range: 10 – 20; t = 2.85, p = .029).

Modified Snowy Pictures Task (MSPT). In this task participants had to identify a line drawing of an object degraded by a snow-like pattern. The snowy pictures were shown for 1000ms followed by a pattern mask for 250ms, then a blank screen for 500ms. There were 24 items in the original version of this test and we used 13 of the items after the results of the pilot study. Herschel could name the objects presented in only two snowy pictures, while controls managed to recognize an average of 6.0 objects (*SD* = 2.7; range: 2 – 9; t = 1.41, p = .198).

Experiment 19: blurred objects

We created a Modified version of the Blurred Pictures Task (MBPT) presented in Viggiano, Costantini, Vannucci, & Righi (2004). The original set of stimuli consists of 62 basic level objects, each object with 10 images varying in contrast, from extremely blurry to crystal clear. The test trials starts with the most blurred image of an object for 250 ms, then the second most blurred image for 250 ms and so on, until the participant correctly identifies the object or all 10 images are displayed. After that, the images of the next object are presented. The onset of each image is controlled by participants and participants have unlimited time to provide an answer. The score is calculated by adding the number of images displayed before each object is recognized; higher scores represent worse recognition performance. In a pilot study, we used the same procedure as in Experiment 16-18 to select 27 stimuli for our main experiment. The control data came from 7 age-matched participants (6 female, age range 48 - 61 years, $M = 54.0$ years old). Herschel needed a total of 134 images to recognize the selected objects, while controls needed on average 90.6 images ($SD = 15.7$; range: 73 – 109; $t = 2.59$, $p = .041$).

Together (see Figure 2.5B), Herschel's scores suggest an impaired ability to recognize basic objects from impoverished visual stimuli, in contrast to his normal performance when nonliving objects were presented unobstructed. His visual recognition system appears to be impaired at inferring the actual form of objects when this information is incomplete.

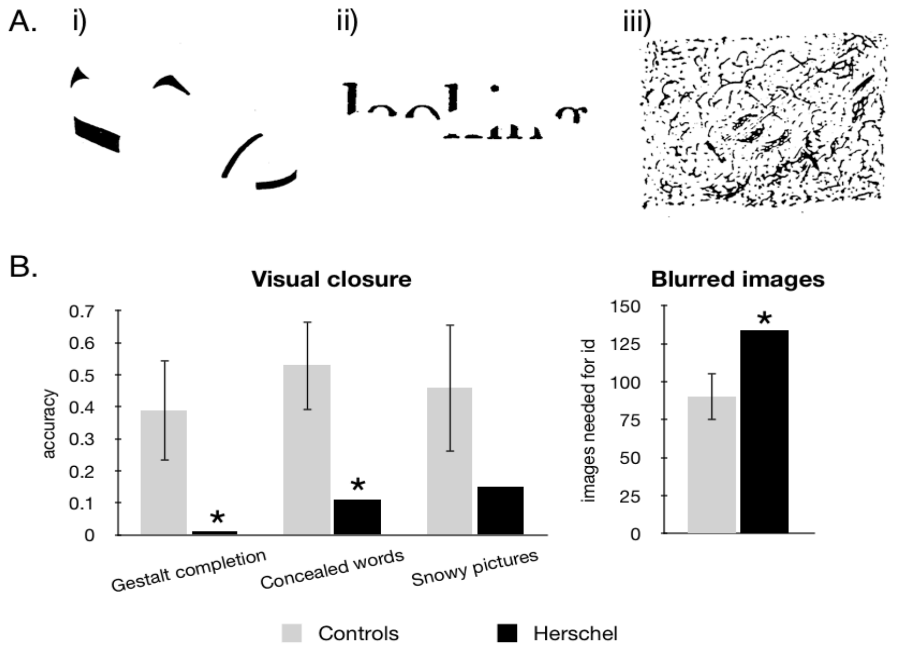


Figure 2.5. Herschel's performance at visual closure. A) Examples of stimuli presented in the visual closure tests: i) MGCT, ii) MCWT, iii) MSPT. The objects/word depicted are: shoe, looking and guitar. B) Herschel was impaired at two tests of visual closure and in the lower range at the other. He was also impaired at the Blurred images test. Error bars represent ± 1 SD.

Experiment 20: Navon hierarchical letters

We also tested Herschel's configural processing using a Navon task (Navon, 1977). The Navon hierarchical letters are compound letters consisting of a number of small capital Ss or Hs (local letters) configured to form either a global S or H (Figure 2.6A). The letters can be consistent (the global and local letters are identical) or inconsistent (the global and local letters are different). When asked to identify either the local or global letters, participants typically show an advantage for global processing (faster identification of global letters) and an interference effect (slower identification when global and local letters are inconsistent) (Navon, 1977; Behrmann, Avidan, Marotta, & Kimchi, 2005).

Herschel and 14 control participants (average age = 41.7 years) were tested in two back-to-back sessions, and the order of the blocks within a session was local, global, global and local for a total of 384 trials. The local and global letters were consistent on half the trials and inconsistent on the other half (Figure 2.6A). Herschel performed normally in this task (Figure 2.6B). He was on average faster for global than for local trials, and he showed the typical global interference effect (he also showed a large local interference effect). He was 100% accurate for all trials (controls averaged 98%). These results suggest Herschel does not have a global processing deficit.

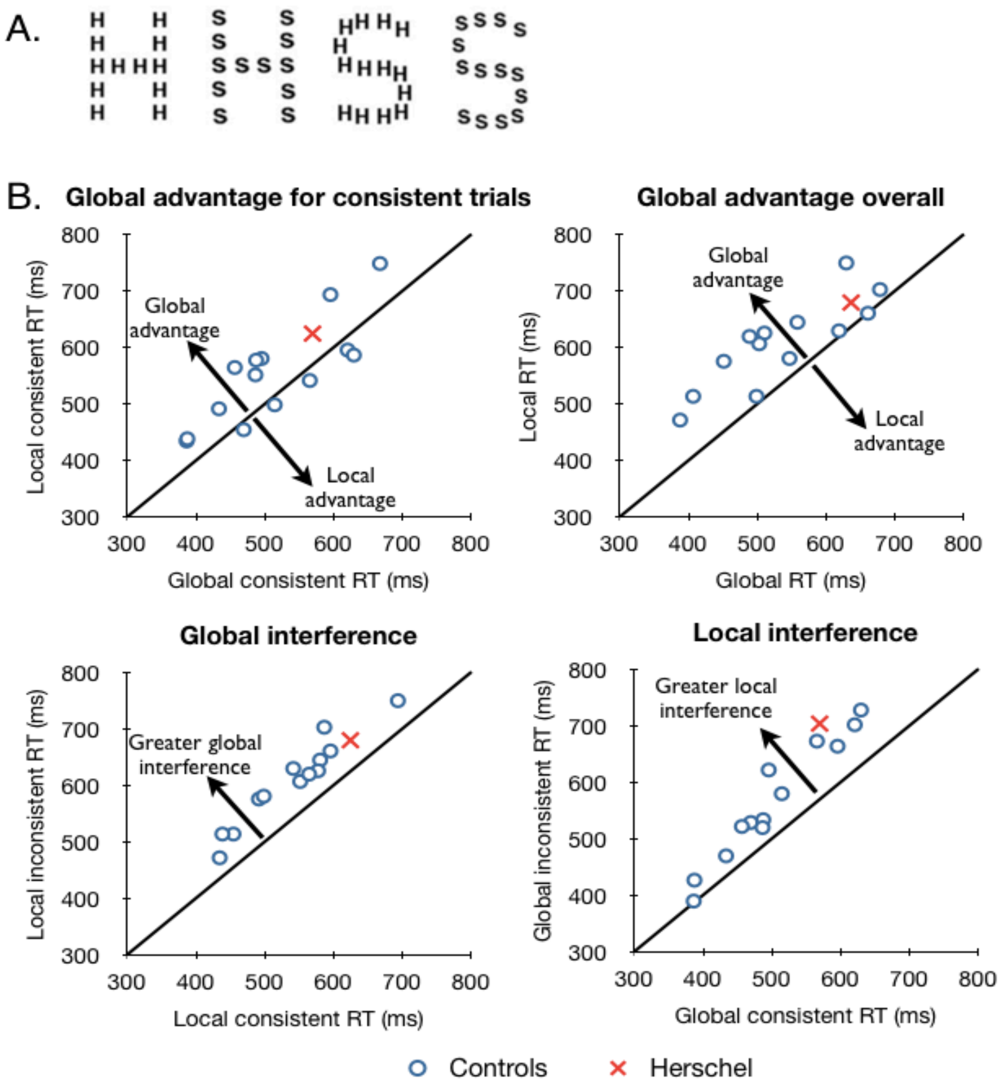


Figure 2.6. Herschel's configural processing in the Navon task. A) The compound stimuli used in the global-local Navon task. B) Individual RTs (averaged across the two sessions) of controls and Herschel on the global – local task. Each circle represents one control participant. Herschel is represented by a red cross. The diagonals in the top figures separate participants who showed a global advantage from those who showed a local advantage; the larger the distance to the diagonal the higher the respective advantage. The diagonals in the bottom figures represent points for which there is no global/local interference. The larger the distance to the diagonal the higher the global/local interference. Herschel showed the largest local interference, but it was not significantly different from controls and his global interference and global advantage were comparable to those of controls.

Experiment 21: Composite faces

Herschel's configural face processing was assessed with the composite faces test (Young et al., 1987). The version we used was adapted from Susilo et al. (2010). Composite stimuli were created by combining top halves and bottom halves of different faces. Participants were presented pairs of composites sequentially (first stimulus for 300ms, blank screen for 400ms, second stimulus for 300ms) and were asked to say whether the top halves matched. Half of all trials presented aligned composites, with the top halves and bottom halves neatly arranged to form new faces, and the other half presented misaligned composites, with the top halves shifted to one side (left or right) compared to the bottom halves. Holistic face processing should make trials with aligned faces more difficult because of automatic processing of bottom halves. The bottom halves in each pair were always different, while the top halves were the same (60 trials) or different (30 trials). We analyzed only 'same' trials because 'different' trials are difficult to interpret (Robbins & McKone, 2007). The composite effect (indicating holistic face processing) was measured as the difference in performance between misaligned and aligned 'same' trials.

Herschel was correct on 90% of the misaligned 'same' trials and on 70% of the aligned 'same' trials. Seven controls (age range: 49 – 56 years old, mean age 54.3, 4 female) averaged 95.2% and 73.3%, respectively. Herschel's composite effect of 20% was not significantly different from controls' average composite effect of 21.9% ($t = 0.11$, $p = .919$), indicating normal configural/holistic face processing.

DISCUSSION

We have presented a new case of acquired prosopagnosia that displays an interesting and surprising cognitive profile. Herschel is a 56-year old British man with a degree in astronomy. Following two strokes, he suffered extensive lesions in the occipitotemporal cortex, especially severe in the right occipital cortex. Despite his intellect, he was unable to recognize famous faces, remember faces previously shown to him, match or order faces based on similarity. Herschel was also impaired with facial expressions and facial gender judgments. In contrast, he showed normal memory for a wide variety of objects in several paradigms and normal ability to discriminate between highly similar items within a novel object category; he was also successful with fine discrimination between human bodies. Furthermore, he was normal at recognizing basic-level objects (with the exception of mammals) in brief presentations when the images were intact. Interestingly though, when visual noise was added or parts information was removed from the images, Herschel had substantial difficulties identifying objects at the basic level. His general global processing (measured with the Navon task) and his holistic face processing (measured with the composite test) were normal.

Normal within-class object recognition in prosopagnosia

Faces play a vital role in our daily interactions, and considerable evidence indicates the brain contains mechanisms dedicated to face recognition (Duchaine et al., 2006; Farah, 1996; Kanwisher et al., 1997; McNeil & Warrington, 1993; Moscovitch et al., 1997; Tsao et al., 2006). However, not all researchers share this view. One alternative possibility is that the brain uses the same mechanisms to recognize both faces and objects (Damasio et al., 1982). This view predicts that prosopagnosics will show impairments with any non-face object task of comparable complexity and within-class similarity to faces.

Herschel's lesions appear to have selectively affected mechanisms used for face processing, leaving object recognition largely intact. Experiments 9-15 showed a clear dissociation between performance with faces and other objects. Herschel recognized unfamiliar cars, houses, horses, tools, sunglasses and guns normally, but he was severely impaired with faces. He scored within the normal range in challenging tasks measuring the ability to distinguish between highly similar members of complex natural classes (cars and hairstyles). His within-class discrimination of bodies and novel objects created to match faces for complexity and similarity was also normal, in stark contrast with his face skills. Herschel demonstrated these dissociations in a variety of paradigms: the Cambridge Memory Tests were memory-based, while the matching tests were perceptual in nature. These results suggest his severe prosopagnosia is not associated with deficits relating to within-level recognition of non-face objects.

Herschel's results mirror results from other patients with acquired prosopagnosia who could learn and/or discriminate between visually similar exemplars of various, complex non-face objects. PS was normal at discriminating between highly similar novel shapes and between exemplars of common object classes - such as cars, dogs, cups, shoes - parametrically manipulated for similarity (Busigny, Graf et al. 2010). GG performed in the normal range for within-category discrimination of birds, boats, cars and chairs (Busigny, Joubert et al., 2010). The case of WJ reported by McNeil and Warrington (1993) is also quite remarkable; following the stroke that led to severe prosopagnosia, he became a farmer and could accurately identify his own sheep and learn unfamiliar sheep from face photographs. The normal performance at fine-grained discrimination of non-face objects or non-human faces displayed by these cases support the specificity of mechanisms implicated in human face processing.

Normal basic-level recognition of nonliving objects

As noted in the introduction, it is possible that the normal basic-level object recognition reported for many cases of prosopagnosia in the literature was due to ceiling effects in basic-level tests. To avoid ceiling effects, Herschel was asked to identify objects from line drawings presented for only 50 ms, which made the task challenging even for control participants. Herschel's performance was in line with controls, with the exception of mammals, of which he identified only 47% (controls identified 82% of mammals on average). His results show a striking resemblance to patient RM (Sergent & Signoret, 1992), who was also severely impaired at recognizing 'feline animals' (50% success rate) while being perfectly normal with other objects (success rate 96%). Selective deficits with living but not with nonliving objects have been noted to co-occur with prosopagnosia before (LH: Levine & Calvanio, 1989; Farah, McMullen, & Meyer, 1991; MB: Farah et al., 1991). The living-nonliving distinction has also been reported in many cases with semantic deficits (Caramazza & Shelton, 1998; Pillon & d'Honincthun, 2011; Warrington & Shallice, 1984), but Herschel's deficits may be due to category-selective visual problems. The existence of such a visual deficit would fit well with functional imaging studies demonstrating that lateral regions of the ventral visual stream show a stronger response to living objects compared to tools (Chao, Haxby, & Martin, 1999; Noppeney, Price, Penny, & Friston, 2006). Future testing of his visual and semantic abilities with living objects will address this possibility.

Configural processing in prosopagnosia – questions and an interesting hypothesis

Another goal of our study was to examine Levine and Calvanio's (1989) suggestion that prosopagnosia is generated by a deficit in general-purpose configural processing. Face perception has been shown to rely on configural processing, but this could be face-specific configural processing or configural processing applied to many visual categories (Farah, 2004; Behrmann, Avidan et al., 2005; Levine & Calvanio, 1989). Here we tested Herschel on three modified visual closure tasks used by Levine and Calvanio (1989) to measure general configural processing, which they defined as the ability to represent and recognize an object based on overall shape rather than the individual features. We also added a fourth test looking at Herschel's ability to recognize objects from blurred images. His configural face processing was examined with the composite faces test.

Herschel's normal performance on most within-category object tests led us to believe that he would score normally on the tests of visual closure, but he proved to be severely impaired in all tests: with objects, as well as with words; with images of objects occluded, obscured by visual noise or blurred (all his scores were significantly different from controls, with the exception of Snowy Pictures; in this test he recognized only two images). His marked impairment at visual closure mirrors that of patient LH (Levine & Calvanio, 1989) and several other prosopagnosics (De Renzi, Faglioni, & Spinnler, 1968; Benton & van Allen, 1972; Bauer & Trobe, 1984). All these seem to point towards a general impairment with configural processing, consistent with Levine and Calvanio's (1989) hypothesis. However, no configural processing deficit was apparent in the Navon task; Herschel showed typical global advantage and global interference effects, challenging the generality of his configural problems. His normal Navon results mirror those of PS, a well-documented case of acquired prosopagnosia with face-selective deficits (Busigny & Rossion, 2011). One interpretation of this dissociation in performance on tests thought to measure configural processing is that there are two types of general configural processing: one that requires integration of visible local features and the other that requires integration of visible and occluded local features. In one case, there is no physical difference between the information available for the global versus local constructs; in the other case, the global construct requires filling-in.

While Herschel's prosopagnosia does not seem to be a result of deficits with general configural processing, it could still be linked to a face-specific configural processing deficit. However, Herschel displayed a normal sized composite face effect, providing preliminary evidence that this is not the case. Further studies are needed to more firmly establish this finding as it would be inconsistent with a suggestion in Busigny, Joubert et al. (2010) that AP cannot occur without abnormal holistic face processing.

It is conceivable that Herschel's poor results at the visual closure tests may be due to his left and right quadrantanopia; damaged upper visual fields may have caused Herschel to see fewer parts of the objects when they were presented for 250 ms. Considering that available visual information had already been reduced to the minimum necessary to allow recognition, further stimulus loss may have made the tasks extremely challenging for Herschel because of low-level deficits rather than recognition deficits. However, we believe this is not the case; Herschel was still unable to recognize the objects from the Gestalt Completion test and the Snowy Pictures test when we allowed him unlimited viewing time, which presumably could compensate for his visual field defect. He could not 'see' the objects in the images even after we named them.

Another possibility is that the type of configural processing deficits that Herschel has (and that is apparent only in visual closure tests) disproportionately affects perception of living, dynamic objects, such as faces and animals (for which Herschel had different degrees of impairment). This may occur because Herschel cannot integrate the non-nameable features that characterise living things into a coherent whole (Levine & Calvanio, 1989) or because of other aspects specific to the perception of living things. For example, constructing a flexible representation necessary to accommodate the dynamics of the second-order spatial relations between the features of a particular living item (e.g. a face must be recognized even when facial expressions modify the face parts and their spatial layout) may require some ability to correct the representation so it can be matched to memory representations. The absence of this ability would compromise recognition of living items, but would not be noticeable in normal object recognition tests, because no correction is required. Spatial relations between individual features of a nonliving object are expected to stay the same and thus a static mental representation (allowing for differences in orientation, viewpoint etc.) would be fine. The deficit would become apparent only when identification depends on reconstructing the spatial configuration of individual parts from incomplete images, as is the case in visual closure tests. It is also possible that Herschel's lesions disrupted lower-level processes only required when any stimulus is degraded, regardless of its animacy.

Herschel's abnormal results in visual closure tests suggest impairments in the system responsible for object perception, impairments which were not apparent in object recognition tests with unaltered images of non-mammal objects (experiment 8). This raises an interesting possibility: within-class recognition or discrimination (for which Herschel was normal, see experiments 9, 11, 12, 15) may not require intact basic level recognition. In other words, visual recognition of objects appears not to follow the hierarchy implied by models of object processing (e.g. Rosch et al., 1976), with successful within-level discrimination necessarily dependent on normal basic level recognition stage. Perhaps only certain aspects of basic level recognition are critical for the within-level recognition, which may still function normally despite impairments affecting basic level recognition. Farah, Levinson and Klein (1995) argued that prosopagnosic LH, impaired with visual closure (Levine & Calvanio, 1989) and basic object recognition (Levine et al., 1980), similarly showed normal within-level discrimination. However, LH was tested with only one category of non-face objects (glasses) and floor effects might have masked subtle impairments (he scored 63% with glasses while controls averaged 69%). In contrast, Herschel showed substantially better recognition of houses, horses, tools, sunglasses, guns (experiment 9), cars (experiments 9 and 12), hairstyles (experiment 11), bodies and novel objects (experiment 15) than of faces (experiments 9, 10, 13) in tests whose difficulty levels for faces and objects were well matched. Therefore, we believe Herschel is the first well-documented case to have intact within-object with compromised basic level recognition.

CONCLUSION

The marked contrast in Herschel's performance with faces and non-face objects represents additional support for at least partly distinct processing for faces. While Herschel had difficulties in recognizing objects at a basic level under difficult conditions, he was very good at learning and discriminating between similar exemplars of several non-face object classes. His basic level recognition of objects was also fine with the notable exception of mammals. Herschel's prosopagnosia does not appear to relate to a general or face-specific configural processing; even though Herschel failed to "see" objects beyond the sum of their (incomplete) individual parts in the visual closure tests, he showed normal configural processing in the Navon task and in the composite faces test. His atypical functional activation across the whole cortical face-selective network precludes us from linking his deficits to a particular face area.

2.2. NORMAL ACQUISITION OF EXPERTISE WITH A NOVEL OBJECT CLASS IN TWO CASES OF ACQUIRED PROSOPAGNOSIA

Introduction

Cognitive neuroscientists generally agree that the visual mechanisms involved in face recognition are different from those involved in most other types of object recognition. However, the question of how specialised these mechanisms are is a long-running debate. Are the mechanisms specific to faces, or are they also engaged by objects sharing certain properties with faces? The expertise hypothesis (Diamond & Carey, 1986; Gauthier & Tarr, 1997) argues for the latter possibility by suggesting that the mechanisms involved in face processing operate on all objects with which an observer has had extensive experience.

Although some studies have examined experts with real objects (Diamond & Carey, 1986; Tanaka & Curran, 2001; Busey & Vanderkolk, 2005), much of the research arguing for the expertise view has involved participants trained in the laboratory to become experts with “greebles”, an artificial class of objects designed to place face-like demands on recognition mechanisms (Gauthier & Tarr, 1997; Gauthier et al., 1998). The greeble training procedure is relatively fast and simple, which makes it an attractive method to investigate the acquisition of expertise. After seven to ten sessions of learning to identify individual greebles as well as the families or gender of the greebles, most participants become proficient at recognizing the greebles and reach the criterion claimed to indicate expertise - response times for recognizing individual greebles which are comparable to response times for recognizing greeble families or gender (Gauthier & Tarr, 1997). Importantly, studies with greebles allow participants to be tested before and after becoming experts. Several studies claimed that greeble processing elicited face-like perceptual (Gauthier & Tarr, 1997; Gauthier et al., 1999; Tarr, 2003) and neural (Gauthier et al., 2000) effects after but not before training, therefore challenging the face-specificity hypothesis.

A fundamental prediction of the expertise account of face processing is that individuals with severe face recognition deficits should also be impaired at acquiring expertise with other objects, such as greebles. We tested this prediction in two cases of acquired prosopagnosia, Florence and Herschel. Herschel's results are doubly relevant to the expertise account because he has atypical functioning of the fusiform face area (FFA), an area selectively responsive to faces but which has been claimed to also mediate expertise more generally - including greeble expertise (Gauthier et al., 1999; Gauthier & Tarr, 2002; McGugin et al., 2012).

Florence and Herschel were trained in a standard greeble procedure (Duchaine et al., 2004; Gauthier & Tarr, 2002; Rossion et al., 2002), involving eight sessions during which participants learned and were tested on 20 greebles belonging to five families. To ensure that this procedure elicited impaired performance in Florence and Herschel when faces rather than greebles were used, we also created a similar training procedure with computer-generated faces which was matched for difficulty with the greeble training.

Materials and Methods

Prosopagnosic participants

Florence is a right-handed female nurse from Canada who was 29 years old at the time of testing (born 1982). She was previously described in Fox et al. (2011) as R-AT1. In 2006 she became prosopagnosic after a resection of her right amygdala and right hippocampus to control epilepsy. Neuropsychological assessment and functional MRI were conducted in 2007 (see Fox et al., 2011, for details). Florence performed normally on a battery of cognitive and visual tests. Despite her face impairments, a static face localizer (faces - objects) revealed face activation bilaterally in all face-selective areas: fusiform face area (FFA), occipital face area (OFA) and superior temporal sulcus (STS). In 2008 she underwent a second operation that removed most of her right anterior temporal lobe, sparing the areas previously found to be face-selective (Figure 2.7). Florence has noted no visual changes since her second surgery and her normal performance on tasks described below suggest that her early visual processes were unaffected by the second procedure.

Herschel is a right-handed British male with a degree in astronomy. He was 55 years old at the time of testing (born 1956). In 2008, following two strokes that lesioned mainly his right occipitotemporal cortex and right hippocampus (Figure 2.7), he became severely prosopagnosic and lost the upper left visual field and a large part of his upper right visual field. A dynamic functional localizer (Pitcher et al., 2011) found atypical face activation bilaterally in all core face areas (except left OFA which could not be identified), with lower percent signal changes (PSCs) compared to controls. Herschel's activation to objects in the lateral occipital cortex bilaterally appeared to be normal (see Rezlescu, Pitcher, & Duchaine, 2012, for fMRI details and neuropsychological assessment).

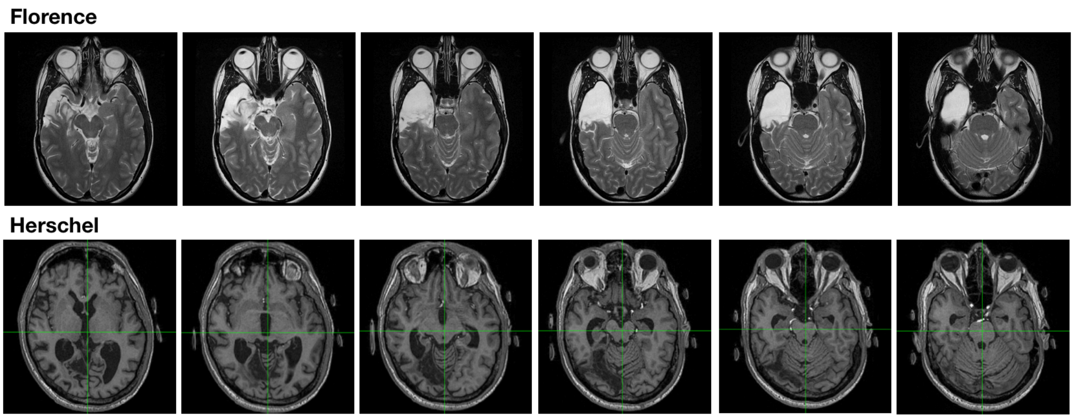


Figure 2.7. Brain lesions of Florence and Herschel. Structural scans are shown in radiological orientation (right hemisphere on the left).

The tests confirming Florence and Herschel's prosopagnosia are summarised in Table 2.2 (see Section 2.1 for test details). In addition to face perception and face memory deficits, Florence shows mixed performance with object recognition. Herschel's impairments however are largely restricted to faces.

Patient	Birth year	Famous faces	CFPT	CFMT	CHMT	CCMT	Sequential matching: Faces	Bodies	Objects
Florence	1982	<u>5%</u>	<u>39%</u>	<u>29%</u>	74%	69%	<u>64%</u>	79%	93%
Herschel	1956	<u>5%</u>	<u>33%</u>	<u>43%</u>	61%	75%	<u>54%</u>	<u>74%</u>	81%
	Control M	86.7%	73.9%	80.4%	74.6%	79.8%	78.1%	84.9%	84.9%
	Control SD	8.6%	8.2%	11.0%	8.1%	8.3%	5.9%	5.0%	7.3%

Patient	Birth year	Old/New recognition (A'): Faces	Cars	Horses	Houses	Tools	Scenes	Glasses	Guns
Florence	1982	<u>0.68</u>	<u>0.83</u>	<u>0.87</u>	0.90	<u>0.83</u>	<u>0.88</u>	<u>0.81</u>	<u>0.74</u>
Herschel	1956	<u>0.68</u>	0.96	<u>0.86</u>	0.97	0.97	0.98	0.88	0.89
	Control M	0.96	0.94	0.94	0.96	0.95	0.97	0.91	0.91
	Control SD	0.02	0.04	0.03	0.03	0.03	0.03	0.04	0.04

Table 2.2. Face and object perception for Florence and Herschel. Impaired scores are underlined. Herschel's problems are largely face-specific; note that when compared with age-matched participants (see section 2.1), his scores at horse old/new recognition and body sequential matching are not impaired. CFPT=Cambridge Face Perception Test; CFMT=Cambridge Face Memory Test; CHMT=Cambridge Hair Memory Test; CCMT=Cambridge Car Memory Test. CCMT norms are for male participants; female norms are lower (Dennett et al., 2012). Scores for Old/New recognition are expressed as A' which is unbiased nonparametric measure of sensitivity.

Control participants

Control data were provided by two age-matched groups of six participants each. Florence's control participants were between 27 and 32 years old (M = 29.3; five female). Herschel's control participants were between 47 and 56 years old (M = 52.3; four female).

Stimuli

The greeble stimuli were selected from those used by Gauthier and Tarr (1997). Greebles share a common first-order relational configuration between parts (Figure 2.8). They can be identified at the family level based on their overall shape or at the individual level based on the sizes and shapes of their parts.

The face stimuli (Figure 2.8) were selected based on a pixel-based similarity matrix (i.e. highly similar faces were preferred) from a large set of computer-generated male faces produced by FaceGen Modeller software (Singular Inversions, 2008).

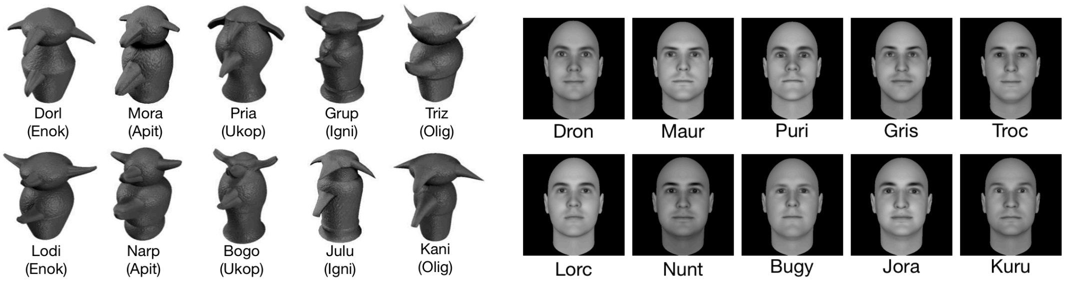


Figure 2.8. Ten of the 20 greebles and faces learned during the training experiment. Individual greebles belong to one of five families (in parentheses) and could be identified at the individual level and at the family level. Faces were identified only at the individual level. Faces and greebles had abstract four-letter names starting with a consonant (family names started with a vowel).

Procedure

The greeble training procedure was modelled after that used by Gauthier and Tarr (2002) and involved eight sessions, one per day. In the first four sessions, 20 individual greebles and the five families to which they belonged (four greebles per family) were gradually learned. Families had four-letter names each starting with a different vowel, and greebles had four-letter names each starting with a different consonant. Successful learning was tested with *naming* and *verification* trials. In naming trials, participants were asked to identify individual greebles by pressing the key corresponding to the first letter of their names. In verification trials, participants were presented with a name, which could be an individual or a family name, followed by a greeble, and asked to indicate if the name and the greeble matched. Participants received auditory feedback for incorrect answers. The final four sessions were testing sessions and included only naming and verification trials. Each learning session lasted about 60 minutes whereas each testing session lasted about 20 minutes. Sessions were completed in consecutive days (one control participant had a one-day break between two sessions).

A similar training procedure was created with computer-generated faces instead of greebles, with the difference that faces were not grouped into families (thus there were no family learning or testing trials). In pilot testing we adjusted the difficulty of the face training procedure to match that of the greeble training. Sessions one to four lasted approximately 45 minutes, while sessions five to eight lasted approximately 15 minutes.

Herschel, Florence and her control group completed first the greeble training and then the face training. Herschel's control group completed first the face training and then the greebles. All participants completed the two training sessions with a break of at least four weeks in between.

Results

To compare performance within and between participants, we computed an average score for the naming trials and an average score for the verification trials for the last four sessions. We selected the last four sessions because these were the testing-only sessions designed to measure recognition of all previously learned stimuli. Because participants are expected to become experts during the later sessions, this is also where we should observe a separation between controls and prosopagnosics in terms of performance according to the expertise hypothesis. Differences between prosopagnosics and controls were evaluated for statistical significance using Crawford's modified t-test for single case studies (Crawford & Garthwaite, 2002).

Performance in testing sessions

As expected, Florence and Herschel had severe difficulties learning faces (Figure 2.9). Florence responded correctly on only 41% of the naming trials and 58% of the verification trials (50% was chance level for verification trials). Herschel scored 30% for naming and 68% for verification. All scores were substantially and significantly lower than controls' average scores (significance levels are shown in Table 2.3).

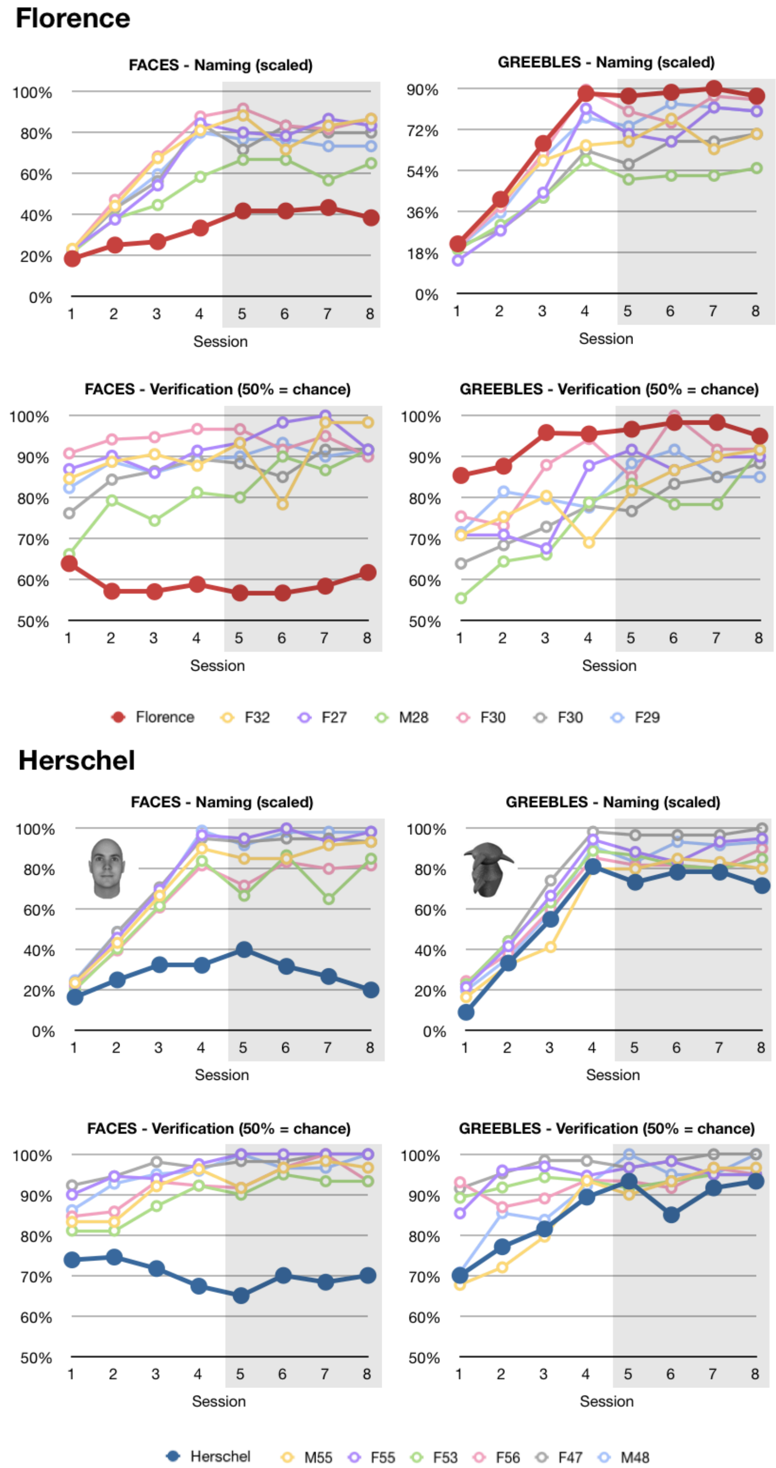


Figure 2.9. Individual scores during the greeble and face training procedures (testing sessions are shaded). Herschel (in blue) and Florence (in red) were severely impaired at learning faces but showed normal learning of greebles. The naming scores were scaled to reflect the varying difficulty of each session, corresponding to the total number of individuals learned up to that point (participants were tested for five individuals in session one, ten in session two, 15 in session three and 20 thereafter). The naming convention for controls reflect gender (M or F) and age. Each prosopagnosic was compared to his/her age-matched control group.

		Young controls		Florence		Old controls		Herschel	
		M	SD	Score	*p*	M	SD	Score	*p*
Faces	**Naming**	78.0%	7.8%	**<u>41.3%</u>**	*.007*	88.5%	9.1%	**<u>29.6%</u>**	*.002*
	Verification	91.5%	3.1%	**<u>58.3%</u>**	< *.001*	96.9%	2.7%	**<u>68.3%</u>**	< *.001*
Greebles	**Naming**	70.3%	10.9%	87.9%	*.195*	87.8%	6.0%	75.4%	*.114*
	Verification	87.2%	3.6%	97.1%	*.052*	95.8%	2.1%	90.8%	*.079*
	Verification family	90.7%	7.3%	96.7%	*.481*	95.3%	7.9%	96.7%	*.876*

Table 2.3. Average scores for the last four session. These are the testing sessions, after all greebles/faces were presented. P-values were calculated using Crawford's t-test for single case studies (Crawford & Garthwaite, 2002). Impaired scores are in bold and underlined.

In contrast to their performance with faces, both Florence and Herschel showed greeble learning curves comparable to controls' curves (Figure 2.9). Florence's average performance for the last four sessions, after all 20 greebles and their families were learned, was 88% for naming, 97% for verification individual, and 97% for verification family. Herschel's average scores were 75%, 91% and 97%, respectively. These scores were not significantly different from controls' (Table 2.3).

Florence and Herschel's better performance with greebles than faces cannot be attributed to task difficulty. If anything, controls' data suggest faces were easier to learn than greebles. Paired samples t-tests showed no significant differences between Herschel's controls' average scores at naming (88.5% and 87.8%, respectively; p = .784) or verification (96.9% and 95.8%, respectively, p = .115) trials with faces and individual greebles. Florence's group fared better with faces than greebles at both naming (78.0% versus 70.3%, p = .045) and verification (91.% versus 87.2%, p = .002).

The dissociation between face learning (impaired) and greeble learning (normal) for Florence and Herschel was confirmed by the Bayesian Standardized Difference Test (Crawford & Garthwaite, 2007), which estimates the percentage of control population exhibiting a more extreme difference between two tasks than a patient. With naming and with verification, for both Florence and Herschel these estimates were below 0.01%.

Response times

To check whether their normal accuracy could be explained by speed-accuracy trade-offs, we compared Florence and Herschel's response times (RTs) to controls' RTs. For each participant, session, and trial type (individual naming, individual verification and family verification), we computed a mean RT for correct responses for 'hit' trials (Gauthier & Tarr, 2002). Means were computed after excluding trials with RTs more than two standard deviations away from the mean to remove outliers and prevent statistical equivalence due to high variability (Figure 2.10). For the testing sessions (5-8), Herschel's average RT for individual verification trials was 1882 ms (controls: M = 1480 ms, SD = 527 ms; p = .512) and for family verification trials was 1690 ms (controls: M = 1288 ms, SD = 801 ms; p = .661). Florence's average RTs were 1241 ms for individual verification (controls: M = 1017 ms, SD = 194 ms; p = .334) for and 890 ms for family verification (controls: M = 814 ms; SD = 153 ms; p = .663). For naming trials, Herschel averaged 3428 ms (controls: M = 2074 ms, SD = 607 ms; p = .094), while Florence averaged 2128 ms (controls: M = 1619 ms, SD = 158ms; p = .030). Note that naming RTs are slower and have limited informational value because participants need to find the right key (therefore they reflect more than perceptual differences between participants). In the face learning task, Florence and Herschel were substantially slower for both verification (2534 ms and 2978 ms, respectively) and naming (4441 ms and 6559 ms, respectively), while controls' RTs remained comparable to those recorded with greebles (e.g. Florence's controls: 938 ms and 1747 ms, respectively).

Expertise criterion

According to Gauthier and Tarr (1997), participants undergoing greeble training are considered greeble experts when average RTs for correct responses at individual-level recognition, initially slower, become comparable (i.e. not statistically different) to average RTs for correct responses at family-level recognition. Herschel reached this criterion in sessions 6, 7 and 8 (p = .748, p = . 946, and p = . 469), while Florence reached the criterion in the final session (p = .

113) (Figure 2.10). As a more stringent criterion for expertise, Bukach et al. (2012) recommended adding an extra constraint: the absolute difference between family and individual level response times should be lower than 95 ms (corresponding to a 95% interval of response times differences from a previous greeble study, Gauthier & Tarr, 2002). Herschel's RTs satisfied this criterion in sessions six and seven (54 ms and -16ms, respectively). Florence did not meet this criterion.

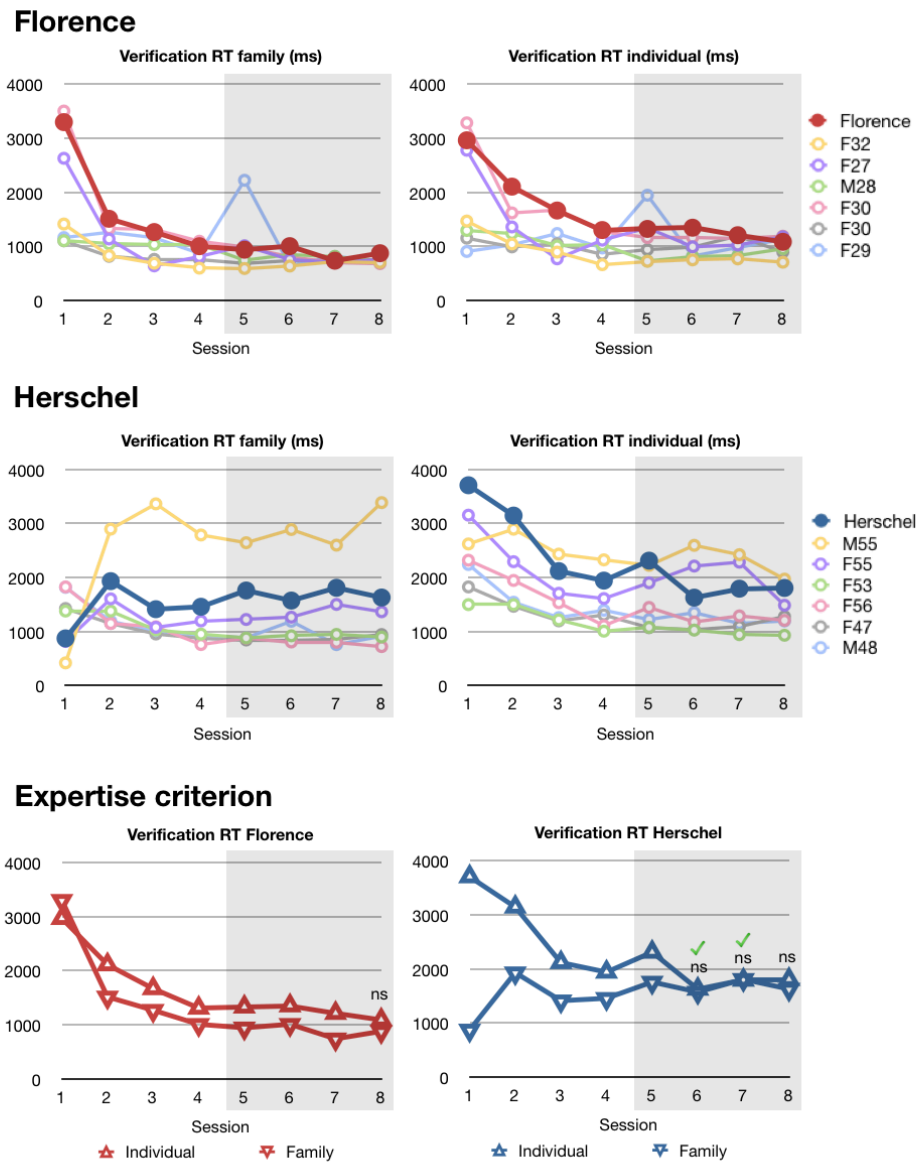


Figure 2.10. Response times during the greeble training procedure. Testing sessions are shaded. Herschel and Florence's displayed comparable RTs to their age-matched control groups, indicating normal greeble learning cannot be explained by speed-accuracy trade-offs. Most importantly, in three (Herschel) and one (Florence) of the final four sessions, there were no significant differences between individual and family verification RTs. According to Gauthier and Tarr (1997), the absence of a difference indicates that Herschel and Florence were greeble experts. ns = difference not significant. The green ticks mark sessions where the absolute difference between individual and family level hit trials was below 95ms, the most stringent expertise criterion (Bukach et al., 2012).

Discussion

Our results demonstrate a clear dissociation between greeble learning and face learning. Greeble recognition after training has been claimed to engage the same mechanisms used for face processing (Gauthier & Tarr, 1997; Gauthier et al., 1999; Rossion et al., 2002), and this predicts lesions impairing face recognition will also disrupt greeble recognition after training. Here we show this is not the case: two acquired prosopagnosics performed normally in a standard greeble training procedure while exhibiting severe impairments with a comparable face task. Additionally, both Florence and Herschel showed reaction time effects which are putative markers of greeble expertise.

Prosopagnosic individuals undergoing the greeble training procedure have been previously reported, but these studies had limitations that complicated their interpretation. Two acquired cases were unable to develop greeble expertise (Behrmann, Marotta et al., 2005; Bukach et al., 2012), but they also suffered from object recognition deficits (Bukach et al., 2012; Gauthier et al., 1999), which makes their results unclear. They may have failed with greeble training because they have object recognition problems rather than a disruption of expert mechanisms. Consistent with this interpretation, the reported cases were impaired with greeble recognition from the beginning of training, whereas deficits limited to mechanisms necessary for expert processing would be expected to become apparent only in later sessions after control participants have acquired expertise. Edward, a developmental prosopagnosic with intact object recognition, performed normally in greeble training (Duchaine et al., 2004), but his performance with greebles was not contrasted with performance on a parallel face training paradigm so it is not clear that the procedure would elicit deficits in Edward if faces were used. We based our study on individuals whose severe agnosias were mostly face-specific, both showing normal processing of several other object categories (especially Herschel; see 1B), including sequential matching of novel objects (not greebles). This made them excellent candidates to reveal a dissociation between face expertise and greeble expertise. Furthermore, the direct comparison between performance with greeble learning (normal) and face learning (impaired) should offset concerns recently raised by expertise researchers who noted: “In particular, evidence for separate mechanisms [for faces and greebles] would require a demonstration that [prosopagnosics] could learn to individuate 20 greebles, but be unable to learn to individuate 20 faces using the same task” (Bukach et al., 2012, p. 290).

Our findings have implications for the debate concerning the specificity of the mechanisms that carry out face processing. The face-specific hypothesis posits the existence of brain mechanisms specialised for face perception, and supports this claim with behavioural (Tanaka & Farah, 1993; Yin, 1969; Young et al., 1987), neurophysiological (Kanwisher et al., 1997; Tsao et al., 2006) and patient (see review in Busigny, Joubert et al., 2010) studies. In contrast, the expertise hypothesis argues that, rather than being face-specific, mechanisms underlying face processing support within-category recognition of objects for which individuals have acquired substantial expertise (Diamond & Carey, 1986; Gauthier & Tarr, 1997). The current study provides evidence that undermines a set of related claims based on greeble research. If we assume, like proponents of the greeble framework do, that i) the greeble training procedure can develop expertise with a novel class of objects and that this expertise is comparable to real-life expertise with certain objects such as faces; and ii) the ‘expertise’ criterion (i.e. similar response times for identification at the family and individual level) is a valid indicator of acquired expertise, then the dissociation between face expertise and greeble expertise is strong evidence against the expertise hypothesis.

Herschel’s normal performance with greebles is especially interesting, given claims about the neural basis of greeble expertise. It has been suggested that the fusiform face area (FFA), a functionally-defined region that responds more strongly to faces than to other stimuli (Kanwisher et al., 1997; Pitcher et al., 2011), is the locus of mechanisms mediating expertise with all objects, including faces, greebles and cars (Gauthier et al., 1999; Gauthier et al., 2000; McGugin et al., 2012). Atypical functioning of the FFA, as seen in Herschel, would thus be expected to interfere with acquisition of expertise. Contrary to this prediction, Herschel's normal greeble learning suggests greeble expertise is not dependent on the FFA and the neural mechanisms supporting face expertise and greeble expertise differ. This is consistent with previous research that failed to find a correlation between activation in the FFA and the expertise level with novel objects (Kung, Peissig, & Tarr, 2007; Moore, Cohen, & Ranganath, 2006; Op de Beeck, Baker, DiCarlo, & Kanwisher, 2006; Yue, Bosco, & Biederman, 2006) or real objects (Grill-Spector et al., 2004). In two of these studies (Op de Beeck et al., 2006; Yue et al., 2006), changes associated with the acquisition of expertise seemed to occur in the lateral occipital complex, a region that showed normal activation in Herschel. A small but significant effect of car and bird expertise on FFA activation was reported by Gauthier et al. (2000), however the effect extended beyond FFA. More recently, behavioural expertise was shown to correlate with FFA activation to cars (McGugin et al., 2012), but the effect was not specific to face-selective voxels (at least not in the right FFA). Other imaging findings apparently supporting the expertise hypothesis may have been confounded by an imprecise localization of the FFA (Harley et al., 2009; Gauthier et al., 1999), particular measures of activation and statistical interpretation of results (Gauthier et al., 1999) or attentional bias (Xu et al., 2005).

If one believes that the greeble framework can produce rapid expertise, then our results conclusively reject the rapid expertise hypothesis. However, one needs to be cautious in extending these findings to expertise with real objects, which develops over a much longer period. Even though some of the initial claims made by proponents of the long-term expertise were disproved (face inversion effects that take ten years to develop; Carey, 1992) or failed to replicate (the face-like inversion effects for dogs in dog show judges; Diamond & Carey, 1986), it could be that the rapid expertise purportedly acquired in greeble training and the extended expertise that might be acquired for everyday objects rely on distinct mechanisms. Neuropsychological evidence though suggests face processing dissociates from long-term object expertise as well. After brain lesions causing prosopagnosia, WJ learned to distinguish tens of sheep (McNeil & Warrington, 1993), while RM retained his superior ability to recognise car makes and models

(Sergent & Signoret, 1992). Conversely, several brain-damaged individuals lost their expertise with non-face objects while their face recognition remained normal (CK: Moscovitch et al., 1997; MX: Assal, Favre, & Anderes, 1984).

To conclude, our current findings add to the evidence supporting the face specificity hypothesis. Acquired prosopagnosics who are able to learn greebles provide straightforward evidence that face and greeble recognition rely on different mechanisms. Claims from the greeble literature about common functional and neural substrates underlying faces and expert objects recognition should therefore be reconsidered.

Chapter 3

FACIAL TRAIT PERCEPTION IN ACQUIRED PROSOPAGNOSIA

In Chapter 1.2 I presented three models of face perception and their take, explicit or implied, on face impressions. The Bruce and Young (1986) model proposed a functional division between mechanisms implicated in identity recognition and mechanisms involved in all other aspects of face perception. According to this model, facial trait perception dissociates from face recognition, but not from perception of emotional expressions or sex (although the authors allowed for further divisions within the functional components responsible for non identity-specific aspects). The Haxby model (Haxby et al., 2000) included a core system and an extended system for face perception. In the core system, there is a functional and neural division between mechanisms involved in processing invariant facial aspects (for example, those leading to identity recognition) and mechanisms for processing changeable facial aspects (e.g. perception of eye gaze and facial expressions). From the core system, some information about changeable aspects is passed on to the part of the extended system for face perception that mediates emotion recognition (amygdala/insula/limbic lobe). In the Haxby model there is no explicit reference to person impressions, but one may infer these are carried out along the same route leading to emotion recognition. In this case, trait perception will dissociate from identity recognition but not from expression recognition. The third model presented (Oosterhof & Todorov, 2008) is an explicit model of trait perception. According to this model, face evaluations can be reduced to two main dimensions best approximated by trustworthiness and dominance judgments. More importantly for our purposes, Oosterhof and Todorov (2008) suggested there are no dedicated mechanisms for perception of facial trustworthiness or dominance. Their overgeneralisation hypotheses stated that trustworthiness judgments are by-products of mechanisms involved in perception of happy and angry facial expressions, and that dominance judgments are by-products of mechanisms for face-based sex recognition.

Overall, these three models imply that i) face evaluations and facial identity recognition rely on distinct functional and neural routes, and ii) face evaluations are performed by the mechanisms responsible for happy/angry expression recognition (the trustworthiness dimension) and sex recognition (the dominance dimension). Previous behavioural, computational, imaging and patient studies summarised in Section 1.2 revealed a mixed picture with regards to these two hypotheses. The goal of the studies presented here was to test these hypotheses in four individuals with acquired prosopagnosia. A necessary interim step was to create a new battery of tests of face perception.

In **Section 3.1** I present the development of ten new tests for perception of facial trustworthiness, attractiveness, aggressiveness, happiness and anger (the reasons for measuring aggressiveness rather than dominance are presented at the beginning of section 3.3). These tests, together with other tests already available in the literature, ensured the results are robust and that performance across different aspects is measured homogeneously.

Section 3.2 tests whether face evaluations and face recognition are dissociable. Four individuals with severe face recognition deficits were assessed with multiple tests for perception of face trustworthiness, attractiveness and aggressiveness. The prediction was that normal trait perception can occur in the absence of identity recognition.

In **Section 3.3** I extended the investigation of one acquired prosopagnosic to include multiple tests for face-based perception of happy and angry expressions, and of sex. The aim was to test the overgeneralisation hypotheses in facial trait perception. The prediction was that normal trustworthiness and aggressiveness judgments do not depend on intact mechanisms for facial expression and sex recognition. Evidence supporting this hypothesis would be inconsistent with the Todorov model.

3.1. TEN NEW TESTS OF FACE PERCEPTION

Rationale

Previous studies examining facial trait perception used ratings (e.g. Oosterhof & Todorov, 2008; Quadflieg et al., 2012; Todorov & Duchaine, 2008). Participants were shown one face at a time and asked to rate it on a Likert scale for a particular trait. To increase the robustness of our findings and facilitate comparison with other aspects of face perception, we wanted to assess trait perception with multiple tests that would have similar format across all dimensions of interest. For example, testing the emotion overgeneralisation hypothesis in trait perception required a comparison between perceptions of facial trustworthiness and of happy and angry expressions. However, no suitable tests were available for individual expression perception; some of the existing perceptual tests of expression may be used to derive a score for each expression, but the interpretation of these scores is limited by the reduced number of trials and interaction with other expressions (Pitcher, Garrido, Walsh, & Duchaine, 2008). Furthermore, these tests were not suitable to adapt for examining trait perception.

Therefore we developed a new battery of tests that were specifically designed to overcome shortcomings in previous comparisons of different facial deficits in prosopagnosia (Calder & Young, 2005). First, the level of difficulty of each test was adjusted to avoid floor and ceiling effects that can mask subtle impairments. Second, tests of different facial aspects had identical format and task demands. Third, within each format, tests of different facial aspects were comparable in difficulty. Fourth, all tests generated substantial inversion effects (i.e., a larger drop in discrimination performance for upside-down faces than for upside-down objects that is characteristic of face processing, Yin, 1969).

Our new battery of tests consists of two formats: a sorting task modelled after the Cambridge Face Perception Test (CFPT - Duchaine et al., 2007) and a categorical Odd-One-Out task. Both formats were used to assess perception of trustworthiness, attractiveness, aggressiveness, happy expression, and angry expression. Because facial trait judgments are not intrinsically objective, the judgments of control and prosopagnosic participants were compared to the consensus judgments of a large group of people with normal face perception (Adolphs et al., 1998; Todorov & Duchaine, 2008).

Face stimuli

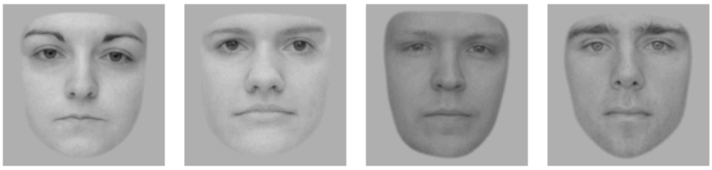
Face stimuli were selected from three databases. For the trait tests, we used the Karolinska Directed Emotional Faces (KDEF - Lundqvist, Flykt, & Ohman, 1998) and a subset from Glasgow Unfamiliar Face Database (GUFD - Burton, White, & McNeill, 2010). For the expression tests, we used a subset from the Radboud Face Database (RFD - Langner et al., 2010).

The KDEF set contains 70 individuals, photographed from five angles while displaying a neutral expression and six other emotional expressions. We selected the neutral frontal head-shots of 66 individuals (33 female) that were used by Oosterhof and Todorov (2008). To remove the potential confounding effects of hair and skin colour on trait judgments (Stephen et al., 2011; Swami et al., 2008), face images were converted to grayscale and cropped to show only the internal features (eyes, nose, mouth) (see Figure 3.1). These faces were rated for trustworthiness by 140 online participants, and for aggressiveness by 48 online participants, recruited through Amazon's online labour market called Mechanical Turk (www.mturk.com). Results from web-based samples of participants have been shown to have comparable means, standard deviations and internal reliability as those from lab-based samples (Germine et al., 2012). The ratings were on a scale from 1 to 9. Mean trait ratings for female and male faces were comparable (all *p*s > .19). Trustworthiness ratings were negatively correlated with aggressiveness ratings ($r = -.64$, $p < .001$). For each trait, we normalised the scores per participant and computed the average z-score for each face. These average z-scores were then used as an indicator of the perceived trustworthiness and aggressiveness of each face. Because attractiveness ratings were not collected, we used the average ratings published by Oosterhof and Todorov (2008). Note that, for trustworthiness, the current ratings of grayscale cropped faces correlated highly with the ratings of un-cropped colour faces obtained by Oosterhof & Todorov (2008). For female faces, the correlation was .80 ($p < .001$), while for male faces it was .62 ($p < .001$). Therefore, we considered the attractiveness ratings of colour un-cropped faces to be a good proxy for the perceived attractiveness of grayscale cropped faces.

Trait	Target gender	Cronbach	M	SD	t	p	
Trustworthiness	F	0.86	3.98	0.41	0.96	0.341	-.64 (<.001)
	M	0..85	4.07	0.40			
Aggressiveness	F	0.86	4.20	0.76	1.31	0.196	
	M	0.90	4.47	0.88			

Figure 3.1. Face stimuli from the KDEF database. Faces were converted to grayscale and cropped to show only the internal features. Evaluations of trustworthiness and aggressiveness were highly reliable, with Cronbach's α between .85 and .90. Ratings for male faces did not differ significantly from ratings for female faces. Trustworthiness ratings negatively correlated with aggressiveness ratings ($r = -.64$).

The GUFD set contains 303 individuals, each with three frontal head-shots taken with different cameras. We selected 200 images (half female) of young to middle aged Caucasians that had closed mouth, full face visible (i.e., no hair covering the eyes), and no facial hair (for male). Images were converted to grayscale and faces were cropped to show only the internal features (eyes, nose, mouth) (see Figure 3.2 for examples). We then asked 150 Mechanical Turk participants (MTurks) to rate these faces on a scale from 1 to 9 for trustworthiness, attractiveness or aggressiveness (50 MTurks per trait). Inter-rater agreement (as measured by Cronbach's α) was high for all traits, ranging between .92 and .95 (Figure 3.2). Female faces were on average perceived more positively than male faces, with higher mean ratings for trustworthiness and attractiveness, and lower mean ratings for aggressiveness. We normalised the scores per participant and stimulus gender (i.e. scores for female faces were normalised based on the mean and standard deviation of ratings for female faces), and then computed the average z-score per face per trait. We used the average z-scores as an indicator of the perceived trustworthiness, attractiveness and aggressiveness of each face. Trustworthiness ratings were positively correlated with attractiveness ratings (r = .55, p < .001) and negatively correlated with aggressiveness ratings (r = -.84, p < .001). Attractiveness and aggressiveness ratings were negatively correlated (r = -.29, p < .001).



Trait	Target gender	Cronbach	M	SD	t	p
Trustworthiness	F	0.92	4.55	0.86	4.66	<.001
	M	0.92	3.98	0.86		
Attractiveness	F	0.95	4.24	0.97	2.14	0.033
	M	0.93	3.97	0.77		
Aggressiveness	F	0.93	4.24	0.97	2.73	0.007
	M	0.94	4.63	1.08		

Correlations:
- Trustworthiness – Attractiveness: .60 (<.001)
- Attractiveness – Aggressiveness: -.32 (<.001)
- Trustworthiness – Aggressiveness: -.82 (<.001)

Figure 3.2. Face stimuli from the GUFD. One hundred female (F) and 100 male (M) faces were rated for trustworthiness, attractiveness and aggressiveness. Ratings were highly reliable, as evidenced by Cronbach's α. Female faces received on average more positive evaluations; t and p values are for inter-gender comparisons. Trait evaluations were significantly correlated with each other (note that the correlations reported here are computed between actual ratings, and are slightly different from the correlations between z-scores reported in the text).

The RFD includes colour pictures of males and females with eight expressions (including neutral) at different viewpoints. We selected the frontal views of 18 males and 18 females in neutral, happy and angry expressions, converted them to grayscale, and created morphed images using FantaMorph (Abrosoft, 2002). For each individual we created one morph continuum from the neutral to the happy image, and another from the neutral to the angry image (see Figure 3.3). Based on pilot testing (see next section), we selected the morphed images best fitted to the difficulty of each expression test.

Figure 3.3. Examples of morphed images created from the Radboud Face Database. The faces look very similar, but in fact they vary slightly on how happy they look (if you cover the faces in the middle and look only at the faces at the two ends, the difference will be more noticeable).

Sorting tests

The sorting tests were modelled after the Cambridge Face Perception Test (CFPT) for identity perception (Duchaine, Germine, & Nakayama, 2007; Duchaine, Yovel, & Nakayama, 2007). Each test had ten trials, presented in a pre-randomised fixed order, and in each trial, participants had one minute to sort six faces on a given dimension (e.g., from the least trustworthy to the most trustworthy). Traditionally, the CFPT score represents a sum of deviations from the correct ordering for each trial, with higher scores denoting poorer performance. In the current study, we expressed performance as percentage of correct sorts using the formula: [% correct = 1 - sum of deviations/maximum number of deviations possible]. With six faces to sort, the maximum number of deviations possible is 18, so the formula becomes: [% correct = 1 - sum of deviations/18]. With this formula, higher scores denote better performance. Using simulation data, chance performance was calculated to be 11.6 deviations per trials, or 35.6% correct sorts.

The CFPT Trustworthiness, CFPT Attractiveness and CFPT Aggressiveness (Figure 3.4) featured GUFD faces. Each test had ten trials, and one trial featured six faces of the same sex with roughly equal spread in terms of z-scores differences (0.3 to 0.5) on that particular trait. Half of the trials presented female faces. Each face was used only once in a test, for a total of 60 faces (30 female) per test. Faces used by CFPT Trustworthiness were not repeated in the other tests. CFPT Attractiveness and CFPT Aggressiveness shared some of their stimuli, but their allocation within each trial was different. Note that the correct ordering of faces on all tests was determined by their consensus ratings (e.g. what most control participants regarded as trustworthy or untrustworthy faces). The tests were validated by three control groups (n_{tw} = 32, n_{att} = 31, n_{agg} = 30) of Dartmouth College undergraduate students (Table 3.1).

Figure 3.4. Examples of trials from CFPT Trustworthiness CFPT Aggressiveness. Participants had one minute to order six identities according to perceived trustworthiness or aggressiveness. CFPT Attractiveness used the same format and stimuli from the same database.

In the CFPT Happy and CFPT Angry (Figure 3.5), each trial presented six morphed images of one Radboud individual on a continuum between the neutral and happy (angry) version of him/herself. We conducted extensive pilot studies to find the optimum distances between the morphs to reach the target performance score of 75% (similar to average performance level for CFPT Identity, Duchaine, Germine, & Nakayama, 2007). These distances were 3% for the CFPT Happy and 8% for CFPT Angry. Each 'happy' trial thus presented the 0%, 3%, 6%, 9%, 12% and 15% morphs of the neutral and happy versions of an individual, while each 'angry' trial presented the 0%, 8%, 16%, 24%, 32% and 40% morphs of the neutral and angry versions of an individual. Morphs were shown from left to right in a predetermined scrambled order, different from trial to trial but comparable in terms of distance from the correct order. Control participants were 30 undergraduate students from Dartmouth College (Table 3.1).

Figure 3.5. Examples of trials from CFPT Happy and CFPT Angry. Participants had one minute to order six faces (same identity) according to how happy/angry they looked.

Categorisation tests

For all measured abilities but sex perception, the categorisation was an odd-one-out (OOO) task. Participants were presented with a line-up of three faces and were asked to select the face that is different from the other two on the dimension of interest. As an example, for the OOO Happy, the correct response would be to select the neutral face if the other two faces are happy, or the happy face if the other two faces are neutral. To improve scores for the OOO Trustworthiness/ Attractiveness/Aggressiveness (which suffer from the lack of an objective decision criteria), we blocked the trials in two halves: one that required participants to identify the trustworthy/attractive/aggressive face from the three, the other that required identification of the untrustworthy/unattractive/unaggressive face from the three. The OOO Happy and Angry trials were unblocked, such that participants had to flip between selecting the happy/angry faces and the neutral faces. Presentation time for each line-up was 3 seconds, but participants had unlimited time to answer. Chance performance was 33.3%.

The OOO tests for trait perception used Glasgow and Karolinska faces. For OOO Trustworthiness and OOO Aggressiveness, we ordered the selected 200 Glasgow faces (half of them female) on attractiveness according to the ratings collected from the MTurk participants. We then created, per sex, 10 equal bins starting from the least attractive to the most attractive face, such that each bin included 10 faces of comparable attractiveness. From each bin we selected three faces such that one face was markedly different from the other two on trustworthiness/aggressiveness. For OOO Trustworthiness (Figure 3.6), we created 16 trials with female faces and 16 trials with male faces, half of them with a trustworthy target (the other two faces are untrustworthy) and half with an untrustworthy target (the other two faces are trustworthy). The same procedure was used to create 32 trials (half with female faces) for OOO Aggressiveness. Faces were not repeated across tests. In addition to the 32 Glasgow trials, we created 24 Karolinska trials. From the Karolinska faces, we selected, per gender, the nine least trustworthy/aggressive and the nine most trustworthy/aggressive faces that were used to create 12 female and 12 male trials for the OOO Trustworthiness/ Aggressiveness. Each face was repeated once. In contrast to the Glasgow trials, the Karolinska trials did not have the faces equated for attractiveness. OOO Attractiveness was created in a similar fashion, with the Glasgow faces (but not the Karolinska faces) within one trial equated for trustworthiness. Final control data came from 30 online participants (M = 30.1 years old, SD = 16.4; 17 female).

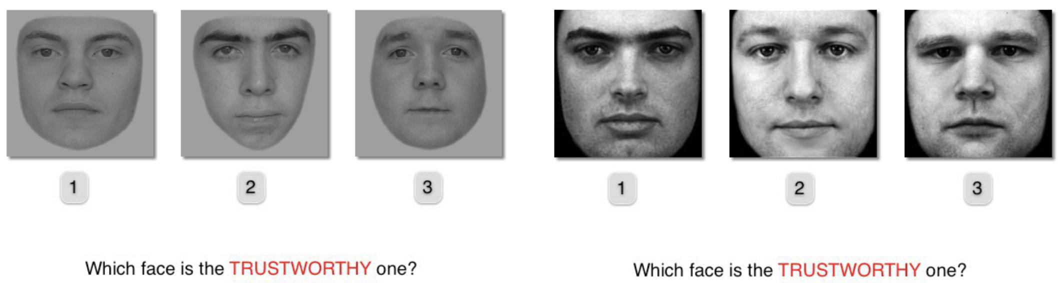


Figure 3.6. Examples of trials from OOO Trustworthiness. Participants had three seconds to select the face that looked trustworthy from the three faces presented. The first trial present GUFD faces, while the second presents KDEF faces. OOO Attractiveness and OOO Aggressiveness used a similar format.

For OOO Happy and OOO Angry we used 36 Radboud individuals (18 female) to generate 36 trials. Half of the trials presented one neutral face and two expressive faces, and the other half one expressive face and two neutral faces. Faces within one trial belonged to different individuals. We used three faces per individual (one neutral and two expressive, or one expressive and two neutral). Each individual's expressive face was a morph between the neutral and the happy/angry picture of that individual. The emotional intensity of each face (i.e., the point on the morph continuum from which the image was taken) was selected based on pilot testing and varied from one individual to another. To avoid the feature-based strategy noted in pilot studies (i.e. for OOO Happy, the choice between happy/neutral was based on whether teeth were visible or not; for OOO Angry, the choice between angry/neutral was based on the creases between the eyebrows), we limited presentation time to three seconds per trial. Moreover, for OOO Happy we masked the mouth area with a grey rectangle, and for OOO Angry we blurred the area between the eyebrows to make creases less conspicuous (Figure 3.7). These adjustments necessarily made the tests more difficult, which could potentially reveal finer impairments. They are also consistent with the aim of the present study, because the emotion overgeneralisation hypothesis suggests that subtle (rather than obvious) resemblance to emotional expressions is what drives trustworthiness judgments. Control data were provided by 30 online participants (M = 35.1 years old, SD = 11.7; 15 female).

Figure 3.7. Examples of trials from OOO Happy and OOO Angry. Participants had three seconds to decide which face was different from the other two in terms of how happy/angry they looked.

Test format	Facial aspect	Controls n	Controls M	Controls SD
CFPT chance: 36% max: 100%	Happy expression	30	75%	8%
	Angry expression	30	76%	9%
	Trustworthiness	32	71%	8%
	Attractiveness	31	67%	7%
	Aggressiveness	30	63%	6%
OOO chance: 33% max: 100%	Happy expression	30	75%	10%
	Angry expression	30	77%	10%
	Trustworthiness	30	72%	10%
	Attractiveness	30	60%	10%
	Aggressiveness	30	68%	9%

Table 3.1. Ten new tests of face perception. Two test formats, sorting (CFPT) and categorisation (OOO) were used for each aspect measured. These aspects were two expressions (happy and angry) and three traits (trustworthiness, attractiveness and aggressiveness). We aimed to equate difficulty level across tests, with most controls' means around 75%.

3.2. FACIAL TRAIT PERCEPTION WITHOUT FACIAL IDENTITY RECOGNITION

Two recent neuropsychological studies suggested that facial trait judgments do not rely on the same perceptual mechanisms that carry out facial identity recognition. One study showed that two individuals with developmental prosopagnosia were able to judge trustworthiness normally (Todorov & Duchaine, 2008), while another study showed PS, a well-studied woman with acquired prosopagnosia, made normal judgments for several traits (Quadflieg et al., 2012).

These studies, however, might have suffered from two methodological issues. First, trait judgments were measured with one format only, namely rating, which is not the way identity perception was assessed. The different task demands prevent direct comparisons between identity and trait results and increase the chance of reporting spurious dissociations. Second, face stimuli in one study were in colour and included hair (Quadflieg et al., 2012), both of which influence trait judgements (Todorov & Duchaine, 2008; Stephen et al., 2011; Swami et al., 2008), and are absent in most tests of identity recognition. To overcome these limitations, here we systematically investigated perceptions of facial trustworthiness, attractiveness, and aggressiveness for greyscale faces without external facial cues. Each trait was assessed using three formats: rating, sorting, and categorisation.

Methods

Prosopagnosic participants

We tested four individuals with acquired prosopagnosia.

Florence (described previously in Section 2.2 and in Fox et al., 2011) is a 29-year old Canadian nurse. In 2006 she became prosopagnosic following a right amygdalohippocampectomy to treat epilepsy. In 2010 she underwent a second operation that removed most of her right anterior temporal lobe.

Grace is an American pharmacist born in 1968 (43 years old at the time of testing). At 15 she contracted herpes simplex viral encephalitis and she was treated with acyclovir after a right temporal lobe biopsy. She is severely prosopagnosic and does not reliably recognise even family members. Her husband told us that one time she picked up the wrong child from the kindergarten. She also reports having problems with animal identification ("I ask myself, is that really a cow?") and colour perception: she can reliably recognise only red and yellow, but no other colours (black and white are fine). She claims she can perceive colour differences (without naming the colours), even though she failed the online Munsell Hue Test. She reports good navigation skills. She was previously presented as B-OT/AT1 in Dalrymple et al. (2011), where she performed well on most neuropsychological tests. Structural MRI scans revealed a large lesion of the right anterior temporal lobe extending to the middle fusiform gyrus, and a small lesion of the mid-fusiform gyrus in the left hemisphere.

Kepler is an American man born in 1958 (53 years old at the time of testing). He became prosopagnosic following a stroke that lesioned his right inferior occipitotemporal lobe. He experiences visual deficits in the upper left visual field and the left half of the foveal representation.

Sandy is an American woman born in 1975 (36 years old at the time of testing). She became prosopagnosic following a resection for epilepsy in 1997. Her problems with faces are severe; for instance, she is unable to reliably recognise her husband or her children. To be able to pick up the children from school she buys them backpacks from other states to ensure they are distinct from other children's backpacks. Interestingly, she reports "losing" faces as they age, apparently not being able to update their representation. When she goes back to her hometown, she can recognise colleagues who have not aged a lot, but not the ones who have put on weight or changed significantly. Similarly, she cannot recognise her mother because she aged considerably during the last decade, yet she can recognise her father who aged mostly before Sandy's problems started. She can recognise herself in photos when she was young, but not now when she sees herself in the mirror. She says she never really "knew" her children because they change so quickly. Sandy was convinced she would do fine on a famous faces test with faces from before 1997 (e.g. Demi Moore in the movie "Ghost"), but this was not tested. She has a full left hemianopia. Structural MRI scans were not available.

All four individuals had severe deficits processing facial identities, as confirmed by their impaired performance on the Cambridge Face Memory Test, Face Old/New Test and Cambridge Face Perception Test (Table 3.2).

Patient	Birth year	Sex	Cause	Lesions	CFMT	Face Old-New	CFPT
Florence	1982	F	Resection for epilepsy	R Temporal lobe, including hippocampus and amygdala	26%	0.68	39%
Grace	1968	F	Herpes Encephalitis	Bilateral Occipito-Temporal and Anterior Temporal	42%	0.66	43%
Kepler	1958	M	Stroke	R Inferior Occipito-Temporal lobe	47%	0.71	42%
Sandy	1975	F	Resection for epilepsy	n/a	33%	0.85	42%
				Control M	80.4%	0.96	73.9%
				Control SD	11.0%	0.04	8.2%

Table 3.2. List of acquired prosopagnosics (APs). The four APs have been evaluated with the Cambridge Face Memory Test (CFMT), a Face Old-New Recognition test, and the Cambridge Face Perception Test (CFPT). CFMT introduces participants to six target faces, which are then presented in test items with distractor faces in different poses and lighting (Duchaine & Nakayama, 2006). CFMT scores represent percentages of correct identifications out of 72 trials. In the Face Old-New test (Duchaine & Nakayama, 2005), participants memorize 10 target faces and then discriminate between those faces and new faces. Scores are A' values (an unbiased measure of discrimination). CFPT is described in section 3.1. Note that CFPT scores have previously been expressed as a sum of errors, but here we present percentages of correct orderings. All CFMT, Old-New and CFPT scores reported above are impaired. L=left, R=right, n/a = not available.

Control participants

For ratings, control data were provided by 30 students from University College London and Dartmouth College (mean age 19.1 years, 14 female). For the other tests, details about each control group are provided in Section 3.1.

Results

We used Crawford's modified t-test (Crawford & Howell, 1998) to determine whether prosopagnosics' performance deviates from controls' in a statistically abnormal manner. Figure 3.8 shows the results of the nine tests of facial trait perception (3 traits x 3 formats) completed by each prosopagnosic participant.

Florence showed trait judgments in line with controls in all nine tests.

Grace was normal with trustworthiness judgments. For attractiveness, she was impaired with ratings ($t = 2.28$, $p = .030$), but normal with CFPT and OOO. For aggressiveness, she provided typical ratings, but was impaired with CFPT ($t = 5.74$, $p < .001$) and OOO ($t = 2.95$, $p = .006$).

Kepler was normal with trustworthiness and attractiveness, but impaired with aggressiveness on CFPT ($t = 2.88$, $p = .009$) and scored in the lower range for on OOO ($t = 1.41$, $p = .166$).

Sandy was impaired at CFPT Trustworthiness ($t = 3.81$, $p < .001$) and CFPT Attractiveness ($t = 3.51$, $p = .001$), but fine with rating and categorisation (i.e. OOO). With aggressiveness, she displayed normal performance in all three tests.

Discussion

Overall, our findings robustly demonstrate that facial trait judgments can be normal when facial identity recognition is impaired. Despite their severe face recognition deficit, all prosopagnosics made judgements about at least one trait normally in all three test formats. Florence in particular was completely normal with all measured traits. Grace was normal with trustworthiness judgments, Kepler with trustworthiness and attractiveness, and Sandy with aggressiveness. These results indicate that the perceptual mechanisms underlying trait judgments are dissociable from those used in recognising identities.

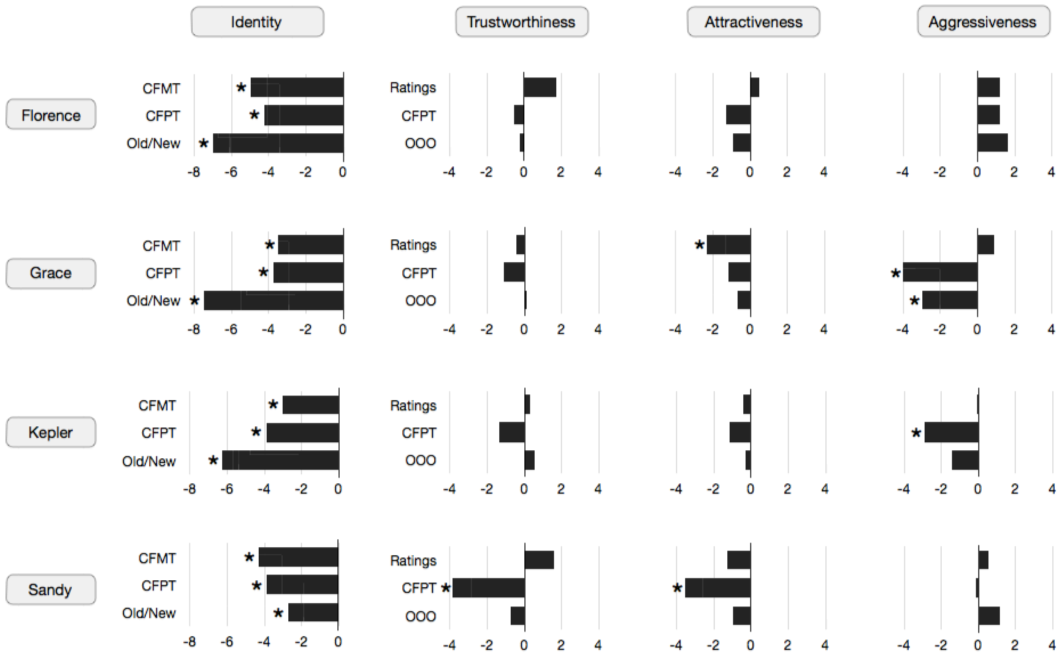


Figure 3.8. Facial identity and trait perception in four acquired prosopagnosics. Bars represent performance in z-scores relative to controls (Grace's score on CFPT Aggressiveness was -5.9, but for ease of comparison we kept the scale minimum at -4). Note the difference scales between identity and trait tests. Stars show significant differences to controls (Crawford's modified t test). Facial identity deficits were confirmed by the Cambridge Face Memory Test (CFMT, Duchaine & Nakayama, 2006), Face Old/New Discrimination Test (Duchaine & Nakayama, 2005) and the Cambridge Face Perception Test (CFPT, Duchaine et al., 2007).

Our results are consistent with and extend previous reports of identity-trait dissociations in prosopagnosia. Todorov and Duchaine (2008) presented four developmental prosopagnosics who made facial trustworthiness judgments in line with controls for a heterogeneous set of faces. Importantly though, when the face stimuli were controlled for emotional expressions and non facial cues such as hairstyle and skin blemishes, two prosopagnosics failed to replicate the normal judgments obtained with richer stimuli. Two prosopagnosics showed normal trait judgments across all stimuli sets, but they were tested on only one trait perception (trustworthiness) and only using ratings. Additionally, findings from individuals who do not have normally developed mechanisms for face perception are not warranted to be informative about the organisation of normal brains.

Quadflieg et al. (2012) extended the number of traits measured and tested PS, a thoroughly studied case of acquired prosopagnosia (Rossion et al., 2003). PS showed normal ratings for several facial traits (aggression, attractiveness, confidence, intelligence, sociability, trustworthiness, typicality) but not for dominance. However, she was tested with faces displaying various hairstyles, and, as Todorov and Duchaine (2008) showed, normal trait judgments of faces including hairstyles do not necessarily translate into normal trait judgments when only facial information is presented. The stimuli were also presented in colour, and skin tone has been shown to influence at least perception of attractiveness (Fink, Grammer, & Matts, 2006; Jones, Little, Burt, & Perrett, 2004; Matts, Fink, Grammer, & Burquest, 2007).

By eliminating some of these potential confounds (e.g. our study presented only grayscale images, cropped so that only facial information was available), we present strong evidence for a dissociation between facial identity and trait perception. Importantly, we used the same sorting format (i.e., CFPT) to examine both identity and trait processing, and some prosopagnosics were impaired only with identity perception. In addition, a critical contribution of our study is the use of three test formats to assess trait judgments. Because several participants performed normally with ratings for a trait but were impaired with our sorting and categorisation tasks, our results suggest the ratings format used in previous studies may have limitations that make it too insensitive for use in neuropsychological investigations: a relatively low average of correlations to consensus judgments (between .42 and .58) and large standard deviations (.16 to .29) in controls. Subtle impairments in trait judgments may be masked by floor effects. For example, Grace showed normal judgments of aggressiveness only for rating, but not for sorting and categorisation. Similar problems with ratings (i.e. relatively low means with large standard deviations) were also noted in the other studies, with mean correlations to consensus judgments typically between .50 and .60, and standard deviations frequently up to .20 (Todorov & Duchaine, 2008; Quadflieg et al., 2012). In general, spurious findings are more likely to arise whenever a single test is used.

The dissociation between facial trait and identity perception noted in acquired prosopagnosia is consistent with findings from neuroimaging and other patient studies which have posited different critical areas for the two abilities. Functional MRI studies suggest that changes in facial identity are primarily reflected in increased activation in the fusiform gyrus (Fox, Moon, Iaria, & Barton, 2009; Rotshtein, Henson, Treves, Driver, & Dolan, 2005; Winston, Henson, Fine-Goulden, & Dolan, 2004), while facial trait evaluations are thought to depend on other areas. For example, variations in perceived attractiveness are tracked by activation changes in the medial and lateral orbitofrontal cortex (Cloutier, Heatherton, Whalen, & Kelley, 2008; Kranz & Ishai, 2006; O'Doherty et al., 2003; Winston, O'Doherty, Kilner, Perrett, & Dolan, 2007), nucleus accumbens (Aharon et al., 2001; Cloutier et al., 2008) and anterior cingulate cortex (Winston et al., 2007; Cloutier et al., 2008). Facial trustworthiness in turn has been linked to amygdala activation by several fMRI studies (Engell et al., 2007; Said, Baron, Todorov, 2009; Todorov et al., 2008; Winston et al., 2002). In addition, Adolphs et al. (1998) presented three patients with complete lesions of both amygdalae (SM, JM, RH) who displayed abnormal perception of facial trustworthiness and approachability. One of them, SM, was able to recognise famous faces (Adolphs, Tranel, Damasio, & Damasio, 1994), learn new faces (Adolphs, Tranel, Damasio, & Damasio, 1995) and discriminate between similar faces (Tranel & Hyman, 1990), providing the reverse side of a double dissociation between facial identity and trait perception from our prosopagnosic patients.

In this context, Florence's results are somewhat puzzling. Although she shows a marked dissociation between facial identity and trustworthiness perception, the dissociation is the opposite of what it should be expected given that she does not have a right amygdala. More specifically, while previous studies featured amygdala patients with impairments in facial trustworthiness judgments but normal identity recognition (Adolphs et al., 1994; Adolphs et al., 1995; Adolphs et al., 1998), Florence displays the reverse pattern, with impaired face recognition but normal trustworthiness judgments. We note that even though currently Florence has lesions affecting a large part of the right anterior temporal lobe (which is linked to face recognition), she became prosopagnosic before that, after the first operation that resected the right amygdala and hippocampus. It is however possible that other damaged regions beyond right amygdala contribute to her severe prosopagnosia. Furthermore, she has an intact left amygdala, while previous cases with abnormal trustworthiness perception had bilateral amygdala lesions. It could be the case that the left amygdala is sufficient for normal trait judgments.

As can be seen in Figure 3.4, the patterns of results for the trustworthiness and attractiveness judgments are quite similar. Considering the high correlation usually found for these judgments (e.g. Oosterhof & Todorov, 2008), one concern for tests of trustworthiness and attractiveness perception is that they measure the same construct (or that responses rely on the same visual cues). However, this is not the case for our tests. When controls sorted the CFPT Trustworthiness and Attractiveness with the instructions switched (i.e. faces from CFPT Trustworthiness were sorted for attractiveness, and faces from CFPT Attractiveness were sorted for trustworthiness), performance was outside normal range. Instead, the fact that judgments of attractiveness – a positive trait – are closely linked to judgments of trustworthiness is consistent with the Oosterhof and Todorov (2008) model of facial trait perception.

Finally, our findings also speak to dissociations within trait judgments. For example, Oosterhof and Todorov (2008) suggest that all trait judgments can be accounted for by two distinct components: trustworthiness (representing positive traits such as attractiveness) and dominance (representing 'status' judgments such as aggressiveness and competence). Our data are consistent with this idea: Grace was normal with trustworthiness and impaired on two of the three tests of aggressiveness, while Sandy showed normal aggressiveness but was impaired with trustworthiness (at least on the CFPT).

3.3. DISSOCIATIONS IN PERCEPTION OF FACIAL TRAITS, EXPRESSIONS AND SEX IN ACQUIRED PROSOPAGNOSIA

We have shown that normal trait perception is possible without intact face recognition. The next question is whether trait perception is dependent on other routes within the face system, such as the expression or sex recognition route. A recent facial trait perception model posits that two primary dimensions, namely trustworthiness and dominance, account for the majority of variance in judgments of multiple facial traits (Oosterhof & Todorov, 2008). Oosterhof and Todorov developed an empirically validated computer algorithm that could orthogonally manipulate perceived trustworthiness and dominance in faces. They observed that increasing trustworthiness made faces appear happy, while decreasing trustworthiness made faces appear angry. Inspired by this observation, they argued that subtle resemblance to happy and angry expressions drives perceptions of trustworthiness of neutral faces, and that trustworthiness judgments are in fact by-products of mechanisms responsible for recognising emotional expressions. This emotion overgeneralisation hypothesis was supported by subsequent studies (Said, Sebe, & Todorov, 2009; Oosterhof & Todorov, 2009; Engell et al., 2010). Similarly, because increasing dominance made faces appear more masculine, while decreasing dominance made faces appear more feminine, judgments of facial dominance were suggested to be by-products of mechanisms involved in recognition of sex (i.e., the sex overgeneralisation hypothesis). Two predictions naturally follow from these hypotheses: (i) individuals with impaired facial expression perception will have deficits with trustworthiness judgments, and (ii) individuals with impaired sex perception will have deficits with dominance judgments.

Acquired prosopagnosics, individuals with severe face recognition problems following brain damage (Bodamer, 1947), provide a powerful means to evaluate the overgeneralisation hypotheses. In addition to their definitional impairment with facial identity, acquired prosopagnosics exhibit a variety of other face deficits. For example, some prosopagnosics are impaired with facial expressions (Fox et al., 2011), while others cannot recognise the sex of a face (Sergent & Signoret, 1992) or both (Rezlescu, Pitcher, & Duchaine, 2012; Rossion et al., 2003; Sergent & Signoret, 1992). A diverse constellation of face deficits other than identity can reveal informative associations and dissociations between mechanisms operating on different facial aspects that may be involved in facial trait processing.

The goal of the following experiments was to test the predictions of the overgeneralisation hypotheses with one prosopagnosic individual, Florence, who showed trait judgments comparable to controls in all nine tests. We examined perception of happy and angry expressions, sex, and perceived trustworthiness and aggressiveness. We chose aggressiveness rather than dominance for several reasons. First, dominance seems more a complex, artificial construct than a natural trait spontaneously inferred from faces. When asked to produce unconstrained evaluations of facial traits, none of the 55 participants in Oosterhof and Todorov (2008) used dominance to describe the faces they saw. Second, aggressiveness had the highest loading (0.66) than any other trait on the dominance dimension of the Oosterhof and Todorov model. Third, dominance is a rather ambiguous construct: it may refer to physical or social attribute and, depending on the context, it can be a positive or a negative trait.

Methods

Participants

Florence, the prosopagnosic individual, was described previously in Sections 2.2 and 3.2. For tests of facial expression and trait perception, details about control groups are provided in Section 3.1 that describes the development of these tests. For CFPT Sex, control participants (n = 36) are described in Chatterjee and Nakayama (2013). For the sex categorisation task (see below), control data were provided by 18 participants (age range 23 – 43, mean age = 28.9 years; 11 female).

Procedure

Florence's perception of face trustworthiness, aggressiveness, happy expression, angry expression and sex was tested with two tests per ability. One was a sorting test (CFPT) and the other a categorisation test (OOO, with the exception of the categorisation test for sex perception described below). The CFPT for sex perception was created by Chatterjee and Nakayama (2013). Each trial presented six versions of one individual varying in terms of how masculine or feminine they looked. The initial scrambled order of the six faces was predetermined. Participants had to sort the faces from most masculine to most feminine. The morph level for each trial was chosen to set overall mean control performance at 75%. Control data came from 36 individuals (Chatterjee & Nakayama, 2013). In the categorisation task for face-based sex perception (Rezlescu, Pitcher et al., 2012), participants had to categorise 60 faces as male or female. Faces were cropped below the eyebrows to avoid reliance on diagnostic individual features. Performance was measured with A' (Macmillan & Creelman, 1991), a bias-free measure that varies between 0.5 and 1.0 with higher scores indicating better discrimination ability between male and female faces.

Results

We used Crawford's modified t-test (Crawford & Howell, 1998) to determine whether Florence's performance deviates from controls' in a statistically abnormal manner. Results are displayed in Figure 3.9.

Florence's results on tests of facial trait perception showed that her ability to judge trustworthiness and aggressiveness in faces is intact despite her prosopagnosia (see previous section).

Florence's perceptions of happy and angry expressions were mixed. She achieved only 61% correct sorts (controls: $M = 75\%$, $SD = 8\%$; $t = 1.72$, $p = .096$) and 53% correct categorisations (controls: $M = 75\%$, $SD = 10\%$; $t = 2.16$, $p = .039$) of happy faces. These results indicate that her perception of happy expression was weaker than controls, bordering on impaired. Florence's perception of angry expression, however, was in the normal range: her scores were 62% for sorting (controls: $M = 77\%$, $SD = 9\%$; $t = 1.64$, $p = .112$) and 81% for categorisation (controls: $M = 77\%$, $SD = 10\%$; $t = 0.39$, $p = .697$). Together, these scores suggest that Florence may have a selective deficit in perceiving happy but not angry expressions.

Florence was impaired on both tests of sex perception. She scored 56% on the sorting task, significantly lower than controls' average of 76% (SD = 9%, t = 2.19, p = .037), and .81 on the categorisation task, substantially lower than controls (M = .97, SD = .02; t = 7.79, p < .001).

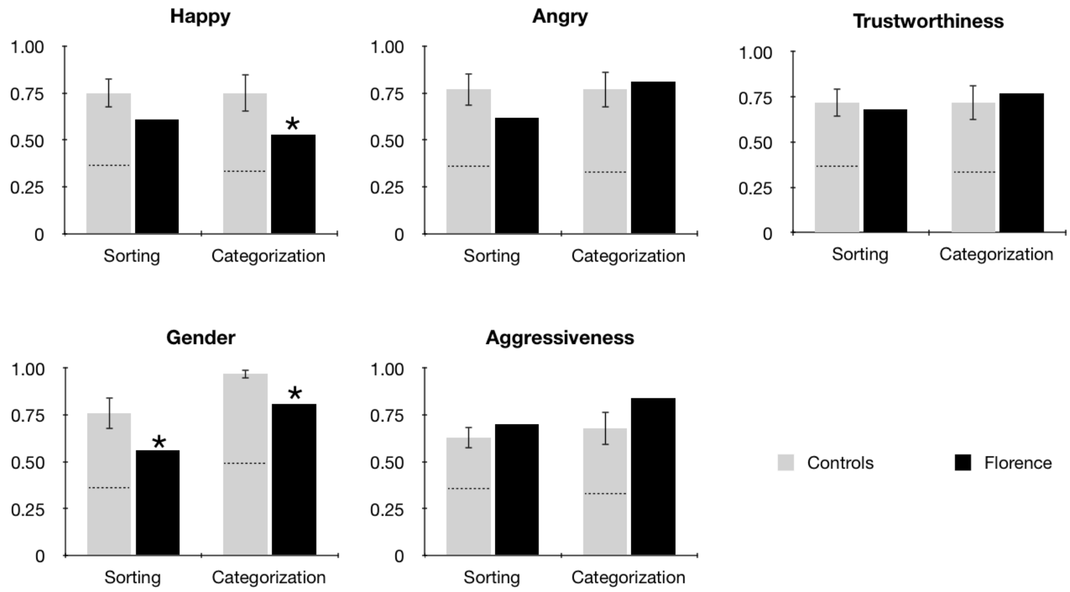


Figure 3.9. Florence's performance in sorting and categorisation tests of expression, sex and trait perception. All scores are expressed in percentages, with the exception of the sex categorisation score expressed as A'. Dashed lines represent chance levels. Error bars show ±1 SD. Stars show significant differences (using Crawford's modified t-test) between performance of Florence and controls.

Discussion

In Oosterhof and Todorov's (2008) computer model of trait perception, varying facial trustworthiness was associated with changes in how happy or angry faces appeared, and varying dominance was linked to changes in how masculine or feminine a face appeared. Todorov and colleagues (Oosterhof & Todorov, 2008; Said, Sebe, & Todorov, 2009; Oosterhof & Todorov, 2009; Engell et al., 2010) suggested that trustworthiness and dominance judgments were by-products of mechanisms responsible for perception of happy/angry expressions and sex, respectively (emotion and sex overgeneralisation theory). Their conjecture predicts that people with expression recognition deficits will also be impaired with trustworthiness judgments, and people with sex recognition deficits will necessarily be impaired with dominance judgments.

We tested these predictions in Florence, a case of acquired prosopagnosia. Despite her severe deficits in recognising identities, Florence was able to make normal judgments of facial trustworthiness and aggressiveness. Critically, these judgments were made in the absence of intact perception of expression and sex, thus challenging the overgeneralisation hypotheses. Florence was impaired at the categorisation test for happy expression, and scored in the lower normal range for two other tests (she was fine at the categorization test for angry expression); although these results do not reflect severe impairments, they question the integrity of her expression perception. For sex perception, results were unambiguous: Florence scored outside the normal range in both tests, indicating severe deficits.

Several considerations suggest our results are robust. Previous studies of facial trait perception in prosopagnosia have only used rating agreements (e.g. Quadflieg et al., 2012; Todorov and Duchaine, 2008). Here we extend these studies by measuring trait perception using two non-rating formats, namely sorting (CFPT) and categorization (OOO), increasing the robustness of our results. We also used the same test formats to assess different aspects of face perception, ensuring that observed dissociations are not caused by different task demands. Our results thus indicate that Florence's normal aggressiveness judgments are the result of perceptual mechanisms distinct from those responsible for sex perception. A dissociation between mechanisms involved in facial trustworthiness and expression perception was partly supported.

GENERAL DISCUSSION

The main goals of the studies reported in this chapter were to investigate a possible dissociation between facial identity and facial trait perception noted in previous studies (Quadflieg et al., 2012; Todorov & Duchaine, 2008) and to test two predictions of the overgeneralisation hypotheses in trait perception: (i) individuals with impaired facial expression perception will have deficits with trustworthiness judgments, and (ii) individuals with impaired sex perception from faces will have deficits with aggressiveness judgments.

We present strong evidence that facial trait perception is dissociable from other aspects of face perception, such as identity, expression and sex recognition. Using three test formats per trait, we showed that four acquired prosopagnosics could provide at least one trait judgment (confirmed by all three tests) in line with controls. One case, Florence, succeeded with all traits measured: trustworthiness, attractiveness and aggressiveness. In contrast, she was impaired with face-based sex recognition and showed mixed performance with expressions of happiness and anger. The results challenge the overgeneralisation hypotheses, according to which trustworthiness and dominance judgments are by-products of mechanisms involved in perception of happy/angry expressions and sex, respectively. Previous studies postulated a critical role for amygdala in facial trustworthiness judgments. Florence does not have a right amygdala, so it is possible that normal trustworthiness perception is achieved solely by the left amygdala. Even though previous imaging studies indicate an involvement of the fusiform gyrus in perceived trustworthiness (e.g. Winston et al., 2002), the results from Florence, Grace and Kepler confirm that (at least the right) occipitotemporal and anterior temporal lobes are not necessary for normal trustworthiness judgments.

Implications for general models of face processing

Our findings suggest that contemporary models of face perception, which tend to focus on the division between subsystems for identity and expression processing (Young & Bruce, 2011), or between subsystems for processing of invariant and changeable aspects in faces (Haxby & Gobbini, 2011), need to broaden their account to explicitly incorporate processing of facial traits. Facial traits may be handled by the subsystem for processing changeable face aspects, because the amygdala (which has been implicated in processing of facial trustworthiness, e.g. Adolphs et al., 1998; Winston et al., 2002) receives information from mechanisms that are responsible for processing dynamic information in faces such as expression and gaze (Haxby & Gobbini, 2011), and because visual aftereffect studies have observed overlapping neural representations of facial trait and facial expression (Engell et al., 2010). Alternatively, because trait is an enduring and invariant feature of the face (repeated judgments of facial traits are highly correlated, Oosterhof & Todorov, 2008), they may be processed by the subsystem that carries out invariant computations in faces, most notably those for recognising identities. A third possibility is that facial trait is primarily processed by its own dedicated subsystem.

That Florence (and other prosopagnosics) was able to provide normal trait judgments despite severe deficits with facial identity and sex recognition, as well as compromised perception of expression, is most consistent with the third possibility. However, her results are also consistent with the idea of a finer division within the invariant subsystem. Perhaps there is one set of overlapping mechanisms responsible for processing all kinds of invariant aspects in faces, only that some mechanisms are more involved in analysing facial identity and others more involved in analysing facial trait. This idea is consistent with evidence from single-cell studies in monkeys, showing that the same set of face neurons in the face-selective middle face patch respond to multiple kinds of basic facial cues (e.g., eye height, face aspect ratio) (Freiwald, Tsao, & Livingstone, 2009). If this idea is correct, facial identity and facial trait perception are identical from a computational point of view; they differ only in terms of which facial cues are more informative for one task over the other (see next section).

What explains the identity vs. trait dissociation?

What might account for Florence's impaired identity but normal trait perception? One explanation is that facial trait judgments, unlike facial identity recognition, do not vary dramatically with subtle changes in facial structure. This explanation is consistent with the nature of the tasks on which Florence was not able to perform normally: Florence was impaired or scored in the lower normal range when she had to sort highly similar face exemplars from the same morph continuum (CFPT Identity, CFPT Happy, CFPT Angry, CFPT Sex). Moreover, Florence was also impaired on a match-to-sample task involving pairs of morphed faces (Susilo et al., 2013). These results suggest that identity recognition involves computations over subtle face cues, whereas trait judgments do not. Note that this does not necessarily mean identity computations are more difficult than trait computations, because in this study they have been matched.

Another possibility is that identity and trait computations involve dissociable cues. One framework relevant to this issue is face space, which suggests that in a psychological space, all faces are mapped and organised according to their values on multiple dimensions that represent the difference between those faces (Valentine, 1991). Face space has traditionally been used to explain how facial identities are recognised (Rhodes & Jeffery, 2006), but it has recently been adapted to account for trait perception as well as social evaluations in faces more generally (Said, Dotsch, & Todorov, 2010). Recent studies have observed that neural responses in face-selective regions are primarily driven not by computations regarding certain aspects of the face (e.g., identity vs trustworthiness), but rather relatively simpler computations of face typicality, or how far and distinctive a face is from an "average" face in the centre of the space (Said, Haxby & Todorov, 2011), making the distinction between identity and trait processing less clear.

CONCLUSION

Using acquired prosopagnosia, we showed that normal trait perception can occur without facial identity recognition. Furthermore, we presented evidence that facial trait perception dissociates from expression and sex recognition, challenging the emotion and sex overgeneralisation theories in trait perception.

Chapter 4
TRUSTWORTHINESS IMPRESSIONS

This chapter concerns four topics in face trustworthiness judgments; the experiments addressing each topic are summarised in separate sections.

It has recently been proposed that perceived face trustworthiness depends on facial width-to-height ratio (WHR), independent of attractiveness (Stirrat & Perrett, 2010). In **Section 4.1** I report three experiments aiming to replicate and extend this finding. The experiments explore implicit and explicit impressions of trustworthiness for two sets of faces: the original stimuli created by Stirrat and Perrett (2010) (male only) and a new set of standardised faces (male and female) manipulated for facial WHR according to the metrics from Stirrat and Perrett (2010). Possible confounds (targets' attractiveness and femininity) and individual differences in terms of propensity to trust, risk-aversion, dominance and attractiveness are also considered.

Face trustworthiness judgments influence consequential decisions (e.g. Rezlescu, Duchaine et al., 2012), but are they valid? Previously, to establish their accuracy, face impressions were typically compared with self-reported personality measures (Bond et al., 1994) or behaviour in laboratory experiments (Fetchenhauer et al., 2012; Gollwitzer et al., 2012). In **Section 4.2** I use data from a high-stake television program in which contestants have an incentive to deceive to examine whether deceptive behaviour can be inferred reliably from faces (static and dynamic stimuli). The predictive power of face impressions in this real situation is compared with their predictive power in a laboratory-based economic game.

Face trustworthiness has been shown to influence economic interactions when people do not have other information about potential partners (e.g. van't Wout & Sanfey, 2008). However, this is not a realistic scenario; in real life, people have access to more than just the face. **Section 4.3** explores the role of face trustworthiness on investment decisions when reliable information about economic partners is available.

Voices are rich social stimuli, in many respects similar to faces. Therefore, it is likely that reliable trait impressions are also formed based on voices. **Section 4.4** is concerned with voice impressions and their integration with face impressions when forming a person impression. The focus is not solely on trustworthiness, but extends to include perceived attractiveness and aggressiveness. Possible correlations between face and voice-based impressions are also investigated.

4.1. FACIAL WIDTH-TO-HEIGHT RATIO AND PERCEIVED TRUSTWORTHINESS

Previous studies suggested explicit and implicit evaluations of face trustworthiness can be explained by face attractiveness (Oosterhof & Todorov, 2008; Wilson & Eckel, 2006), resemblance between observer and target (DeBruine, 2002), structural similarity to expressions of happiness and anger (Oosterhof & Todorov, 2009) or brown eyes (Kleisner et al., 2013). Some of these signals may be deceptive though; Takahashi, Yamagishi, Tanida, Kiyonari, & Kanazawa (2006) found that male faces judged to be more attractive were more likely to abuse trust in economic games.

Recently, a sexually dimorphic (Weston et al., 2007) facial metric related to testosterone level (Verdonck et al., 1999) - the facial width-to-height ratio (WHR) - was shown to predict male aggressive behaviour in the lab and on the ice-hockey ring (Carré & McCormick, 2008). WHR also predicted face-based evaluations of aggressiveness (Carré et al., 2009; Carré et al., 2010). Consistent with the negative correlation between dominance/aggressiveness and trustworthiness judgments (Oosterhof & Todorov, 2008), Stirrat and Perrett (2010) found that facial WHR also predicted perceived and actual trustworthiness of male students. Female participants high on a self-reported scale of submissiveness were more sensitive to WHR in their trustworthiness evaluations of male faces. However, the relation between facial WHR and trustworthy behaviour is not straightforward. Depending on the context, men with wide faces were found to engage in both more and less prosocial behaviour than men with narrow faces (Stirrat & Perrett, 2012).

The following experiments present further investigations of the links between facial WHR and face trustworthiness.

EXPERIMENT 1[†]

This experiment aimed to replicate the original findings relating facial WHR to perceived trustworthiness (Stirrat & Perrett, 2010) with a new, better-controlled set of stimuli. The stimuli were computer-generated faces created to match the facial WHR of the original faces. They included both male and female faces to examine if the presumed effect of WHR on evaluations of male faces extends to female faces. We also examined if individual differences in terms of propensity to trust, risk-aversion, dominance and attractiveness predicted differences in trustworthiness evaluations. The face stimuli were binned in two groups, one with 'narrow' faces (i.e. low WHR) and the other with 'wide' faces (i.e. high WHR). The groups matched in terms of attractiveness and femininity. We wanted to control for attractiveness because attractiveness correlates highly with explicit ratings of trustworthiness (Oosterhof & Todorov, 2008) and was found to be a strong predictor of implicit trustworthiness measured by investments in trust games (Wilson & Eckel, 2006). In Stirrat and Perrett (2010), attractiveness explained five times more variance (30% compared to 6%) in trust decisions than facial WHR. Increased cooperative behaviour was also associated with femininity (Perrett et al., 1998).

Methods

Participants

Twenty participants (age range: 19 to 36 years old, mean age: 26.6 years; 13 female) were recruited from the University College London's subject pool. Participants were paid £5 at the conclusion of the second session. To increase motivation, one randomly selected participant received a bonus of £10 (see

[†] The results of this experiment were previously reported in the MSc thesis of Katarzyna Borowska, supervised by Adam Harris and Constantin Rezlescu.

procedure for details).

Face stimuli

One hundred and sixty facial characters (80 female) were randomly generated using FaceGen Modeller 3.3 software (Singular Inversions, 2008), with the following constraints: Caucasian faces, average age 20 years old, caricature features between attractive and typical, and asymmetry between symmetrical and typical. We then used available controllers in FaceGen to create one face version with low width-to-height ratio (WHR) and one face version with high WHR per character (Figure 4.1), by: i) narrowing the face and/or increasing the distance between eyes and lips; ii) widening the face and/or decreasing the distance between eyes and lips. From this set of 320 faces we eliminated characters for which the narrow or wide faces looked unrealistic and characters who looked very similar to each other (for a wider diversity). Furthermore, we selected only the characters with WHR for the narrow faces between 1.65 and 1.85, and WHR for the wide faces between 1.85 and 2.05. Per character, wide faces had an WHR which was on average 1.10 times larger than that of the narrow faces (range: 1.05 to 1.15), which was comparable to the WHR difference between wide and narrow face stimuli in Stirrat and Perrett (2010). In the end, we retained the narrow and wide faces from 40 male and 20 female characters.

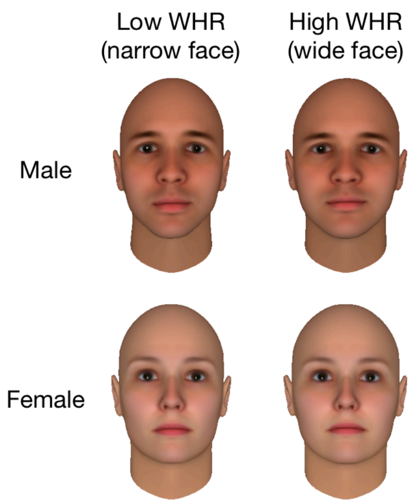


Figure 4.1. Examples of computer face stimuli varying on WHR. Two face versions, one with low WHR ('narrow face') and one with high WHR ('wide face'), were created for each male and female character.

The female faces were then rated for attractiveness and the male faces for attractiveness and femininity/masculinity by 104 online volunteers recruited on Psychological Research on the Net (http://psych.hanover.edu/research/exponnet.html). The femininity scale ranged from 1 (very feminine) to 7 (very masculine) and the attractiveness scale from 1 (not at all attractive) to 7 (very attractive). Each participant rated either the narrow or wide version (randomly determined) of each character. Attractiveness ratings were lower for male faces (M = 2.70) than for female faces (M = 3.14). To match attractiveness level across male and female faces, we ordered the male faces on attractiveness and further selected only the ten faces immediately below the average attractiveness level for female faces (i.e. 3.14) and the ten faces above that level. Our final set of stimuli consisted of narrow and wide faces from 20 female and 20 male characters. Table 4.1 presents details about facial WHR, perceived attractiveness and femininity. Note that the four groups of stimuli (male narrow, male wide, female narrow, female wide; each with 20 faces) do not differ in terms of attractiveness, and the two male groups do not differ on femininity/masculinity. Male narrow and female narrow were also matched in terms of facial WHR, as were male wide and female wide. The average ratios between the wide and the narrow versions of the same characters were also comparable for male and female faces. The only significant difference was between the facial WHR for narrow and wide faces, according to our manipulation.

Face stimuli		Facial WHR				WHR Wide/ WHR Narrow				Attractiveness		Femininity	
		Min	Max	M	SD	Min	Max	M	SD	M	SD	M	SD
Male	Narrow	1.71	1.85	**1.81**	0.04	1.04	1.12	1.09	0.02	3.13	0.30	4.98	0.58
	Wide	1.85	2.04	**1.97**	0.05					3.06	0.54	4.87	0.42
Female	Narrow	1.67	1.84	**1.78**	0.04	1.03	1.17	1.09	0.03	3.14	0.69	-	-
	Wide	1.88	2.02	**1.94**	0.04					3.09	0.57	-	-

Table 4.1. Facial width-to-height ratios (WHR), attractiveness and femininity ratings for the selected 80 face stimuli (20 faces per sex per width). Femininity ratings for female faces were not collected. Note that the stimuli groups are matched in terms of facial WHR between sexes, ratios between Wide and Narrow faces, attractiveness and femininity (all *p*s > .05). The only significant differences are for facial WHR between male narrow and male wide ($p < .0001$) and female narrow and female wide ($p < .0001$).

Facial WHR did not correlate with perceptions of attractiveness for neither male nor female faces (absolute *r*s < .14, *p*s > .42). For male faces, attractiveness and femininity were also uncorrelated (r = -.20, p = . 213).

Procedure

The experiment had two parts. In the first session subjects were asked to play a series of trust games with faces seen on the screen. The trust games were introduced as part of an online investment game played simultaneously between students at various universities. Participants were told they would play the role of the investor, and decide whether to invest their money in the trustees whose faces they would see on the screen. On each trial, participants had 100 virtual pounds (VP) available for investing. If the decision was to invest, this amount tripled automatically (i.e. became 300 VP) by the time it reached the trustee. Participants were told it was then up to each trustee to return or not half of that tripled amount to the investor. Therefore participants had an incentive to invest only in trustees looking trustworthy (i.e. likely to share the money). If participants decided not to invest, they "banked" the 100 VP and moved on to the next trustee. In sum, each of our trust games allowed for three outcomes: i) participants invested 100 VP and trustee shared the tripled amount; in this case, the investment was successful, participants made a profit of 50 VP and banked 150 VP; ii) participants invested 100 VP but the trustee did not share the money; the investment is unsuccessful, participants lost 100 VP and banked 0 VP; iii) participants did not invest; in this case, they banked 100 VP. To ensure participants believed they were interacting with real people, before starting the experiment each participant was photographed, their image uploaded in FaceGen and processed to create a "computerised" version of their face, similar to the face stimuli in the experiment. Then they had a few practise trials to allow familiarisation with the rules of the game. During these practice trials, participants were given feedback about each trustee's decision (to share or not) and the end result of their decision (how much they 'banked'). No feedback was provided during the main experiment, but participants were told that a summary for all trials would be presented at the end and that the most successful investor (the one who accumulated the largest sum of money on all trials) would receive a bonus of £10.

The trustees seen during the experiment were in fact faces of the computer-generated characters described in the previous section. The 20 male characters and the 20 female characters were split in two equal groups per sex. Each participant saw the narrow faces (n = 10) of one group of male characters and the wide faces (n = 10) of the other group of male characters. The same was valid for female characters. This allowed for a 2 x 2 within-subjects factorial design, with facial WHR (low and high) and gender (male and female stimuli) as independent variables. The dependent variables were the judged probability of reciprocation and number of 'invest' decisions.

The judged probability of reciprocation was a continuous implicit measure of perceived trustworthiness. After seeing the face of the trustee, participants were asked to provide an estimate between 0% and 100% of how likely the respective trustee was to reciprocate. Then participants had to decide whether or not to invest in that trustee. The 'invest' decision was a binary variable of perceived trustworthiness; it also showed whether participants were likely to make consequential decisions in line with their estimations of trustee's trustworthiness.

The second session of the experiment was scheduled seven to ten days after the first session. Participants were told they would play the same trust games again, with the same trustees, and were asked to try to give the same answers as they did in the first session. The session was presented as a memory test. However, the assumption was that participants would not remember their decisions, making the second session effectively a re-run of the first session. Crucially, in the second session the face versions from each character were flipped. Thus, if the narrow face of a character was presented in the first session, the wide face of the same character was presented in the second session. This allowed us to directly measure the impact of facial width of investment decisions; for example, if a participant did not invest in the narrow face of a character in the first session but invested in the wide face of the same character in the second session, it can be concluded that it was the facial width that affected his decision. If participants' decisions to invest did not change from one session to the other, there are two possibilities: facial width did not affect investment decisions, or participants remembered their decisions from the first session (the narrow and wide faces of the same character are similar and may be perceived as the same faces). However, the time interval between the two sessions made the possibility of accurately remembering the decisions linked to 40 highly-standardised, computer-generated faces extremely unlikely. During debriefing, participants confirmed they could not remember their actions from the first session.

At the end of the second session, participants were asked to self-rate their perceived attractiveness and (social and physical) dominance, and fill in three questionnaires measuring propensity to trust, risk-aversion and submissive behaviour (Appendix 1). All participants were debriefed by email several days after all testing had concluded to avoid contamination of the subject pool.

Results

In general, when judging the probability of reciprocation, participants made a good use of the scale available (0%-100%), with 14 participants (out of 20) having a difference between the maximum and minimum judged probability larger than 50%. In the first session, a 2 x 2 repeated-measures ANOVA revealed a main effect of facial WHR [$F(1,19) = 5.89$, $p = .025$, partial $\eta^2 = .24$], with participants estimating wide faces to be more likely to share money (49%) than narrow faces (47%). There was no main effect for face gender [$F(1,19) = 0.01$] or interaction effect between face gender and WHR [$F(1,19) = 0.34$].

If participants evaluate wide faces more favourably than narrow faces as suggested by the results of the first session, we would expect to find a main effect (in the same direction) of facial WHR in the second session as well. If, on the contrary, participants were able to remember the evaluations linked to each character from the first session, we would expect to find a main effect of facial WHR in the opposite direction, with narrow faces judged more favourably (because characters shown with wide faces in the first session were shown with narrow faces in the second session, and vice versa). However, none of these predictions turned out to be true. In the second session there were no significant main or interaction effects (all *F*s < 0.11).

In terms of investment decisions, we note that one participant did not invest in any of the faces he saw on the screen in session one or two, and one participant invested in all instances except one face in the second session. The other 18 participants varied their investment decisions. In the first session, on average, participants decided to invest in 9.10 (SD = 4.01) narrow faces and 10.15 (SD = 4.22) wide faces (from the 20 faces seen in each category). The average number of investments in male faces was 9.90 (SD = 4.67) and in female faces 9.35 (SD = 4.65). A 2 x 2 repeated-measures ANOVA revealed no significant main effect for face gender [$F(1,19) = 0.22$], or face width [$F(1,19) = 2.48$, $p = .132$], and no interaction effect [$F(1,19) = 0.08$].

The pattern of investment decisions in the second session was very similar to the first session. First, a 2 x 2 repeated-measures ANOVA revealed no significant main effect for face gender [$F(1,19) = 0.03$], for face width [$F(1,19) = 0.13$], and no interaction effect [$F(1,19) = 2.16$, $p = .158$]. Second, a 2 x 2 x 2 repeated-measures ANOVA with session (first or second) as the third independent variable revealed no significant main or interaction effects (all ps > .25).

One concern in our experiment might be that participants remembered their decisions in the first session and stuck to them in the second session. However, this was not the case. On average, participants switched their investment decisions from session one to session two for 14.3 characters (of the 40 characters; 20 would be chance). Notably, most of the changes were from ‘invest’ in wide faces to ‘not invest’ in narrow faces (7.9) rather than the reverse (6.4) as findings of Stirrat & Perrett (2010) would suggest. The difference was not significant ($p > .05$).

We saw that the binary variables facial WHR (low/high) and sex (male/female) do not influence investment decisions or implicit trustworthiness judgments. We next analysed whether other variables had an effect. For this, we computed the total number of ‘invest’ decisions and the average estimated probability of reciprocation per face across the two sessions and correlated these with the continuous variables: facial WHR (i.e. the precise ratio not the binary variable used above), perceived attractiveness and perceived femininity (only for male faces). For female faces, estimated reciprocation probability correlated with attractiveness (r = .39, p = .013), but not with facial WHR (r = .05, p = .779). Number of 'invest' decisions did not correlate with either attractiveness or facial WHR (*p*s > .24). For male faces, estimated reciprocation probability correlated with perceived femininity - more masculine faces were trusted less (r = -.43, p = .006) but not with facial WHR (r = .06, p = .738) or attractiveness (r = .17, p = .306). Perceived femininity was also the only variable to correlate with investment decisions (r = -.32, p = .042) (Table 4.2).

Stimuli		facial WHR	attractiveness	femininity
Male faces	reciprocation probability	.06 (*.738*)	.17 (*.306*)	**-.43 (*.006*)**
	invest' decisions	.01 (*.944*)	.18 (*.277*)	**-.32 (*.042*)**
Female faces	reciprocation probability	.05 (*.779*)	**.39 (*.013*)**	-
	invest' decisions	.19 (*.247*)	.06 (*.704*)	-

Table 4.2. Facial WHR, attractiveness, femininity and perceptions of trustworthiness. Implicit trustworthiness measures did not correlate with facial WHR, but were predicted by perceived attractiveness (for female faces) and femininity (for male faces). The numbers are Pearson's correlations (p values in parentheses). Significant correlations are in bold. Note that "reciprocation probability" refers to the estimates provided by participants in their role as investors in trust games, and not to the actual probability of reciprocation by the trustees (in our experiment, trustees were fictional).

Participants' age, gender, propensity to trust, risk aversion, self-rated dominance and attractiveness did not influence reciprocation probability or investment decisions. Participants' submissiveness correlated with both reciprocation probability (r = .36, p = .005) and number of investments in female faces (r = .22, p = .005), showing that submissive individuals were more likely to trust female faces.

Discussion

Following Stirrat and Perrett's (2010) suggestion that facial WHR influences face trustworthiness judgments, we investigated whether people take potentially costly actions based on facial WHR. More specifically, we measured the number of 'invest' decisions in trust games featuring trustees whose faces were manipulated for WHR. The trustees were also judged for probability of reciprocation (another implicit measure of trustworthiness). In contrast with Stirrat and Perrett (2010), our results indicate that faces with low WHR are not favoured in decisions based on trust. If anything, we noted a trend in the opposite direction, with wider faces judged to be more likely to reciprocate trust. Increasing facial WHR also tended to determine more positive switches (from 'not invest' to 'invest') than decreasing it.

Instead, our measures of implicit trustworthiness negatively correlated with perceived masculinity of male faces, i.e. more feminine faces were perceived as more trustworthy. This result is consistent with previous findings showing that increased masculinity of male faces decreased perceptions of cooperativeness (Perrett et al., 1998).

What can be the reasons behind the different results related to the effect of facial WHR on perceived trustworthiness obtained by Stirrat and Perrett (2010) and us? One possibility concerns the different stimuli used. Our stimuli were computer generated, while Stirrat and Perrett (2010) used natural-looking faces. Furthermore, manipulation of WHR in our stimuli occurred mainly by increasing the facial width ('stretching' the face horizontally), while in Stirrat and Perrett (2010) the WHR was modified predominantly by increasing the facial height ('stretching' the face vertically).

However, another interesting possibility is that the trustworthiness-based choices in Stirrat and Perrett (2010) were not driven by the facial WHR, but by the apparent femininity. Narrower faces in Stirrat and Perrett (2010) tended to appear more feminine; if facial WHR correlated with perceived femininity and this in turn influenced participants‘ responses, facial WHR might appear as a predictor of trustworthiness choices if femininity is not partialled out. To investigate this possibility, we attempted an exact replication of Stirrat and Perrett's (2010) study with an extra variable collected. After choosing the most trustworthy-looking faces from various pairs, participants were also asked to select the most feminine-looking from the same pairs.

EXPERIMENT 2

Methods

Participants

We recruited 109 online volunteers through Psychological Research on the Net (http://psych.hanover.edu/research/exponnet.html).

Face stimuli

We used a subset of the original stimuli presented in Stirrat & Perrett (2010). The stimuli were of 12 male composite characters. Each image was manipulated in shape to create one version with high WHR and one version with low WHR (more details about the stimuli are given in Stirrat & Perrett, 2010).

Procedure

Participants were presented with pairs of faces consisting of the two versions (narrow and wide) of each character, and were asked to indicate which image looked: i) more feminine; and ii) more trustworthy. The trials were blocked according to the trait judged; femininity judgments were first. Position of narrow/wide faces to the left or right of the screen was randomised.

Results

Of all participants, 54 (50%) chose more often the narrow faces and 43 (39%) chose more often the wide faces when judging femininity (12 chose an equal number of narrow and wide faces). For trustworthiness judgments, 56 (51%) chose more often the narrow faces and 42 (39%) the wide faces (11 showed no preference). Chi-square tests revealed both distributions were significantly different from chance (p = .020 and p = .007, respectively).

For each participant and dimension measured, we calculated the proportion of choices for narrow faces (chance level: 50%). On average, participants showed a significant preference towards narrow faces when judging femininity: M = 55%, SD = 24%; $t(108) = 2.30$, $p = .023$. Similarly, when judging trustworthiness, participants preferred on 53% (SD = 26%) of the trials the narrow faces, although the result was not significantly different from chance: $t(108) = 1.24$, $p = .218$.

In 9 out of 12 pairs, the narrow face was chosen more often for both perceived femininity and trustworthiness. There was a large correlation between the number of times a narrow face was judged more feminine in a pair and the number of times the same narrow face was judged more trustworthy: $r = .90$, $p < .001$.

Discussion

We replicated Stirrat and Perrett (2010) findings that faces with lower WHR are judged as more trustworthy. Additionally, we also found a preference towards perceiving faces with lower WHR as more feminine. The high correlation between choices for trustworthy and feminine faces indicates there is a strong link between the two dimensions. Therefore, it is possible that trustworthiness judgments depend on variations in perceived femininity rather than facial WHR. In Stirrat and Perrett (2010), attractiveness was excluded as a potential confound (by including it in the regression), but femininity was not.

The correlation between responses favouring the narrow faces in judgments of trustworthiness and femininity might have been artificially inflated by our design. Same participants were asked to complete both tasks and thus faces selected in the first task might have become the ‘default’ option for the second task. To exclude this possibility, we conducted a short follow-up experiment in which a different group of participants were asked to complete only the trustworthiness task (which was the second task in the current experiment).

EXPERIMENT 3

Methods

Participants

We recruited 32 online participants (mean age 33.2, 16 female) through

Amazon's Mechanical Turk (www.mturk.com). Their participation was paid $0.10.

Face stimuli

Same as in Experiment 2.

Procedure

In a two-alternative forced choice (2AFC) task, participants indicated which image from pairs of faces consisting of the narrow and wide version of each character looked more trustworthy.

Results and discussion

Of all participants, 14 (44%) chose more often the narrow version and 13 (41%) chose more often the wide version as the more trustworthy face (5 were equally split). Per participant, the proportion of choices for narrow faces was on average 52% (SD = 6%), which was not significantly different from chance: $t(31) = 0.48$, $p = .634$. However, the number of times a narrow face was judged to be more trustworthy in this experiment correlated highly with the number of times the same face was judged to be more feminine in experiment 2: $r = .83$, $p = .001$. This is compelling evidence for a strong correlation between femininity and trustworthiness judgments for the face stimuli from Stirrat & Perrett (2010) and lends additional support to the possibility that facial WHR does not influence perceived trustworthiness directly, but through changing perceived femininity.

GENERAL DISCUSSION

Experiment 1 showed no clear preference for faces with low WHR when participants were asked to make investments in trust games, implying that narrow faces were not perceived as more trustworthy than wide faces. If anything, we noted the opposite trend, with wide faces attracting more trust decisions. Our results are inconsistent with those obtained by Stirrat and Perrett (2010) who found that investments in trust games negatively correlated with facial WHR of the male counterparts, and that narrow face versions were preferred to wide face versions in 2AFC paradigms.

Our stimuli were computer-generated faces, while Stirrat and Perrett (2010) used either natural faces or composite images of natural faces, so one concern is that replication was not successful because of the different nature of the stimuli. However, we believe this is not the case. First, faces created using FaceGen have been used in numerous studies before and were shown to elicit trait judgments similar to natural faces (e.g. Oosterhof & Todorov, 2008). Second, our stimuli were carefully controlled to match the facial WHRs of the faces used by Stirrat and Perrett (2010). We matched not only the average WHRs for the narrow and wide faces, but also the average WHR difference between face versions derived from the same character.

One difference between our stimuli and Stirrat and Perrett's (2010) stimuli concerned the method by which facial WHR was varied; we varied facial width while keeping facial height relatively constant, while Stirrat and Perrett (2010) did the opposite. It may be that perceived trustworthiness is increased by either increasing facial height (as in Stirrat & Perrett, 2010) or increasing facial width (as in Experiment 1), independent of the facial width ratio. One simple explanation is that people select the bigger face to be more trustworthy. Another explanation may be that increasing facial height and increasing facial width produced similar changes in expressions making the faces appear happier (or less angry). Subtle cues resembling happy or angry expressions were shown to bias trustworthiness perception (Oosterhof & Todorov, 2008).

A more likely possibility suggested by our results is that, at least for male faces, trustworthiness judgments are driven by perceived femininity. Femininity was previously shown to influence perceptions of trustworthiness (Perrett et al., 1998) and female faces are generally perceived as more trustworthy (Kleisner et al., 2013). In Experiment 1, we found a significant correlation between femininity and both investment decisions and judged probability of reciprocation (implicit measures of trustworthiness), but not between facial WHR or attractiveness and our dependent variables. We therefore argue that it is perceived femininity, rather than facial WHR, that is driving trustworthiness judgments.

In Experiment 2 we verified this claim with the original set of faces used by Stirrat and Perrett (2010). First, we successfully replicated the results reported by Stirrat and Perrett (2010), with the narrow faces attracting significantly more choices in terms of perceived trustworthiness. Second, we showed that the narrow faces were also judged to be the more feminine ones from the pairs. Third, and most importantly for our purpose, we found a very high correlation ($r = .90$) between the number of times a narrow face was judged more feminine in a pair and the number of times the same narrow face was judged more trustworthy. Experiment 3 confirmed this high correlation was not an artefact of the within-subjects design used in Experiment 2.

It is important to note we do not argue that facial WHR is unrelated to impressions of trustworthiness for certain faces. In fact, our second experiment confirmed Stirrat and Perrett's (2010) original findings that narrow faces are perceived to be more trustworthy. However, the relation is not direct and the effect appears to be mediated by perceived femininity, such that a preference for narrow faces is not seen any more when faces vary in WHR but not perceived femininity (experiment 1).

In real life, facial WHR may correlate with a more masculine (i.e. aggressive) behaviour. Men with wider faces were found to be more aggressive (Carré & McCormick, 2008) and more likely to exploit trust in laboratory games (Haselhuhn & Wong, 2011; Stirrat & Perrett, 2010). Although these findings were recently challenged (Özener, 2011), there is a plausible biological mechanism proposed to account for the relation between facial WHR and behaviour (Verdonck et al., 1999; Weston et al., 2007), by which testosterone level influences both. We note that our experiments did not examine actual behaviour and thus our results cannot contribute to this debate.

To conclude, we showed that facial WHR is not a direct predictor of face trustworthiness. Instead, facial WHR may be linked to perceived femininity which in turn influences trustworthiness judgments. Future studies should attempt to further disentangle the effects of facial WHR from the effects of perceived femininity on perceptions of trustworthiness. For example, our experiments may be replicated with faces that vary orthogonally on facial WHR and femininity. A simple way to modify perceived femininity of a face is to interfere with the eyebrows, which were shown to be a strong cue to facial sex (Bruce et al., 1993). If our hypothesis is correct, wide faces with less pronounced eyebrows (i.e. more feminine) will be judged as more trustworthy than narrow faces with prominent eyebrows (i.e. more masculine).

4.2. ACCURACY OF FACE TRUSTWORTHINESS JUDGMENTS

INTRODUCTION

Although people tend to agree on their evaluations of faces along many social dimensions (Oosterhof & Todorov, 2008), there is no conclusive evidence to support a correlation between face impressions and the actual personality or behaviour of the targets. Some studies found that certain self-reported personality traits can be inferred from faces (Little & Perrett, 2007; Berry, 1991; Bond et al., 1994; Penton-Voak et al., 2006), as can be membership to religious (Rule et al., 2010), political (Rule & Ambady, 2010) or sexual minority groups (Rule, Ambady, & Hallett, 2009; Rule & Ambady, 2008; Rule, Macrae, & Ambady, 2009). However, other studies found no relations between face impressions and personality or behaviour (Hassin & Trope, 2000; Zebrowitz et al., 1996; Zebrowitz et al., 1998; Pound et al., 2007).

Accuracy of facial trustworthiness judgments received less attention, possibly because of the difficulty to have an objective criterion to measure actual trustworthiness. Some studies (Porter et al., 2008; Valla, Ceci, & Williams, 2011) investigated whether people can infer group membership of targets thought to vary dramatically in their trustworthiness (e.g., criminals and non-criminals). They found slightly higher than chance performance but, in at least one of the studies (Valla et al., 2011), heterogeneity of stimuli facilitated group membership recognition (criminals' photos were mugshots).

However, membership to a criminal group does not necessarily imply lack of trustworthiness (not more than receiving a Nobel prize certifies one's trustworthiness). More recent studies examined whether people can reliably predict trustworthy (i.e., cooperative) behaviour in economic games, such as the trust game, played in the laboratory. Results were mixed. When participants saw short face clips of the targets, usually between 5s and 20s ("thin slices" of behaviour; Ambady & Rosenthal, 1992), some authors reported significant correlations between face trustworthiness impressions and cooperative behaviour (Fetchenhauer et al., 2010), but others could not replicate this effect (Gollwitzer et al., 2012).

Stiratt and Perrett (2010) suggested that people's impressions of trustworthiness vary with the targets' facial width-to-height ratio and that this metric is a significant predictor of reciprocation rates in trust games (signalling trustworthiness). The implication is that cooperative behaviour can also be inferred from static face images. Some evidence supporting this view came from two studies (Bonnefon et al., 2012; Verplaetse et al., 2007) which showed that, under certain conditions (with photos taken at the moment of deciding to cooperate or not; or with black-and-white cropped photos to exclude external cues) participants were better than chance at identifying cheaters and cooperators.

Trust games were shown to be a good proxy for the trust investors place on trustees (Berg et al., 1995), but they are not an ideal measure of trustees' trustworthiness. At no point during the game do trustees commit to returning money to the investors; on the contrary, they are usually told that they have no obligation to do so. In the absence of an explicit commitment to cooperate, it is difficult to classify their behaviour as deceptive or untrustworthy. A better proxy to (un)trustworthiness is provided by situations in which explicit social commitments are broken for a personal gain at the expense of the other party. A television game show broadcasted in the Netherlands in 2002 provided just such a setup (Belot, Bhaskar, & van de Ven, 2012). In the final stage of the show, contestants were asked to play a prisoner's dilemma game for the prize money they had earned up to that point (usually in the order of thousands of Euros). Crucially, before deciding whether or not to share the money, they had an opportunity to make explicit promises to share. These promises were later kept or broken, providing a measure of each contestant's trustworthiness.

Using data from this television game show and from a series of trust games played in the laboratory, we investigated the accuracy of facial trustworthiness impressions under various conditions. In **Experiment 1**, we showed participants still photos of contestants on the television game show and asked them to predict which ones kept their explicit promise to share the prize money. **Experiment 2** aimed to compare face-based judgment accuracy for static (still photos) versus dynamic (5s video clips) stimuli, and for two very different interaction situations: i) a high-stakes prisoner's dilemma game on a television show, and ii) a lower-stakes trust game in a laboratory experiment. The criterion for trustworthiness in the laboratory trust games was the average rate of reciprocation. **Experiment 3** was a within-subjects replication of Experiment 2 using a sub-sample of the face stimuli (both static and dynamic) from the television show.

EXPERIMENT 1

Methods

Participants

Sixty-six participants (41 female, age range: 18 - 62 years, M = 24.1, SD = 7.1) were recruited from University College London's psychology subject pool. They received £3 for their participation. Moreover, an additional £20 was awarded to the most accurate person.

Face stimuli

We used data (recorded in 2002) from 61 episodes of a Dutch television game show called "Will (s)he share or not?" (see Belot et al., 2012). In this show, five players compete for money by answering trivia questions. The least accurate players are progressively eliminated until only the two most successful contestants are left. In the final round of the show, the prize money accumulated by the two remaining players (based on their answers to the trivia questions) is pooled. These two players then play a variant of the prisoner's dilemma game in which they each secretly decide whether to share or 'steal' the pooled prize money. If both players chose to share, their pooled gains are evenly split (i.e., they each get an equal share of the total). If one player decides to share but the other decides to 'steal' then the latter takes all the money while the former goes home empty-handed. Finally, if neither player chooses to share (i.e., they both decide to 'steal') then they both go home empty-handed. Critically for our study, the two final contestants could discuss their strategies with each other and make explicit promises to share (or not) before secretly making their decisions. Two Dutch-speaking judges independently identified 47 Caucasian finalists from the show who made firm and unambiguous promises to share (we ignored finalists who did not make firm commitments to share). For each of these finalists, we extracted a still photo of their face from the beginning segment of the show where each contestant introduced him/herself. The stills we selected were frontal shots with closed mouth, neutral expression and standard lighting (see Figure 4.2). Of the 47 finalists, 19 (5 female) broke their promise and chose to 'steal' (i.e., they were untrustworthy), while 28 (13 female) kept their promise and shared (i.e., they were trustworthy).

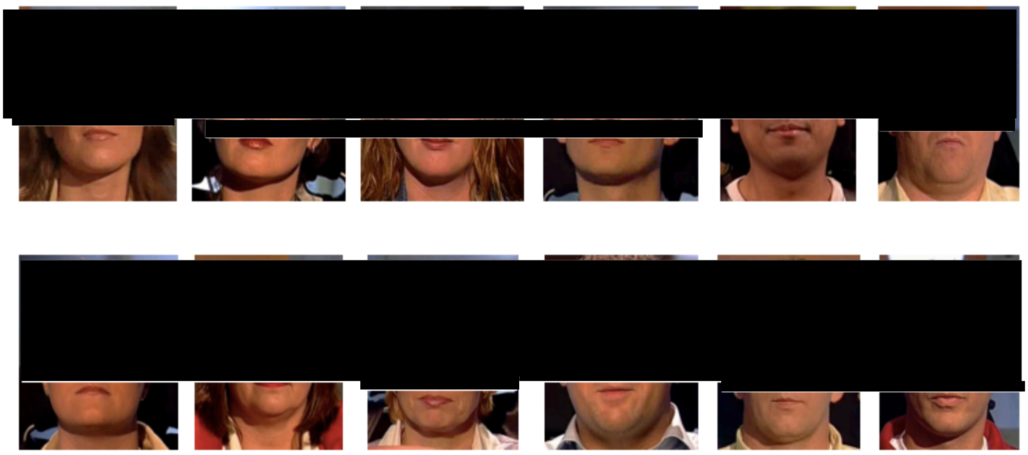


Figure 4.2. Examples of face stimuli from the Dutch television game show. All selected contestants made firm promises to share the prize money. Contestants in the top row kept their promise and shared (trustworthy targets), while those in the bottom row broke their promise and chose to 'steal' (untrustworthy targets).

Procedure

Participants were first given details about the television game show and how the face stimuli were selected. The instructions explained that they would be seeing these photos, presented one at a time, and that they would have to guess whether each person shown in the photo had kept his/her promise to share the prize money. The base rate (i.e., the overall proportion of trustworthy players) was not disclosed. Before starting the experiment, participants were asked to predict their performance - predicted accuracy rate (AR). They were then shown the stimuli, one by one (presentation order was randomised), and asked to guess whether the target had chosen to share or 'steal' (binary variable) by clicking on the corresponding button (the position of the buttons was counterbalanced between participants). After each choice, participants reported how confident they were in their judgment (using a 0–100% scale). No feedback was provided during the trials. After completing all trials, participants were asked to estimate how well they thought they did (estimated AR). At the end, they reported basic demographic characteristics, self-assessed ability to "read" faces (a binary variable), and familiarity with the television program and contestants (all reported seeing the stimuli for the first time).

Results

Results are summarised in Table 4.3. The average accuracy rate (AR = number of correctly identified cheaters and cooperators / number of trials) achieved by participants was 53.2% (*SD* = 8.0%). The result was significantly greater than chance: $t(65) = 3.26$, $p = .002$. A 2×2×2 mixed ANOVA with target behaviour (trustworthy/untrustworthy) and target gender (male/female faces) as within-subject factors and participants gender as a between-subjects factor revealed a significant main effect of target behaviour on accuracy [$F(1,64) = 25.10$, $p < .001$, partial $\eta^2 = .28$], with higher rates for trustworthy targets (57.4%) than for untrustworthy targets (47.1%). No other main effects or interactions were significant (all *p*s > .37).

The ability to infer behaviour from faces was overestimated by participants, of which 66% claimed to be able to "read" faces. The average predicted AR collected before the experiment (63.8%) and the average estimated AR collected after completion of all trials (58.6%) were both significantly greater than chance (both *p*s < .001). Both values were also significantly greater than the actual AR (53.2%): t(63) = 6.02, p = 1.00×10-7, η^2 = .37, and t(65) = 4.40, p = 4.19×10-5, η^2 = .23, respectively. Furthermore, 67% of participants estimated they would be above chance before seeing the stimuli and this percentage remained high (62%) even after seeing the stimuli.

	Accuracy				Discriminability		Response bias	
Targets	Trustworthy	Untrustworthy	ALL	*p*	A'	*p*	B"	*p*
Male faces	.56 (.14)	.47 (.15)	.52 (.10)	*.163*	.53 (.16)	*.194*	**.16 (.36)**	*<.001*
Female faces	.59 (.18)	.46 (.19)	**.56 (.13)**	*.001*	**.55 (.19)**	*.048*	**.22 (.49)**	*.001*
ALL	**.57 (.12)**	**.47 (.11)**	**.53 (.08)**	*.002*	**.54 (.13)**	*.025*	**.20 (.31)**	*<.001*
p	*<.001*	*.029*	*.002*					

Table 4.3. Accuracy, discriminability and response bias in identifying trustworthy and untrustworthy targets from the television game show face stimuli. Significant deviations from chance levels are in bold and underlined (standard deviations are presented in parentheses).

Because of the larger number of trustworthy than untrustworthy targets in our sample, a higher than chance accuracy rate may be achieved if participants have a bias towards positive evaluations. To exclude this possibility, we computed A', a nonparametric measure of judgment sensitivity independent of response bias (Stanislaw & Todorov, 1999). A' typically ranges from .5, corresponding to no ability to distinguish signal from noise, to 1, reflecting perfect discrimination. The average A' in our experiment was .54 (SD = 0.13), which was significantly higher than .5: t(65) = 2.29, p = .025. The average response bias, B", was .20 (SD = .31), showing a significant preference towards providing more "trustworthy" responses: t(65) = 5.10, p < .001.

Individual differences in the ability to discriminate between trustworthy and untrustworthy targets did not correlate with how well participants thought they did. The average point-biserial correlation between response accuracy (correct/incorrect) and confidence level (continuous variable between 0 and 100%) was .03 (SD = .27; p = .353). Furthermore, accuracy rates achieved by participants did not correlate with predicted ARs (r = .07, p = .595) or estimated ARs (r = .06, p =.623).

The inter-rater reliability of responses in our sample was low, Fleiss' kappa = .07. Fleiss' kappa calculates the degree of binary scale agreement between raters over that expected by chance and ranges from less than 0 (chance-level agreement) to 1 (perfect agreement).

Discussion

Our results suggest that people can infer trustworthy behaviour in real-life situations from faces, at slightly better than chance levels. The effect may be driven by individual differences between observers (how good they are at "reading" faces) or targets (how easy to "read" the faces are). If certain observers are better 'face readers' than others, one might expect to see this reflected in their confidence when categorising faces. This is not what we found. In general, confidence in the ability to "read" faces exceeded actual ability, and there was no correlation between confidence and accuracy of responses. If certain faces are easier to "read" than others, this would imply that there are generally useful diagnostic facial cues to trustworthiness, cues that are apparent to most observers. This in turn would imply that observers should agree on who looks trustworthy and who looks untrustworthy. Inconsistent with this idea, our inter-rater agreement measure was relatively low.

In the next experiment we compared accuracy of face trustworthiness impressions from static and dynamic stimuli. In addition, we examined whether our findings replicate with a second set of data: faces and behavioural data from participants in laboratory trust games.

EXPERIMENT 2

Methods

Participants

Eighty participants (43 female; age range 18 - 54 years old, M = 24.4, SD = 6.4) were recruited from the psychology subject pools of University College

London and the University of Warwick. They were paid £3 for their participation.

Face stimuli

We had two sources for our stimuli and we created two types of stimuli from each source: still photos and silent video clips of five seconds. The first source were recordings of the Dutch television program presented in the previous experiment. For each of the 47 Caucasian contestants selected in Experiment 1, who made strong and unambiguous promises to share, we aimed to have two stimuli. The first was a still photo, frontal shot with closed mouth, neutral expression and standard lighting, from the individual presentations at the beginning of the show. The second was a five-second silent clip, frontal shot, cropped from the moment they committed to share the prize money with their partner. Due to certain limitations (e.g., some finalists were not filmed frontally for at least five seconds when making promises to share), we ended up with stills and clips from 38 players (15 female) that promised firmly to their partners to share the earnings. Of these, 19 (10 female) kept their promise and shared, while 19 (5 female) broke their promise and chose to 'steal'.

The second source for our stimuli were recorded interviews of first year Psychology undergraduates students from Bangor University detailing their reasons for choosing to study Psychology. Some of these students took part a few months later in another experiment in which they played trustees in a series of trust games with other students. In trust games (Berg et al., 1995), an investor is endowed with money and given the possibility to invest part or all of it in a trustee, hoping for a higher return. The amount invested is a good proxy for the trust the investor places in the trustee. In line with previous studies (e.g., Stirrat & Perrett, 2010), we used the amount returned by the trustee as a measure of his/her trustworthiness. We selected ten "trustworthy" students (five female) with average return rates above 50% (M = 54.9%, SD = 6.7%) and ten "untrustworthy" students (five female) with average return rates below 50% (M = 26.6%, SD = 6.7%). From their recorded interviews, we created for each student a five-second silent clip and a still photo (neutral expression).

To summarise, we had four groups of stimuli. One set of 38 still photos and one set of 38 clips were created from the finalists in the Dutch television show (we call these the TV stimuli). One set of 20 still photos and one set of 20 clips were created from the interviewed students who played trust games in the laboratory (we call these the Lab stimuli). Within each set, one half of the stimuli were of 'trustworthy' targets and the other half of 'untrustworthy' targets.

Procedure

Participants were allocated, in a counterbalanced order, to one of four conditions, corresponding to the stimuli set seen first: TV stills, TV clips, Lab stills or Lab clips. The second set of stimuli was selected such that the source (TV or Lab-based) and the type (stills or clips) were different from those of the first set. For example, TV stills were followed by Lab clips, while Lab stills were followed by TV clips.

Before the TV stimuli, participants were given details about the TV show and how the stimuli (stills or clips) were selected, including disclosure of the base rates (i.e., 50% trustworthy and 50% untrustworthy targets). They were then shown the stimuli, one by one (presentation order was randomised), and asked to guess whether the target chose to share or 'steal' (binary variable) by clicking on one of two buttons (the position of the buttons was counterbalanced between participants). After this binary choice, participants reported how confident they were in their judgment on an 11-division ruler ranging from "not at all confident" (coded as -5) to "extremely confident" (coded as +5). The division in the middle was coded as 0. In addition, participants were asked to predict their accuracy rate before seeing any TV stimuli (predicted AR) and to estimate their achieved accuracy rate after completing all the TV trials (estimated AR).

Before the Lab stimuli trials, participants received information about the trust games and how we selected the stimuli (stills or clips). They were then shown the stimuli, one by one (presentation order was randomised), and asked to provide an estimate of the return rate of that target (continuous variable, between 0 and 100). Estimated return rates below 50% were coded as "untrustworthy target" responses, while return rates 50% and above were coded as "trustworthy target" responses. After each judgment, participants were asked for their confidence (see previous paragraph).

No feedback was provided during the trials. At the end of the study, all participants reported basic demographic characteristics, self-assessed ability to "read" faces (a binary variable), and familiarity with any faces seen on the screen (all reported seeing the stimuli for the first time).

Results

TV stimuli

Performance (i.e., accuracy of trustworthiness judgments) was measured as proportion of correct responses. Because the base rate of trustworthy to untrustworthy targets was 50/50, the percentage of correct responses can be considered an unbiased indicator of the ability to discriminate between trustworthy and untrustworthy targets (i.e. potential response biases do not artificially increase/ decrease the computed sensitivity).

Our experimental design ensured that each stimuli set was rated as the first set by 20 participants and as the second set by another 20 participants. Independent t-tests failed to find significant differences between the accuracy of participants judging the TV stills first versus participants judging the TV stills second [$t(38) = 0.48$, $p = .631$], or between participants judging the TV clips first versus participants judging the TV clips second [$t(38) = 1.62$, $p = .1451$]. Likewise, independent t-tests failed to find differences in performance related to participants' gender [TV stills: $t(37) = .27$, $p = .786$; TV clips: $t(36) = .06$, $p = .950$]. Therefore, for the following analyses we collapsed all data across order and gender, such that we had 40 participants judging the TV stills and 40 participants judging the TV clips.

As expected, estimated AR and predicted AR correlated (for TV stills: $r = .78$, $p < .001$; for TV clips: $r = .50$, $p = .001$). More importantly, one-sample t-tests revealed they were all significantly higher than chance level (50%), an indication of participants' confidence in their judgments. There were no significant differences between estimated AR and predicted AR for either stills or clips (all scores and significance levels are presented in Table 4.4).

In terms of response accuracy, the average percentage of correct responses was 50.3% for TV stills - indistinguishable from chance [$t(39) = .27$], and 54.5% for TV clips - significantly higher than chance level [$t(39) = 3.77$, $p = .001$]. Accuracy for TV clips was also significantly higher than accuracy for TV stills [$t(78) = 2.48$, $p = .016$]. A mixed 2×2 ANOVA revealed in addition a significant interaction effect between stimulus gender and type [$F(1,78) = 5.16$, $p = .026$, partial $\eta^2 = .06$], with discrimination of male stimuli more successful from clips (56%) than from stills (48%).

To analyse the distribution of participants' responses, we coded each "trustworthy" response as 1, and each "untrustworthy" response as 0. Therefore, an equal distribution of responses between the two options would be reflected by an average of 0.5. There was a slight bias towards selecting the "trustworthy" option for both stills [$M = 53.2\%$; $t(39) = 2.56$, $p = .014$] and clips [$M = 52.8\%$; $t(39) = 2.02$, $p = .050$]. An independent t-test revealed no difference in response bias between stills and clips.

We next examined whether there was any correlation between participants' confidence and their accuracy, trial by trial and at an aggregate level. For each participant, we computed an average of their confidence levels across all trials and correlated these averages from all participants with the average accuracy scores. Confidence scores ranged from -5 (not at all confident) to +5 (extremely confident). The average confidence across all trials for TV stills was 0.48 ($SD = 1.63$), while for TV clips was 0.81 ($SD = 1.28$) (no significant difference between the two groups). The average confidence scores per participant correlated with accuracy scores neither for stills ($r = .14$, $p = .387$) nor for clips ($r = .05$, $p = .762$). For each participant, we computed the point-biserial correlation between confidence (continuous variable) and response accuracy (correct/incorrect). The average point-biserial correlations for TV stills and TV clips were close to zero, .01 for each group (both *p*s > .55).

TV Stimuli	Estimate Before	Estimate After	Accuracy	Response	Confidence	Correlations: confidence & correct (per trial)	Correlations: avg confidence & avg correct
Stills	**.55 (.15)** *.045*	**.58 (.15)** *.002*	.50 (.08) *.788*	**.53 (.08)** *.014*	.48 (1.63)	.01 (.14) *.648*	.14 *.387*
Clips	**.57 (.14)** *.002*	**.60 (.12)** *<.001*	**.55 (.08)** *.001*	**.53 (.09)** *.050*	.81 (1.28)	.01 (.14) *.558*	.05 *.762*
p	*.081*	*.198*	***.016***	*.834*	*.316*	*.939*	

Table 4.4. Summary of participants' performance at identifying trustworthy and untrustworthy targets in the TV show. The face stimuli were still photos or 5s clips. p values to the right (in italics) reflect differences to chance levels, while p values at the bottom reflect differences between performance with stills and clips. Significant results are in bold and underlined.

Lab stimuli

In contrast to the categorical response recorded for the TV stimuli, the response asked for the Lab stimuli was continuous, with possible values between 0% and 100%. However, based on these values, we also computed a categorical variable - response category -with values 1 (i.e., "trustworthy target") for responses of 50% and above, or 0 (i.e., "untrustworthy target") for responses below 50%. Similarly, based on their actual return rates during the trust games, the targets were classified into two categories: "trustworthy", if the actual return rate was 50% or above (half of them), and "untrustworthy", if the actual return rate was below 50% (the other half). To measure the accuracy of participants' response we computed two indicators: the correlation between participants' response and targets' actual return rate, and the accuracy of categorisation (similar to the one computed for the TV stimuli), calculated as the percentage of matches between the participants' response category and the targets' actual return rate category (i.e., trustworthy-trustworthy or untrustworthy-untrustworthy). Responses from participants seeing the Lab stimuli first and from participants seeing the Lab stimuli second were comparable ($p = .511$) and thus pooled together.

Overall, across both stills and clips, the average response given by participants (i.e., estimated reciprocation rate) was 43.1% (SD = 15.7%). This was surprisingly accurate, considering that the average targets' actual return rate was 40.7%. Estimated reciprocation rates did not vary between stills and clips (see Table 4.5). Accuracy of responses may be estimated by computing the correlation between estimated and actual return rates. This correlation was .11 (p = .019) for stills, and -.03 (p = .353) for clips. The correlation for stills was significantly higher than for clips (p = .013).

Next we analysed the computed categorical variables. Categorisation accuracy was not significantly different from chance for either stills or clips (although approaching significance for stills, see Table 4.5). Response category indicated a bias towards estimated rates below 50% (i.e., "trustworthy" responses) for stills, but not for clips. Confidence levels did not differ between stills and clips, and there was virtually no correlation between confidence and categorisation accuracy (Table 4.5). Also, the average accuracy obtained by each participant did not significantly correlate with average confidence level (Table 4.5).

							Correlations			
Lab Stimuli	Estimated reciprocation rate	Correlation btw. response and actual return rate		Categorisation accuracy		Response category		Confidence	confidence & categorisation accuracy (per trial)		avg confidence & avg accuracy	
Stills	.41 (.15)	**.11 (.30)**	*.019*	.53 (.10)	*.053*	**.41 (.28)**	*.045*	.19 (1.78)	.01 (.25)	*.849*	.08	*.626*
Clips	.45 (.17)	-.03 (.21)	*.353*	.52 (.07)	*.147*	.46 (.29)	*.407*	.45 (1.65)	.00 (.25)	*.923*	.07	*.647*
p	*.299*	***.013***		*.486*		*.413*		*.504*	*.839*			

Table 4.5. Summary of participants' performance at identifying trustworthy and untrustworthy targets in the Lab trust games. The face stimuli were still photos or 5s clips. p values to the right (in italics) reflect differences to chance levels, while p values at the bottom reflect differences between performance with stills and clips. Significant results are in bold and underlined.

To better understand the factors behind differences in performance, we submitted our data to a series of 2×2 mixed ANOVA with stimulus type (stills/clips) as the between-subjects factor, and stimulus gender (female/male) as the within-subject factor. The dependent variables were: estimated reciprocation rate, correlation between estimated and actual reciprocation rate, categorisation accuracy and response category (Table 4.6).

For the estimated return rate, there was a significant interaction effect between stimulus type and stimulus gender [F(1,78) = 9.99, p = .002, partial η^2 = .11], with female clips judged as more generous (estimated return rate 47.4%) than female stills (40.5%), male clips (42.4%) and male stills (41.9%). No other effects were significant (*p*s > .08).

For the correlation between estimated and actual return rate, we found a main effect of stimulus type [$F(1,78) = 5.17$, $p = .026$, partial $\eta^2 = .06$] and a strong interaction effect [$F(1,78) = 30.39$, $p < .001$, partial $\eta^2 = .28$]. The correlation was positive for female stills ($r = .26$, $p < .001$) and negative for female clips ($r = -.13$, $p = .007$).

For categorisation accuracy, there was a significant main effect of stimulus gender [$F(1,78) = 8.81$, $p = .004$, partial $\eta^2 = .10$], with higher accuracy for the female stimuli (M = 54.9%) than for male stimuli (M = 50.0%). The interaction effect between stimulus type and stimulus gender was also significant [$F(1,78) = 17.54$, $p < .001$, partial $\eta^2 = .18$], reflecting higher accuracy for female stills (M = 59.0%) compared to female clips (M = 50.8%), male stills (M = 47.3%) and male clips (M = 52.8%). Accuracy for female stills was the only one significantly higher than chance level [$t(39) = 3.98$, $p < .001$; the other three ps were above .09].

For response category, we noted a significant main effect for stimulus gender [$F(1,78) = 11.17$, $p = .001$, partial $\eta^2 = .13$], with female stimuli categorised more often as trustworthy (46.9%) than male stimuli (40.1%). No other effects were significant (*p*s > .22).

Measure	Lab stills		Lab clips		ANOVA results
	Female	Male	Female	Male	
estimated reciprocation rate	.41 (.14)	.42 (.16)	.47 (.18)	.42 (.17)	$F_{type} = 1.09$, $p = .299$, partial $\eta^2 = .01$ $F_{gender} = 3.11$, $p = .082$, partial $\eta^2 = .04$ **$F_{gender\ x\ type} = 9.99$, $p = .002$, partial $\eta^2 = .11$**
correlation between estimated & actual return rate	**.26 (.39)** ***< .001***	-.08 (.33) *.155*	**-.13 (.30)** ***.007***	.06 (.31) *.218*	**$F_{type} = 5.17$, $p = .026$, partial $\eta^2 = .06$** $F_{gender} = 2.17$, $p = .145$, partial $\eta^2 = .03$ **$F_{gender\ x\ type} = 30.39$, $p < .001$, partial $\eta^2 = .28$**
categorisation accuracy	**.59 (.14)** ***< .001***	.47 (.10) *.094*	.51 (.08) *.570*	.53 (.12) *.162*	$F_{type} = 0.49$, $p = .486$, partial $\eta^2 = .01$ **$F_{gender} = 8.82$, $p = .004$, partial $\eta^2 = .10$** **$F_{gender\ x\ type} = 17.54$, $p < .001$, partial $\eta^2 = .18$**
response category (i.e. percentage "trustworthy" responses)	.43 (.28) *.118*	**.39 (.30)** ***.024***	.51 (.30) *.876*	.42 (.32) *.096*	$F_{type} = 0.68$, $p = .413$, partial $\eta^2 = .01$ **$F_{gender} = 11.17$, $p = .001$, partial $\eta^2 = .13$** $F_{gender\ x\ type} = 1.53$, $p = .220$, partial $\eta^2 = .02$

Table 4.6. Responses to lab stimuli per type (stills vs. clips) and sex (female vs. male). Standard deviations are presented in parentheses. P values below each statistic reflect differences to chance levels. Significant results are in bold.

In the end, all ANOVAs reported above were re-run with two additional between-subjects factors: participant gender and stimuli set order (whether the Lab stimuli were seen before or after the TV stimuli). The results did not modify in any notable way: we found no significant main effects or interaction effects involving any of these factors (all ps > .05).

Comparison and correlation between TV and Lab stimuli judgments

In our experiment, participants provided judgments for both TV and Lab stimuli, only the stimulus type differing between the two blocks. Thus, participants either judged TV stills and Lab clips, or TV clips and Lab stills. As we have pointed out above, the results suggest that participants were better than chance at judging (male) clips from the TV stimuli, and female stills from the Lab stimuli. Considering that the same participants who judged TV clips also judged Lab stills, one possibility is that our counterbalanced allocation of participants to conditions produced, by chance, a higher proportion of participants able to distinguish trustworthy from untrustworthy targets among those that saw TV clips and Lab stills. If that were the case, one would expect to find a correlation between performance with TV clips and performance with Lab stills.

The correlations between participants' performance with one type of stimuli versus the other are summarised in Table 4.7. The only significant correlations were between confidence levels, which means that participants confident in their judgments for stills were similarly confident in their judgments for clips. Participants' judgments for TV clips did not correlate with their judgments for Lab stills, in terms of either response bias or accuracy, suggesting that the participants who performed better with TV clips were not also better with Lab stills. The situation was similar for participants that judged TV stills and Lab clips.

Measure	TV stills & Lab clips	TV clips & Lab stills
response category	0.048 *.769*	0.065 *.690*
categorisation accuracy	0.296 *.063*	0.084 *.605*
confidence	**0.841** ***< .001***	**0.806** ***< .001***

Table 4.7. Correlations between performance in the two experimental blocks (TV and Lab stimuli, different types). p values are presented in italics below each correlation. Significant correlations are in bold.

Although our data did not come from a pure between subjects design (each participant completed two out of four conditions, see above), we ran a 2×2 between-subjects ANOVA with stimulus source (TV/Lab) and type (clips/stills) as independent variables and accuracy as dependent variable. Confirming the noted anomaly between higher performance with dynamic stimuli in the TV condition and higher performance with static stimuli in the Lab condition, we found a significant interaction effect between stimulus source and type: $F(1, 156) = 5.83$, $p = .017$, partial $\eta^2 = .04$. Main effects were not significant (*p*s > .11).

Discussion

When asked to categorise targets from a TV show as trustworthy (i.e., keeping their promise) or untrustworthy (i.e., breaking their promise), participants performed better than chance with dynamic stimuli. When static images of the same targets were presented, performance was at chance. The difference between performance with clips and stills was more evident for male faces. This result suggests that, for the TV targets, cues to trustworthiness are extracted from facial motion and not from facial structure.

For the Lab stimuli the situation was reversed. When looking at female still images, participants could distinguish targets who returned half or more of the amount received (i.e., showing trustworthy behaviour) from those who returned nothing or less than half of the money available (i.e., showing untrustworthy behaviour). With male stills and both male and female clips, participants were at chance. The increased accuracy with TV clips and Lab (female) stills was not due to chance allocation of better 'face-readers' to the group seeing these stimuli, compared to the group seeing TV stills and Lab clips. Performance with the two stimulus types did not correlate for any of the groups.

How can we explain that it was possible to distinguish between trustworthy and untrustworthy targets in clips from the TV show but not from the Lab trust games? The difference might be due to how these stimuli were produced. First, the TV clips included the facial sequence when the potential deceptive behaviour occurred, while the Lab clips were taken from before the trust games. The TV clips could thus feature additional information about 'state trustworthiness', while the Lab clips were limited at presenting information about 'general trustworthiness'. In this respect, the Lab clips are not more informative than the Lab stills and might even have a detrimental effect on detecting deception (as is the case for the female stills) by increasing noise. Second, in one case we have stimuli linked to (un)trustworthy behaviour exhibited in the real world, while in the other case the stimuli are related to artificial behaviour in the lab. It is conceivable that people behave differently in these contexts.

Up to now, our attempts to understand if and how trustworthy behaviour in the real world can be predicted from faces (static or dynamic) have produced mixed results. In Experiment 1 we found that trustworthiness can be inferred from static images, but the result did not replicate in Experiment 2. Instead, in Experiment 2 dynamic images seemed to be more informative. Therefore, we conducted a final experiment to inform us about the role of stimulus type (static and dynamic image) and gender (female and male targets) on the accuracy of trustworthiness impressions in real-world situations.

EXPERIMENT 3

Methods

Participants

We recruited 120 online participants (68 female) through Amazon

Mechanical Turk (www.mturk.com). Results from web-based samples of participants have been shown to have comparable means, standard deviations and internal reliability as those from lab-based samples (Germine et al., 2012). Participants were U.S.-based and had a task completion rate of at least 95%, an indicator of good previous performance. Age range was between 18 and 63 years old, with a mean age of 34.0 years old (SD = 10.7). Four participants (three female) were excluded because they did not vary their responses in one or more of the four conditions (see procedure). The experiment took seven minutes to complete and participants were paid $0.60.

Stimuli

From the 38 TV targets (15 female) presented in Study 1, we randomly selected 32 targets (12 female) with the constraints imposed by the desired experimental design (i.e., number of trustworthy female, untrustworthy female, trustworthy male and untrustworthy male should be multiple of four). We had 8 trustworthy females, 4 untrustworthy females, 8 trustworthy males and 12 untrustworthy males. For each target, we had one still photo and one clip, but participants saw one or the other (see below).

Procedure

The experimental design was a 2×2 within-subjects design, with target trustworthiness (trustworthy versus untrustworthy) and stimulus type (stills versus clips) as independent variables. The trials were blocked based on stimuli type and gender, producing four blocks: female stills, male stills, female clips, male clips. The blocks were clearly delimited, with a starting screen announcing the next block (e.g., “Female stills”). The order of blocks was counterbalanced across participants, such that an equal number of participants started with clips or stills (stimuli type counterbalancing) and female or male stimuli (target gender counterbalancing). Each participant saw all characters once, in either a still or a clip (randomly determined). To ensure all characters were seen an equal number of times in stills and in clips, we created yoked pairs; the characters that appeared in stills to one participant appeared in clips to the next participants, and vice versa (keeping the same order of the blocks, but randomising the trials within one block). There were 32 trials in total (8 per condition). Notably, the ratio of trustworthy to untrustworthy characters was different between male and female, but equal between stills and clips if gender was ignored.

Before starting the actual experiment, participants were asked whether they believed they could determine a person's personality from appearance (binary answer) and whether they thought they could spot deceptive individuals when they saw them (binary answer). Following each question, we also collected their confidence levels in the answers given (continuous variable between 0 and 100). Then, after a presentation of the experiment and the task, participants were asked to estimate how well they thought they would do in each block (four estimates between 0 and 100%).

Results

Almost half of our participants thought that faces reveal information about personality or behaviour: 46% said they can infer personality from face (average confidence 50%, SD = 24%) and 44% said they can detect untrustworthy individuals by their face (42% confidence, SD = 25%). A 2×2 repeated-measures ANOVA revealed a main effect of target type on the accuracy estimates [$F(1,115) = 50.69$, $p < .001$, partial $\eta^2 = .31$], with higher estimates for clips (57%) compared to estimates for stills (50%). The main effect of target gender [$F(1,115) = 0.07$] and the interaction effect [$F(1,115) = 0.77$] were not significant.

Because ratio of trustworthy to untrustworthy targets was not equal across the four conditions, we analysed participants' performance using nonparametric measures from signal detection theory. We computed A', a bias-free measure of sensitivity (i.e., the ability to discriminate cooperators from cheaters), and B" or response bias (i.e., the tendency to select one option more often than the other). We coded trustworthy characters as targets, so each trustworthy character correctly guessed was a Hit and each untrustworthy character incorrectly guessed was a False Alarm. A' usually ranges between 0.5 corresponding to zero sensitivity and 1 corresponding to perfect sensitivity to detect the target. B" ranges from -1 to 1 and the absolute distance to 0 reflects the response bias to one option or the other (0 is no response bias).

In terms of sensitivity (i.e., performance), a 2×2 repeated-measures ANOVA with target gender (female/male) and type (still/clip) as independent variables and A' as the dependent variable revealed no significant effects (all *p*s > .12). Adding participants' gender as a between-subjects factor to the above ANOVA produced negligible changes, with neither the main effect nor any interaction effects of gender reaching significance (all *p*s > .10). The average A' for stills was 0.58, while the average A' for clips was 0.56, and they were both significantly higher than 0.50 [$t(115) = 4.40$, $p < .001$; and $t(115) = 3.20$, $p = .002$, respectively].

In terms of response bias, a 2×2 repeated-measures ANOVA with target gender (female/male) and type (still/clip) as independent variables and B" as the dependent variable revealed a significant main effect of gender [$F(1,115) = 25.38$, $p < .001$, partial $\eta^2 = .18$], with women perceived to be more trustworthy. The other effects were not significant (*p*s > .29). The average B" for stills and clips was .49 in both cases, showing a tendency to select more "trustworthy" responses (both *p*s < . 001).

Discussion

Our results indicate that trustworthy behaviour can be inferred from faces. In contrast with the results from Experiment 2, we found that judgments about targets' trustworthiness from stills and clips were comparable in terms of accuracy. This suggests that diagnostic cues to trustworthiness may be also linked to facial structure, in which case facial motion, even at the moment when targets are considering deception, adds limited information.

Consistent with the results from the first two experiments and previous studies (DePaulo, Stone, & Lassiter, 1985; Zuckerman, DePaulo, & Rosenthal, 1981), we noted a response bias in favour of selecting the 'trustworthy' option.

GENERAL DISCUSSION

The present studies showed that people were able to predict trustworthy behaviour from faces (see summary in Table 4.8). This capacity was used to discriminate at higher than chance levels between trustworthy and untrustworthy individuals in laboratory trust games and in real-world, high-stake televised prisoner's dilemma games. The implicit trustworthiness judgments were based on static and dynamic face stimuli. Discrimination based on static images was successful for the laboratory behaviour and in two out of three instances for the real-world behaviour, especially with female faces. Discrimination based on dynamic images was better than chance on both occasions when TV stimuli were presented, but failed for Lab stimuli. In this case, it appeared dynamic images degraded performance compared to static images. Target sex did not make a difference.

Stimuli		Sensitivity (A')						Response bias (B")		Correlation confidence & response accuracy	
		Stills			Clips			Stills	Clips	Stills	Clips
		Female	Male	ALL	Female	Male	ALL				
Exp1	TV	**<u>.55</u>**	.53	**<u>.54</u>**			-	**<u>.20</u>**	-	.03	-
Exp2	TV	.52	.48	.51	**<u>.56</u>**	**<u>.58</u>**	**<u>.57</u>**	**<u>.12</u>**	**<u>.11</u>**	.01	.01
Exp2	Lab	**<u>.58</u>**	.51	**<u>.57</u>**	.52	.53	.52	**<u>-.48</u>**	**<u>-.46</u>**	.00	.00
Exp3	TV	**<u>.60</u>**	**<u>.54</u>**	**<u>.58</u>**	.54	.54	**<u>.56</u>**	**<u>.49</u>**	**<u>.49</u>**	-	-

Table 4.8. Summary of sensitivity (A'), response bias (B") and correlations between confidence and response accuracy in three experiments on detection of trustworthy behaviour. Significant results are in bold and underlined. Negative B" values represent response biases towards selecting the 'untrustworthy' response, while positive B" values represent biases towards the 'trustworthy' response.

Comparing performance with dynamic versus static images produced mixed results. The difference was in favour of clips (Experiment 2, TV stimuli), stills (Experiment 2, Lab stimuli) or none (Experiment 3, TV stimuli). The results are inconsistent with the "leakage" hypothesis (Ekman & Friesen, 1969; Zuckerman et al., 1981), according to which deceivers “leak” nonverbal cues to deception that can be observed by their counterparts. Examples of such visual cues are micro-expressions, less eye contact, increased blinking and increased fidgeting.

A common finding in all three studies is that people overestimate their ability to ‘read’ faces. This is consistent with the general overconfidence bias described in the judgment and decision-making literature (Harvey, 1997; Lichtenstein & Fischhoff, 1977), according to which individuals systematically overestimate the accuracy of their responses to cognitive or perceptual problems. However, higher confidence did not equal better performance. In fact, we noted almost null correlations between confidence level and response accuracy, indicating that people had little insight on the correctness of their answers. These results are consistent with the findings from a meta-analysis examining the correlation between judgment accuracy and confidence in studies of deception detection (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997).

Detection of (un)trustworthiness in laboratory and real-world games is one aspect of a more general capacity to detect cheaters in social contexts. Researchers argued that evolution equipped us with a specialised module for cheater detection (Cosmides & Tooby, 1992), with fast and automatic mechanisms that are independent of central cognitive processing. Our evidence that (un)trustworthiness detection can occur without conscious insight supports this view and it is inconsistent with a more recent hypothesis (Sturgis, Read, & Allum, 2010) which suggested success at detecting deceptive behaviour is correlated with intelligence.

A rather puzzling finding is the difference in accuracy between trustworthiness judgments from static images (higher than chance) and dynamic images (at chance) in the case of the laboratory targets (Experiment 2). It seems reasonable to assume that, given the increased richness of dynamic stimuli, accuracy of face-based trustworthiness judgments can only improve. However, recent studies (Bonnefon et al., 2012) showed that richer stimuli interfered with accuracy of face impressions of trustworthiness; external cues such as hairstyle and clothing increased people’s confidence in their impressions but negatively affected their accuracy. In our study, facial motion of laboratory targets was recorded a few weeks before the trust games and thus it was extremely unlikely to be any more informative than facial structure with respect to the behaviour exhibited during the games. Instead, it probably added just noise and made participants' judgments less accurate. In contrast, the clips of the TV targets (discriminated at better than chance level) were recorded at the moment when promises to cooperate were made, increasing the probability that cues to deception were "leaked" into facial motion.

In line with previous studies of lie/cheater detection (Bond & DePaulo, 2006; DePaulo et al., 1985; Zuckerman et al., 1981), we document a "truth bias" in participants - a higher proportion of "trustworthy" than "untrustworthy" responses in the experiments featuring TV stimuli (stills and clips) when people were asked for binary judgments. When participants were asked to estimate the average return rate in laboratory trust games, most responses were below 50%. This tendency reflected in the negative average B'' values. A general propensity to judge most people as trustworthy or untrustworthy may produce significant results in experiments investigating accuracy of trustworthiness impressions when the base rates favour cooperators. To guard against this possibility, we computed sensitivity measures independent of response bias whenever necessary.

One aspect not elucidated by our studies is which facial cues facilitate trustworthiness detection. Stirrat and Perrett (2010) claimed that facial width-to-height ratio (WHR) is a valid signal to trustworthiness that is successfully used by observers to infer targets' behaviour. The appeal of WHR as a cue to behaviour lies on its biological plausibility; higher facial WHRs were linked to increased testosterone levels in men (Verdonck et al., 1999), and testosterone was linked to more aggressive (Carre & McCormick, 2008), thus potentially uncooperative behaviour. However, other aspects (e.g., brown eyes, Kleisner et al., 2013) may also play a role.

Future studies should explore the factors that influence trustworthiness detection. Although it is unlikely that faces can provide sufficient information ever to allow a reliable forecast of trustworthy or cooperative behaviour (if that was the case, cheaters would have been eliminated by evolution long time ago), perhaps trustworthiness detection can be improved beyond the slightly above chance level.

In a series of laboratory studies, Bonnefon et al. (2012) showed that facial trustworthiness judgments are more accurate when faces are converted to black-and-white and cropped to exclude external cues (including hair). It would be interesting to see whether this ‘less is more’ effect replicates with real-world stimuli and behaviour. Bonnefon et al. (2012) also found that trustworthiness detection was successful only for consequential decisions and not for explicit evaluations. A straightforward way to test these hypotheses on real-world data would be to run an experiment in which participants are assigned to the role of investors in trust games (or partners in PD games), with trustees (or partners) represented by black-and-white cropped faces of the TV contestants.

Another factor to consider is that trust may be contextual. For example, in a prisoner’s dilemma game each player is likely to adjust his/her strategy according to the forecasted action of the other player. Especially if a player is perceived as a “grabber” (i.e. selecting to not share in the game), the player might want to punish this behaviour by selecting to not share as well. In the TV show, the two contestants in the prisoner’s dilemma game could communicate and send true or false signals. The contestants might have adjusted their strategy according to how these signals were decoded or interpreted. Therefore, their decision might have to do not only with their actual trustworthiness, but also with their partner’s perceived trustworthiness. In our experiment, we presented the faces of the players one at a time, with no information about their partners. Perhaps a presentation of both partners would increase the accuracy of trustworthiness judgments.

To conclude, we showed that people are slightly better than chance at recognising (un)trustworthy behaviour in laboratory and in real high-stake economic games. Contrary to what might be expected and in contrast with the predictions of the “leakage” hypothesis (Ekman & Friesen, 1969; Zuckerman et al., 1981), in most cases dynamic stimuli did not improve the ability to detect deception. Importantly, we found null correlations between accuracy and confidence, suggesting that relying on face-based judgments to infer trustworthiness on a case-by-case basis is likely to be unsuccessful.

4.3. FACE TRUSTWORTHINESS AND REPUTATION†

Previous studies have demonstrated a causal role of facial cues in economic interactions, with participants investing more in partners with trustworthy-looking faces (Rezlescu, Duchaine et al., 2012; Stirrat & Perrett, 2010; van't Wout & Sanfey, 2008). However, these studies offered participants no information about their partners beyond their faces, a situation rarely encountered in real life (Olivola & Todorov, 2010b). People usually go beyond appearances and learn more about prospective partners before engaging in any social or economic cooperation, and face judgments are known to be quickly updated in line with background information (Todorov & Olson, 2008).

Thus, an important question for confirming the ecological validity of the effects of trustworthy-looking faces is whether they survive in richer informational environments. The current study aims to explore the interaction between face impressions and reputational information in trust games (Berg et al., 1995). To convey reputational information, we created relatively unambiguous behavioural trustee histories, designed to suggest high or low reciprocity in previous rounds of trust games. Rationally, people should focus on the trustees' past behaviour and ignore facial cues.

Methods

Participants

Fifty-two participants (30 female, age range: 18-62 years, Median = 23 years) participated for payment (£4 show-up fee plus variable bonus, see Procedure below).

Face stimuli

We used the 40 computer-generated faces presented in Study 1 from Rezlescu, Duchaine et al. (2012). The faces were originally created by Oosterhof

† Pilot data from this study were previously presented in the thesis submitted in 2009 to University College London for the MSc. degree in Cognitive and Decision Sciences.

and Todorov (2008), who developed a computer model that can manipulate faces to make them less or more trustworthy-looking. Twenty Caucasian faces ('characters') with neutral expressions were generated randomly using FaceGen Modeller 3.3 software (Singular Inversions, 2008). For each character, the model produced two different facial 'identities' at opposing ends of the trustworthiness scale (at -3 and +3 SD away from the original face on the model's trustworthiness dimension; see Figure 4.3). The distance on the trustworthiness scale between the identities was sufficiently large that participants would be unlikely to realise the two identities were derived from the same face, but not so extreme that faces lost their neutral expression (Todorov et al., 2008) or looked unrealistic. In the end we had 20 untrustworthy and 20 trustworthy faces.

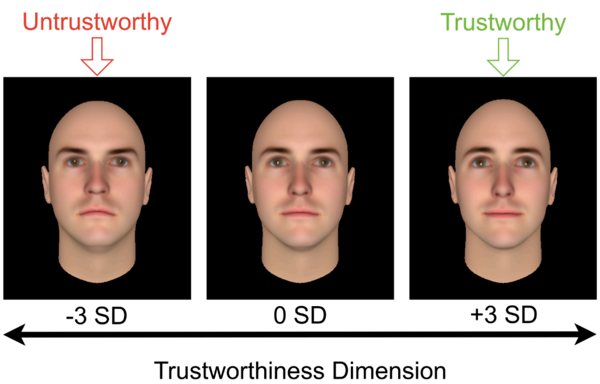


Figure 4.3. Examples of face stimuli. Face identities of the same computer character varied on the trustworthiness scale. For each character, we selected the faces found at -3 SD and +3 SD on the trustworthiness scale (indicated here with arrows).

Behavioural history stimuli

In addition to face identity, participants saw each trustee's behavioural history in the trust game. Behavioural histories were presented as 3 × 3 grids of blue-coloured cells varying in shading (Figure 4.4). Participants were told that these cells represented nine randomly selected return rates in past rounds from the corresponding trustee. Lighter shades of blue corresponded to low return rates and darker shades to high return rates. We used colour rather than numbers to avoid explicit arithmetical operations and simple cutoff-rule investment strategies. Our intention was to provide summary representations of partners' behavioural histories. We also aimed to discourage the belief that these histories were perfect predictors of future return rates, hence the random selection and variable nature of trustee past behaviour.

The behavioural history variable had two levels: Good and Bad, corresponding to predominantly high or low past return rates, respectively. History stimuli were selected by asking ten volunteers to rate 50 Bad and 50 Good behavioural histories, quasi-randomly generated (i.e. following a series of parameters designed to ensure that Bad histories contained predominantly low return rates, while Good histories contained mostly high return rates), on a scale from 1 (not at all trustworthy) to 7 (very trustworthy). Fifteen histories with average ratings between 2.5 and 3 and the lowest SDs (between 0.47 and 0.95) were picked for the Bad condition; 15 histories with average ratings between 5 and 5.5 and the lowest SDs (between 0.67 and 0.97) were picked for the Good condition. Hence we selected 30 histories that were consistently perceived as either "bad" or "good", without appearing extreme. These histories were then rotated clockwise and anticlockwise to produce more variations for a total of 70 stimuli (half "bad").

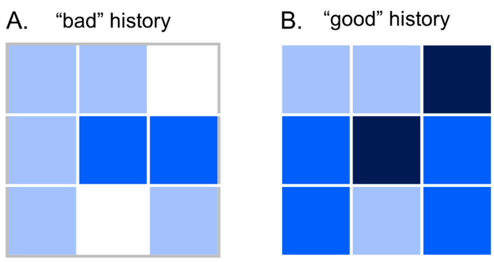


Figure 4.4. Examples of trustee behavioural histories shown to participants (i.e. investors in trust games). (A) Behavioural history from a "Bad" history trial with two very low (white boxes), five medium-low (light blue), two medium-high (intense blue) and no very high (dark blue) return rates. (B) Behavioural history from a "Good" history trial.

Experimental design

We had two independent variables of interest: history (Bad vs. Good) and face identity (Trustworthy vs. Untrustworthy). The total number of trials was 70: 10 trials for each of the 2 × 2 conditions in which faces and histories were presented simultaneously to the participants, plus 30 trials for a control condition where each behavioural history was presented alone ("No-face" condition). Faces and histories were randomly paired according to each condition (e.g. an Untrustworthy face could be presented with any Bad behaviour). Trials were randomised with the constraint that two (trustworthy and untrustworthy) face versions of the same character could not be presented directly one after the other. The dependent variable was the amount invested by participants.

Procedure

Participants were ostensibly engaged in a series of online trust games (Berg et al., 1995) and all were assigned to the role of investor. On each round, they received 100 virtual pounds (VP) and could invest any part of this amount in a trustee whose computerised face and behavioural history (the stimuli described above) appeared simultaneously on the screen. Participants were informed that they would not see their partners' faces on some trials (i.e. the "no face" condition). The amount invested tripled before reaching the trustee. Participants were (falsely) told the trustees were real players from other universities who could decide, without any obligation, to return part of the tripled amount to the investors. Furthermore, participants were (correctly) informed that they would be paid based on their accumulated earnings across 70 rounds of the game (according to an exchange rate of £1 per 1000 VP) so they had an incentive to invest in trustees who would return more than their initial investment. Thus, the amount invested in each partner measured the perceived trustworthiness associated with the corresponding face identity. We stressed the anonymity of the game and that interactions were non-repeating (i.e., only one interaction with each trustee). There was no time limit for decisions, nor feedback provided after each round; the amounts 'returned' by trustees were concealed to avoid subsequent decisions being affected by earlier outcomes. As in experiment 1 from Rezlescu, Duchaine et al. (2012), we took a number of measures to ensure participants believed they were interacting with real trustees. First, we insisted participants arrive on time for the experiment so that they could start at the (allegedly) agreed-upon time with their partners in the game. If they arrived more than five minutes late (or failed to show up), we rescheduled the experiment at a later date. Second, before starting the experiment, participants were photographed wearing a neutral expression and their photo was uploaded into FaceGen to create a "computerised" version of their face. These computerised faces were similar to the face stimuli used in our study: they preserved the facial structure of each participant, yet had no hair or specific face identifiers, and had perfect skin texture. After showing participants their own computerised FaceGen photo, we pretended to upload it for the trustees to see during the game. Thus, participants had a good reason to believe that the FaceGen trustee faces they saw during the experiment were computerised representations of real people's faces whose photos were similarly taken, transformed, and uploaded for the study. Third, between the practice trials and actual games, we intentionally added a delay of several minutes – a fake "waiting time" for other players to (allegedly) join the game – during which the experimenter complained about the difficulties of running such a large scale study. Finally, we added random-length 10-20 second delays between participants' investment decisions and the confirmations they received from trustees (that the latter players' decisions had also been made). This was done to strengthen participants' impressions that they were interacting with real, deliberating human players. To avoid contaminating the subject pool, participants were fully debriefed by email only after all testing had been concluded.

Results

In trust games in which reputational information was presented next to the trustees' faces, investments were influenced by both histories and face identities (Figure 4.5). A 2 × 2 repeated-measures ANOVA, with history (Bad and Good) and identity (Untrustworthy and Trustworthy) as independent variables, revealed significant main effects of behavioural history: $F(1, 51) = 214.48$, $p < .001$, partial $\eta^2 = .81$; and face identity: $F(1, 51) = 5.94$, $p = .018$, partial $\eta^2 = .10$, but no interaction effect: $F(1, 51) = 2.31$, $p = .135$. The average amount invested in Trustworthy identities (of both "Good" and "Bad" trustees) was 6% higher than the average amount invested in Untrustworthy identities (45.2 versus 42.4 VP).

Figure 4.5. Average amounts invested in Untrustworthy and Trustworthy face identities with Good and Bad behavioural histories. Note that behavioral histories are represented on two different scales: the blue scale corresponds to Bad history trials; the green scale corresponds to Good history trials. Main effects of both behavioral history and face identity were significant. Error bars represent standard error.

"Good" histories attracted an average of 67.39 VP, while "Bad" histories attracted an average of 20.65 VP, further confirming that participants considered the coloured history matrices to be informative of their partners' tendency to reciprocate. The study also included trials in which participants were only shown the behavioural history of their partner (i.e., without the face presented). This allowed us to compute the facial trustworthiness bonus for each character, which was the difference between the amount invested in the character's trustworthy identity coupled with a behavioural history and the amount invested in the same behavioural history alone, without any face (i.e., facial trustworthiness bonus = trustworthy face – no face). We also computed the facial untrustworthiness penalty for each character, which was the difference between the amount invested in the character's untrustworthy identity coupled with a behavioural history and the amount invested in the same behavioural history alone (i.e., facial untrustworthiness penalty = untrustworthy face – no face). The mean facial trustworthiness bonus, averaged across participants and characters, was 1.57 VP. The mean facial untrustworthiness penalty was -1.83 VP. A within-participant t-test confirmed that the difference between the trustworthiness bonus and the untrustworthiness penalty (Δ = 3.40 VP) was significant: $t(51) = 2.94$, $p = .005$, $\eta^2 = .15$. However, a within-participant t-test comparing the absolute values of the trustworthiness bonus and the untrustworthiness penalty failed to reveal a significant difference between the two: $t(51) = .16$. Thus facial trustworthiness appears to shift investments symmetrically upwards or downwards (relative to no face) according to its valence.

Discussion

Building on one of our previous studies (Rezlescu, Duchaine et al., 2012; experiment 1) showing that facial configurations perceived to indicate trustworthiness have an impact on economic decisions, here we showed that trustworthy-looking facial features influenced investors' actions even when information about trustees' past behaviour was available. However, the facial trustworthiness premium was much reduced, from 42% when participants saw only the faces of the trustees (Rezlescu, Duchaine et al., 2012, experiment 1) to 6% when reliable 'good' or 'bad' information about trustees was provided.

The lack of an interaction between face trustworthiness and background information suggests that the trustworthiness effect is independent of behavioural history type, so that trustees with "Good" and "Bad" histories benefited equally from trustworthy-looking facial features. Furthermore, we showed that trustworthy and untrustworthy identities contributed equally to the facial trustworthiness effect (but in opposite directions).

The fact that trustworthy-looking faces are still favoured when accompanied by objective cues about trustworthiness is consistent with recent studies which found that perceptions of borrower trustworthiness influenced lending decisions and received interest rates in online peer-to-peer lending where lenders have relatively rich background information about borrowers (Duarte, Siegel, & Young, 2010; Ravina, 2008). In this study though, borrower photos often presented a full body, clothes etc. and sometimes included other people as well (e.g. family members, small children), so it is difficult to assert which aspects associated with trustworthiness influenced lending choices.

Another study (Chang et al., 2010) found that, in laboratory trust games involving multiple interactions with the same trustee, participants invested increasingly more in partners who proved to be trustworthy. In contrast with our findings, the main effect of facial trustworthiness was not significant, but partners who showed trustworthy behaviour (i.e. higher reciprocation rates) still received more money if they had trustworthy rather than untrustworthy looks. The different results obtained by Chang et al. (2010) and us are probably due to how reputational information was acquired. In our experiment, rather than gradually discovering trustee reputations from first-hand interactions, participants saw visual summaries of their partners' past reciprocations (just as one might receive third-party reports about potential business partners). Thus, our participants had simultaneous access to faces and reputational information, so they could integrate both immediately.

The face trustworthiness effect was arguably small, but surprising considering how clear-cut the historical information about the investment partners was. A question to be answered by future studies is how an increased degree of ambiguity related to the partners' past behaviour would alter decision-making and whether we can establish an inverse correlation between the richness and reliability of background information available and the face trustworthiness premium.

4.4. VOICE IMPRESSIONS

As social stimuli, voices are in many respects similar to faces and thus are likely to play a significant role in forming trait impressions. Spectral analysis of U.S. presidential candidates' voices predicted eight out of eight surveyed elections (Gregory & Gallagher, 2002), and voices were shown to influence evaluation and voting of political candidates (Surawski & Ossoff, 2006; Tigue et al., 2012). Men with lower-pitch voices are overwhelmingly perceived to be more dominant (Jones et al., 2010; Puts et al., 2006; Puts et al., 2007) and more attractive (Feinberg et al., 2008; Feinberg et al., 2005; Jones et al., 2010), irrespective of the judge's gender. Vocal characteristics or voice impressions seem to be unrelated to body characteristics (Collins, 2000), but male voice pitch is correlated with testosterone level (Dabbs & Mallinger, 1999) and thus may be a valid cue to dominance.

Despite their importance in social interactions, trait impressions based on voices received considerably less attention than facial trait impressions. In Experiment 1 I validate a new set of voices and faces and investigate whether voices, like faces, lead to formation of reliable trait impressions of trustworthiness, attractiveness and dominance. Furthermore, I study the intramodal correlations between the three trait impressions and the intermodal correlations between voice-based and face-based impressions. Experiment 2 examines how voice impressions combine (or not) with face impressions to form an integrated person perception. The different integration patterns across various traits are explored.

EXPERIMENT 1: Voice and face impressions

Methods

Participants

We recruited online participants through Amazon Mechanical Turk (www.mturk.com). Results from web-based samples of participants have been shown to have comparable means, standard deviations and internal reliability as those from lab-based samples (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Germine et al., 2012; Horton, Rand, & Zeckhauser, 2011). Participants were U.S.-based and had a task completion rate of at least 95%, an indicator of good previous performance.

Two hundred ninety-eight Mechanical Turks (MTs) rated the voice and face stimuli in exchange for $0.60. Fifteen MTs were excluded because of extremely low variance in ratings (below 0.7), denoting lack of involvement with the task (average variance in our sample was 3.43, SD = 1.55). In the end, we analysed the ratings from 283 participants (age range 18 - 67, mean age 31.1, SD = 11.9; 141 female).

Voice and face stimuli

We obtained audiovisual material from 41 volunteers (age range 19 - 49, 19 female) recorded in a sound proof chamber. Volunteers were asked to face the camera and pronounce five English vowels in various expressions (neutral, happy, angry, surprised, disgusted, fearful and sad, in this order). They were instructed to be as expressive as possible in both voice and face and produce sounds lasting approximately one second. We opted for vowels (as opposed to words or sentences) to avoid potential confounding factors related to accents or verbal information. Each vowel-expression was repeated once.

For the current study we selected as voice stimuli the first neutral expressions of vowels A, E and O (we discarded I and U because of the potential verbal loadings). Stimuli were normalised to a standard intensity level using Audacity (http://audacity.sourceforge.net). The face stimuli were full grayscale head shots in neutral expression. The volunteers were grouped into 'white British' (WB) (n = 26; 9 female) and 'others' (Other) (n = 15; 10 female) based on their reported ethnicity.

Procedure

In this between-subjects study, participants were randomly assigned to one of nine (3 x 3) conditions: they were asked to provide trustworthiness (Tw), attractiveness (Att) or dominance (Dom) judgments for the 41 neutral vocalisations of vowels A, E or O. Additionally, they were asked to rate the 41 neutral faces on the same trait assigned for voice judgments. Each condition had at least 29 participants. Face stimuli did not differ according to vowels (i.e. the same faces were presented regardless of the vowel condition), thus the number of ratings per face per trait was approximately triple the number of ratings per vowel per trait. The ratings ranged from 1 (not at all trustworthy/attractive/dominant) to 9 (extremely trustworthy/attractive/dominant). The trials were blocked as follows: WB voice stimuli, WB face stimuli, Other voice stimuli, Other faces. The order within each block was randomised.

Results and discussion

Random allocation of participants to condition groups aimed to ensure the groups were comparable in terms of participant characteristics. Chi-square values for 3-by-3 contingency tables (trait-by-vowel) confirmed the groups did not differ for total number of participants ($p = .900$), number of female participants ($p = .622$), or number of white participants ($p = .460$). A two-way ANOVA confirmed there was no difference in age between groups according to trait ($p = .921$) or vowel ($p = .695$), just a marginally significant interaction effect ($p = .046$, partial $\eta^2 = .04$).

Validation of voice and face stimuli

We first analysed whether target stimuli were rated differently based on their ethnicity (WB or Other). Average ratings of trustworthiness and attractiveness were lower for WB voices than for Other voices (trustworthiness: 5.14 vs. 5.27, $t(100) = 2.36$, $p = .020$; attractiveness: 4.62 vs. 4.85, $t(88) = 2.90$, $p = .005$). WB faces were also perceived as more dominant than Other faces (4.92 vs. 4.73, $t(92) = 2.68$, $p = .009$). The other ratings (dominance for voices, trustworthiness and attractiveness for faces) did not differ between WB and Other targets (all ps > .05).

One of the most common findings in previous studies of face impressions is the high reliability of these judgments. For example, Oosterhof and Todorov (2008) reported reliabilities (Cronbach's α) above .90 and interrater agreements (r) between .26 and .47 for 13 traits. While Cronbach's α is primarily used in psychometrics to indicate the degree to which a questionnaire's items measure the same construct (i.e. how similar are the item scores belonging to the same participant), by substituting respondents with target stimuli and item scores with raters' judgments, it becomes a measure of the similarities between different raters' impressions of the same stimuli. The interrater agreement was calculated as an average of all pairwise correlations between raters' judgments. As can be seen in Table 4.9, we replicate the high reliabilities and interrater agreements for face judgments obtained by Oosterhof and Todorov (2008). Importantly, especially for WB targets, voice judgments appear to also have high reliabilities (above .88) and average interrater correlations comparable to faces (between .22 and .33). Because of their lower reliability, we excluded the Other voice stimuli from the subsequent analyses.

We next examined whether trait judgments across different vowels correlate, in other words, looking at whether individual voices receive similar ratings regardless of the vowel they are asked to produce. If trait perceptions are linked to voice attributes, one can expect high correlations here. Indeed, correlations ranged between .47 and .76 and were significant across all categories (Table 4.10). Therefore in the following sections, for simplification, we collapsed ratings of different vowels into one general rating per voice (and per trait) and report results based on these average voice ratings.

Trait judgment	Stimuli	Sample size (n)	white British targets (26)		other targets (15)	
			Interrater agreement (r)	Reliability (α)	Interrater agreement (r)	Reliability (α)
Trustworthiness	vowel A	30	**0.29**	**0.93**	0.09	0.76
	vowel E	31	**0.23**	**0.90**	0.08	0.71
	vowel O	40	**0.25**	**0.93**	0.08	0.79
	faces	101	0.28	0.97	0.21	0.97
Attractiveness	vowel A	30	**0.24**	**0.91**	0.10	0.77
	vowel E	30	**0.30**	**0.93**	0.10	0.79
	vowel O	29	**0.33**	**0.93**	0.15	0.83
	faces	89	0.33	0.98	0.42	0.98
Dominance	vowel A	29	**0.22**	**0.88**	0.36	0.93
	vowel E	31	**0.22**	**0.90**	0.26	0.92
	vowel O	33	**0.22**	**0.90**	0.19	0.88
	faces	93	0.37	0.98	0.35	0.98

Table 4.9. Interrater agreements (r) and reliabilities (Cronbach's α) for voice and face stimuli. Note that Cronbach's α tends to increase with sample size (number of raters); therefore the slightly higher reliabilities for face stimuli may be simply due to larger sample size. White British targets were selected for subsequent analyses.

Vowels	Trustworthiness (TW)	Attractiveness (ATT)	Dominance (DOM)
A - E	**.625 *(.001)***	**.736 *(.000)***	**.757 *(.000)***
A - O	**.623 *(.001)***	**.688 *(.000)***	**.669 *(.000)***
O - E	**.627 *(.001)***	**.763 *(.000)***	**.471 *(.015)***
mean correlation	***.625***	***.729***	***.632***

Table 4.10. High correlations between trait judgments of same voices under different vocalisations. All correlations were significant (in bold). p values are given in parentheses.

Intramodal correlations

Among the various trait judgments, face trustworthiness and face attractiveness are usually found to correlate the most (e.g. Oosterhof & Todorov, 2008). This was also the case for the voice impressions in our study: Tw and Att judgments of both male and female voices correlated highly (.87 and .83, respectively; see Table 4.11). In contrast, neither Tw - Dom nor Dom - Att correlations were significant.

In the case of faces, Tw and Att correlated when all stimuli were considered (see Table 4.11). However, a closer look revealed that the significant correlation held only for male faces, while for female faces the correlation was close to zero. Attractive female faces were not necessarily perceived as trustworthy. The differences in perception of female and male faces underline the importance of dividing all analyses of trait perceptions by target gender. For male faces, we also found a significant negative correlation between Tw and Do judgments, replicating the findings of Oosterhof & Todorov (2008), where face Tw was negatively correlated with dominance. None of the other correlations reached significance. Notably, the results did not suffer major changes when analysed separately based on participants' gender.

Trait judgments	VOICES			FACES		
	All (26)	Female (9)	Male (17)	All (26)	Female (9)	Male (17)
Tw - At	**.890 *(.000)***	**.832 *(.005)***	**.865 *(.000)***	**.512 *(.008)***	.014 *(.971)*	**.508 *(.037)***
Tw - Do	-.020 *(.923)*	.149 *(.702)*	.344 *(.176)*	**-.630 *(.001)***	-.410 *(.273)*	**-.595 *(.012)***
Do - At	-.334 *(.095)*	-.270 *(.482)*	.032 *(.904)*	-.282 *(.163)*	.349 *(.358)*	-.211 *(.417)*

Table 4.11. Intramodal correlations between voice and face trait judgments. There were 26 stimuli, 9 female and 17 male. Significant correlations are in bold and p values are given in parentheses.

The high correlations between Tw and Att judgments may imply that participants, in general, find it difficult to judge these two dimensions independently. However, the null correlation between Tw and Att judgments of female faces rules out this possibility. Instead, it is more likely that judging one of these dimensions is made difficult only by particular contexts and/or stimuli. In these situations, participants may rely on the "halo effect" and judge faces similarly on all positive (or negative) dimensions (e.g. Dion, Berscheid, & Walster, 1972). For example, it is conceivable that attractiveness is not such a well defined concept for male compared to female faces, and therefore people rate male faces on valence rather than attractiveness. Furthermore, attractiveness may be perceived predominantly as a visual attribute, difficult to "translate" into the auditory domain, which would generate high correlations between voice attractiveness and other positive traits.

Intermodal correlations

There are two reasons to be interested in looking at intermodal correlations of trait judgments. First, because people agree to such a large extent with respect to which faces (voices) are high on a given trait, it is reasonable to assume that there are certain visual (auditory) cues which are consistently seen as signalling that particular trait. The interesting question is whether the same individuals display both the telling visual and auditory cues. The second reason is an extension of the first. It has been claimed (e.g. Ambady, Hallahan, & Rosenthal, 1995; Berry, 1991; Carre & McCormick, 2008; Stirrat & Perrett, 2008) that certain character traits (including trustworthiness) can be accurately inferred from faces. If this were the case, one may expect some diagnostic information to "leak" through both faces and voices.

In our study, an initial comparison of face-based and voice-based ratings per target revealed significant correlations among Tw and Att (but not Dom) judgments (Table 4.12). However, when ratings were analysed separately per target gender, these correlations were no longer significant. This indicates the significant correlations when collapsing across gender were due to the previously documented bias to judge feminine targets more favourably than masculine targets (Perrett et al., 1998). Higher ratings for both female faces and voices than male faces and voices, respectively, led to this apparent correlation between face and voice ratings when gender was disregarded. In effect, voices and faces determined independent perceptions of Tw and Att. Results did not differ according to participants' gender.

Our findings are inconsistent with Collins & Missing (2003) who found that male judgments of female vocal and facial attractiveness are related. Their results might have been confounded by body size, which influenced both voice pitch and perceived attractiveness.

For dominance judgments, we note a negative correlation between face-based and voice-based perception of male targets, suggesting that dominant faces had less dominant voices. This result is puzzling. It has been suggested that male facial width ratio (Verdonck et al., 1999) and voice pitch (Dabbs & Mallinger, 1999) are both influenced by testosterone level, which in turn has been linked with dominant and aggressive behaviour (Ehrenkranz, Bliss, & Sheard, 1974). Therefore, if anything, we would have expected a positive correlation between ratings of voice and face dominance, reasoning that men with wider faces tend to have lower-pitched voices and both wide faces and low-pitched voices would be perceived as dominant. The negative correlation might be explained by a compensatory strategy, according to which men with less dominant looks compensate through their voices. The effect is likely to be specific to men because dominance has been a crucial evolutionary trait for the male population.

	White British targets		
Trait judgment	**All (26)**	**Female (9)**	**Male (17)**
Trustworthiness	**.471 *(.015)***	.096 *(.806)*	.365 *(.149)*
Attractiveness	**.502 *(.009)***	.447 *(.228)*	.086 *(.744)*
Dominance	-.126 *(.540)*	-.026 *(.947)*	**-.515 *(.034)***

Table 4.12. Correlations between voice-based and face-base judgments. Significant correlations are in bold and p values are given in parentheses.

EXPERIMENT 2: Integrated person perception

As we have seen, people tend to associate positive or negative traits with voices, similar to the process taking place with faces. Face impressions were shown to play a significant role in social interactions (e.g. Hamermesh, 2011) even when reputational information is available (Rezlescu, Duchaine et al., 2012), and it is reasonable to assume voice impressions have a similar impact. However, social interactions typically involve simultaneous exchange of visual and auditory information, so a person impression is likely to be based on both the face and the voice (and possibly other information, such as body movements). Our next aim was to examine how face impressions and voice impressions interact to form an integrated person impression.

Methods

Participants

There were three groups of online participants corresponding to the three trait judgments. The 'trustworthiness' group had 41 participants (age range 18 - 59, mean age 30.9, SD = 9.8; 17 female), the 'attractiveness' group had 41 participants (age range 19 - 66, mean age 34.0, SD = 13.4; 18 female) and the 'dominance' group had 84 participants (age range 18 - 51, mean age 28.6, SD = 8.2; 34 female). The considerably larger number of participants in the 'dominance' group was determined by a script error that doubled the requested number of participants compared with the other groups (the results did not differ when keeping only the first 41 participants). The task lasted two minutes and each participant received $0.25.

Voice and face stimuli

This experiment used a sample of the male WB stimuli (n = 17) from Study 1. The voice stimuli were of vowel A. After excluding targets above 40 years old (n = 3), with mild strabismus (n = 1) or facial hair (n = 1), we were left with 12 targets whose faces and voices could be mixed, with all face-voice combinations appearing plausible.

Procedure

First, faces were ordered based on their trustworthiness ratings from Study 1 and binned accordingly into two equal groups, one with 'untrustworthy' (UTw) and the other with 'trustworthy' (Tw) faces. We then did the same for voices, producing one group of UTw and one group of Tw voices. Note that the face and voice of a particular target could fall in different categories (e.g. same target could have a 'trustworthy' face and an 'untrustworthy' voice). Each group had six face (voice) stimuli. The average rating for the UTw faces was 3.9, for Tw faces 5.1, for UTw voices 4.4 and for Tw voices 5.3. After that, for each participant, we randomly paired half of the untrustworthy faces with untrustworthy voices and the other half of the untrustworthy faces with trustworthy voices. Additionally, half of the trustworthy faces were paired with untrustworthy voices and the other half with trustworthy voices. In the end, we had four experimental conditions, with three pairs per condition. After a catch trial ensuring that that they have the sound on, participants were presented each pair (face plus voice) in a randomised order and asked to rate the respective person for trustworthiness.

The same procedure was used to create the pairs of stimuli and collect ratings for the attractiveness and dominance judgments. The average rating per 'unattractive' (UAtt) faces was 2.9, for 'attractive' (Att) faces 4.0, for Uatt voices 3.9, for Att voices 5.0; for 'undominant' (UDom) faces 4.3, for 'dominant' (Dom) faces 5.8, for UDom voices 4.4, and for Dom voices 5.6. Allocation of participants to the trustworthiness, attractiveness or dominance condition was counterbalanced.

Results

We calculated an average rating per condition per participant and submitted these numbers to three 2x2 (face x voice) within-subjects ANOVA, one per trait. For trustworthiness judgments, both main effects of face and voice were significant: face $F(1,40) = 25.06$, $p < .001$, partial $\eta^2 = .39$; voice $F(1,40) = 30.85$, $p < .001$, partial $\eta^2 = .44$. In addition, there was a significant super-additive interaction effect between face and voice trustworthiness: $F(1,40) = 9.67$, $p = .003$, partial $\eta^2 = .20$.

The main effects of face and voice were also significant for attractiveness judgments; face $F(1,40) = 56.08$, $p < .001$, partial $\eta^2 = .58$; voice $F(1,40) = 6.37$, $p < .016$, partial $\eta^2 = .14$. There was no interaction effect between faces and voices for attractiveness ($p = .086$). We note that the main effect size of faces is four times the main effect size of voices. This difference could not be attributed to a larger difference in the average ratings of Att and UAtt faces versus the difference in the average ratings of Att and UAtt voices, because they were identical (see section above). Furthermore, to confirm that the larger effect size of faces is not driven by our artificial separation of stimuli into ‘unattractive and ‘attractive’ categories, we regressed the person ratings given by each participant on the average rating for each face and the average rating of each voice. As expected, both face and voice ratings were significant predictors ($p < .001$ and $p = .003$, respectively). More importantly, the ratio of their standardised coefficients (.318 divided by .126) confirmed the larger impact of the face on person ratings of attractiveness.

For dominance judgments, again both faces and voices had significant main effects; faces $F(1,83) = 16.39$, $p < .001$, partial $\eta^2 = .17$; voice $F(1,83) = 91.61$, $p < .001$, partial $\eta^2 = .53$. The interaction effect between faces and voices approached significance, but the effect size was small: $F(1,83) = 3.57$, $p =.062$, partial $\eta^2 = .04$. We note that in terms of effect sizes, the situation is reversed compared to attractiveness judgments; for dominance, voices have an effect size three times larger than faces. The pattern of results for the three trait judgments is summarised in Table 4.13.

	Main effects			Interaction
Trait	Face	Voice	Ratio	Face x Voice
Trustworthiness	.39	.44	1:1	.20
Attractiveness	.58	.14	4:1	-
Dominance	.17	.53	1:3	-

Table 4.13. Effect sizes (partial η^2) for faces and voices in forming a person impression. Only significant effects are shown.

Transforming participants’ ratings into z scores and reanalysing the data had no impact on the reported results. Similarly, adding participants’ gender as a between-subjects factor to the above ANOVAs produced negligible changes. In all cases, main effects of gender were not significant (all ps > .49) and there were no interaction effects with the other factors (all ps > .13) with the exception of a triple interaction effect with faces and voices for dominance judgments ($p = .032$, partial $\eta^2 = .06$) denoting that the interaction effect between faces and voices was more pronounced for male participants.

Discussion

Our results show that integration of face and voice information in forming person impressions differs according to the judged trait. Thus, while for all measured traits both faces and voices have a significant impact on person impressions, the effect sizes are of different magnitudes. Faces seem to be more important than voices when it comes to attractiveness, while the reverse pattern was observed for dominance judgments. Trustworthiness impressions were equally influenced by faces and voices. Trustworthiness judgments were also the only ones for which we found a significant interaction effect between faces and voices.

Previous studies on bimodal integration found a dominant role for visual information across multiple domains. Interestingly though, this visual superiority effect is not rigid but varies according to signal reliability. For example, Collignon et al. (2008) showed that the visual dominance in emotional processing disappeared when the reliability of the visual stimuli was diminished. In our study, the different contributions of faces and voices to attractiveness and dominance perception may be explained within this framework. On one hand, if we consider that attractiveness is predominantly a visual concept, visual information is likely considered most reliable and thus the key determinant in overall judgments of attractiveness. On the other hand, dominance judgments have been shown to correlate with more masculine aspects (Oosterhof & Todorov, 2008). Because the voice is highly sexually dimorphic (Fitch & Giedd, 1999), it may be considered more reliable when judging a person's masculinity and therefore given higher weight in overall assessments of dominance. Indeed, deep voices are reliably associated with perceived dominance in men (Jones et al., 2010; Puts et al., 2006; Puts et al., 2007) and previous studies found that men with lower-pitched voices were preferred in political contexts (Klofstad, Anderson, & Peters, 2012; Tigue et al., 2012). The voice pitch may even be a valid cue to dominance; it can signal testosterone level

(Dabbs & Mallinger, 1999), and testosterone level is directly linked to actual dominance in men (Mazur & Booth, 1998).

Our results for crossmodal integration of attractiveness perception are partly consistent with Surawski and Ossoff (2006), who found a visual superiority effect and a synergistic combination of face and voice attractiveness in evaluations of other character traits of political candidates. The higher impact of visual information was evident in our study, but we did not find a significant interaction between faces and voices in evaluations of person attractiveness. The difference is probably due to different task demands; while we measured integrated attractiveness, Surawski and Ossoff (2006) measured perceived competence, trustworthiness and leadership. As we have seen in present study, people integrate information differently according to the judged trait.

For combined trustworthiness perception, the fact that audio-visual integration did not show the visual dominance found in emotion perception (Collignon et al., 2008) challenges the claim that trustworthiness and expressions share the same perceptual mechanisms (Oosterhof & Todorov, 2009). The different pattern of visual and auditory information integration for trustworthiness and attractiveness judgments also suggest that, despite a high correlation between these judgments within the same modality, they are still distinct judgments. In other words, when asked to judge perceived trustworthiness, people do not rely on perceived attractiveness (at least, not entirely).

CONCLUSION

Our findings suggest that, while sharing some aspects with face impressions, voice impressions also display specific properties. First, voices, just like faces, can lead to formation of reliable trait impressions of trustworthiness, attractiveness and dominance. Our measures of reliability and inter-rater agreement were high and comparable to those of face impressions reported in the current study or elsewhere. Second, we showed that voice-based judgments of trustworthiness and attractiveness were highly correlated, perhaps due to a valence halo effect used to compensate the mismatch between the domain of the attractiveness judgment (predominantly visual) and the domain of the stimuli (auditory). Third, we showed that voice impressions do not correlate with face impressions (with the exception of perceived dominance for male stimuli). This suggests visual cues are not paired with auditory cues (diagnostic or not) when it comes to trait perception.

For face impressions, Tw judgments of male targets correlated positively with Att judgments and negatively with Dom judgments. In contrast, judgments of female targets did not correlate with each other. As we noted, a separate analysis per gender is necessary to avoid artificial correlations due to gender biases in trait perception. For example, if women are generally perceived as more trustworthy and more attractive than men, Tw and Att ratings will erroneously appear as correlated because of this divide between men and women.

Interpersonal perceptions in humans usually involve a combination of visual and auditory stimuli. In our second study we showed that an integrated person impression assigns different weights to facial and vocal information, depending on the trait. Voices and faces are equally important to an integrated trustworthiness impression, which also shows an interaction effect. In contrast, attractiveness seems to rely more on faces, while dominance seems to rely more on voices. The differential roles of faces and voices in attractiveness and dominance may be explained by the predominantly visual character of attractiveness and by the strong links between voice pitch, masculinity and dominance.

Chapter 5
CONCLUSIONS

The experiments in the current thesis aimed to contribute to a better understanding of several key topics in face perception, mainly revolving around the specificity of face processing mechanisms, the functional and neural bases of facial trait perception, and some outstanding issues related to face trustworthiness impressions. More specifically, I addressed the following questions:

1. Are face recognition mechanisms distinct from mechanisms involved in within-level discrimination of non-face objects? (Chapter 2.1)
2. Are face recognition mechanisms also involved in processing non-face objects for which one has acquired expertise? (Chapter 2.2)
3. Can facial trait perception dissociate from facial identity recognition? (Chapter 3.1)
4. Is normal perception of face trustworthiness and aggressiveness dependent on intact facial expression and sex perception mechanisms? (Chapter 3.2)
5. Is facial width-to-height ratio a reliable predictor of perceived trustworthiness? (Chapter 4.1)
6. How accurate are face trustworthiness impressions in real-life settings and do dynamic images improve accuracy? Does response accuracy correlate with confidence? (Chapter 4.2)
7. Does reputation abolish the effect of face trustworthiness in economic interactions? (Chapter 4.3)
8. How do faces and voices contribute to an integrated person impression? (Chapter 4.4)

In what follows, I will summarise the findings by question, present their implications and directions for future research.

5.1. Are face recognition mechanisms distinct from mechanisms involved in within-level discrimination of non-face objects?

This was a test of one of the alternative hypotheses to face specificity - the individuation hypothesis (Damasio et al., 1982), according to which face recognition mechanisms should not be different from those involved in exemplar recognition within non-face object classes. An extensive investigation (21 experiments) of a new case of acquired prosopagnosia, Herschel, provided evidence for face specific mechanisms. Behaviourally, Herschel was severely impaired with the recognition of familiar faces, discrimination between unfamiliar identities, and the perception of facial expression and gender. His visual recognition deficits were however largely restricted to faces. He showed normal recognition memory for a wide variety of object classes in several paradigms, normal ability to discriminate between highly similar items within a novel object category, fine discrimination between human bodies and intact ability to name basic objects (except mammals). Furthermore, Herschel displayed a normal face composite effect and typical global advantage and global interference effects in the Navon task, suggesting spared integration of both face and non-face information. Intriguingly, despite his fine performance with object recognition from brief presentations of intact images, Herschel failed visual closure tests requiring recognition of basic objects from degraded images. This abnormality in basic object recognition is at odds with his spared within-class recognition and presents a challenge to hierarchical models of object perception (e.g. Rosch et al., 1976), according to which successful within-level discrimination is dependent on intact basic level recognition.

The clear dissociation between Herschel's face recognition (impaired) and within-level recognition of non-face objects (normal) is strong evidence against the individuation hypothesis. However, Herschel's prosopagnosia is not "pure" (i.e. his visual recognition impairments are not strictly confined to faces) and several of his results warrant further investigation. First, additional visual and semantic testing of his difficulties at identifying mammals may provide further insights with respect to the organisation of the visual recognition system - patient studies (Levine &

Calvanio, 1989; Farah et al., 1991) and functional imaging studies (Chao et al., 1999; Noppeney et al., 2006) suggest a distinction between representations of living and nonliving objects. Second, his results at tests of general configural processing were mixed, with normal performance in the Navon task but impaired scores at visual closure. Further testing is necessary to clarify if his prosopagnosia (and maybe also his deficits with mammal recognition) is a result of impaired configural processing. Third, additional probing of Herschel's basic object recognition is necessary before drawing firm conclusions about a dissociation between normal within-level and degraded basic-level recognition.

5.2. Are face-specific mechanisms involved in processing non-face objects for which one has acquired expertise?

The expertise hypothesis - another alternative to face specificity - suggests that face recognition mechanisms operate on all objects with which an observer has acquired expertise (Diamond & Carey, 1986; Gauthier & Tarr, 1997). A widely used method for studying expertise in the laboratory involves a training procedure claimed to lead to the development of expertise with a novel object class named greebles. Challenging the expertise hypothesis, in section 2.2 I presented two cases of acquired prosopagnosia who showed normal accuracy and response times throughout the standard greeble training procedure along with severe deficits on a matched face training procedure. Furthermore, both acquired prosopagnosics fulfilled the criterion claimed to signal successful acquisition of greeble expertise. The results demonstrate that face expertise and greeble expertise rely on separate mechanisms.

The findings are novel - I showed that lesions to a normal brain can produce impairments with faces but leave intact the system for learning individual greebles, highly-similar exemplars within one object category. The two acquired cases who were previously trained with the greebles were unable to develop normal greeble expertise (Behrmann, Marotta et al., 2005; Bukach et al., 2012), but these prosopagnosic cases suffered from object recognition deficits (Bukach et al., 2012; Gauthier et al., 1999) which makes it unclear whether their difficulties with greebles stem from disruption of object recognition mechanisms or a general expertise mechanism. Duchaine et al. (2004) showed normal greeble learning in one developmental prosopagnosic, but this prosopagnosic's performance with faces in a comparable task was not assessed.

The findings are also robust - I presented two cases of acquired prosopagnosia (Herschel and Florence) whose face recognition deficits resulted from different lesion locations. The expertise training procedure for greebles was based on one used by Gauthier and Tarr (1997) and a face learning procedure was designed to match the greeble procedure. Both Herschel and Florence showed clear differences in performance between faces and greebles during the training sessions (1 to 4) and testing sessions (5 to 8). While their face performance was severely impaired, their greeble performance (accuracy and response times) was comparable to controls.

The findings should have a major impact on the field, because the research program involving the greebles has been highly influential. According to Web of Science, the original article introducing the greebles (Gauthier et al., 1997) has been cited 369 times, and two other papers claiming that recognition of expert objects is carried out by the same mechanisms involved in face recognition (Gauthier et al., 1999; Gauthier et al., 2000) have been cited more than 500 times each, with the number of yearly citations increasing. In the current study I ∂dresented compelling evidence that greeble expertise is not dependent on the same mechanisms as face expertise, which conclusively rejects evidence purportedly supporting the expertise account derived from greebles studies.

5.3. Can facial trait perception dissociate from facial identity recognition?

In Section 3.2 I presented compelling evidence that facial trait perception can dissociate from identity recognition. Four individuals with acquired prosopagnosia were tested for perception of face trustworthiness, attractiveness and aggressiveness, with three tests with different task demands (ratings, sorting, categorisation) per each trait. One prosopagnosic (Florence) showed normal trait perception in all tests. Each of the other three prosopagnosics was normal on all tests pertaining to at least one trait.

Successful trait perception in individuals with severe face recognition difficulties has been previously reported (Quadflieg et al., 2012; Todorov & Duchaine, 2008), but they had several limitations. First, they tested trait perception only with ratings, and ratings might not be sensitive enough to detect more subtle impairments. Second, trait perception measured with ratings was compared with identity recognition measured by other formats; it is possible the noted dissociation might have been influenced by the different test formats. Third, Quadflieg et al. (2012) used insufficiently controlled face stimuli (see introduction and discussion from Section 3.2), while Todorov and Duchaine (2008) measured only one trait (trustworthiness). We can also add that the individuals showing a dissociation between trait and identity perception in Todorov and Duchaine (2008) were developmental prosopagnosics, i.e. their face processing mechanisms did not develop normally. Extrapolation of any dissociation found in developmental prosopagnosic to a normal face processing system should be done with caution. In contrast, we tested four acquired prosopagnosics on multiple tests for each facial ability, with one test format (sorting) identical for trait and identity perception. Our stimuli were tightly controlled for non-face cues that might lead to apparently normal face evaluations.

The findings have implications for models of face perception. Trait perception is an important aspect of face processing, with considerable impact on social interactions. The fact that it seems to rely on distinct perceptual mechanisms from those involved in identity perception underlines the necessity to include it in any model aiming to explain how people perceive faces. Whether trait perception is performed by mechanisms which are also different from other aspects of face perception (e.g. expressions, sex) was examined in the following section.

5.4. Is normal perception of face trustworthiness and aggressiveness dependent on intact facial expression and sex perception mechanisms?

The overgeneralisation model of facial trait perception (Oostehof & Todorov, 2008) posits that trustworthiness judgments from faces are by-products of (happy and angry) expression perception mechanisms, and dominance judgments are by-products of sex perception mechanisms. According to this model, impairments with facial expression and sex recognition would also affect trustworthiness and dominance perception. However, contrary to these predictions, one individual with acquired prosopagnosia (Florence) showed normal aggressiveness judgments despite impaired sex perception. Aggressiveness is highly correlated with dominance and had the highest loading on the dominance component in the two-dimensional model of trait perception proposed by Oosterhof and Todorov (2008). Furthermore, Florence provided face trustworthiness judgments in line with controls but she showed mixed performance with expression tests: she failed one test, scored in the lower normal range for two tests, and succeeded at one. Although these results do not show severe deficits, they question the integrity of Florence's expression perception mechanisms.

Perception of each expression, trait and sex was assessed with two tests, one requiring sorting and the other requiring categorisation of faces on the judged dimension. Therefore, observed dissociations were robust and could not be attributed to different task demands. Face stimuli were black and white images cropped to show only internal facial features, so that perception of traits could not rely on skin colour or external cues.

The results are clearly inconsistent with the emotion and sex overgeneralisation theories of facial trait perception, showing that judgments of face trustworthiness and aggressiveness/dominance do not depend on intact expression and sex recognition mechanisms. The findings do not question the fact that subtle cues to happy and angry expressions may influence perceived trustworthiness, similar to how a more feminine/masculine appearance might influence perceived dominance/aggressiveness. They are simply not indicative of a tight relationship between these abilities, suggesting a more comprehensive account of trait perception is needed.

The fact that trait perception can dissociate from identity recognition (Section 3.1), expression and sex recognition (Section 3.2) suggests further divisions in general face models. The new battery of tests of face perception will be extended by developing similar sorting and categorisation tasks examining perception of other traits, all basic expressions, gaze, and also aspects related to physical properties of faces (e.g. shape and reflectance cues, spacing and part cues, cues pertaining to different spatial frequencies). These tests can then be used for comprehensive investigations of more cases of acquired prosopagnosia to reveal associations and dissociations between perception of various facial aspects, leading to a refinement of existing models of face perception.

5.5. Is facial width-to-height ratio (WHR) a reliable predictor of perceived trustworthiness?

Following Stirrat and Perrett's (2010) assertion that facial WHR is a valid cue to trustworthiness that is used by people in their judgments of face trustworthiness, I examined whether WHR influences perceived trustworthiness when faces are controlled for attractiveness and femininity/masculinity. Participants were asked to play trust games with presumed real partners whose faces varied on WHR (but not attractiveness or femininity). Contrary to Stirrat and Perrett (2010), the results did not reveal a preference towards faces with lower WHR. Instead, we found that perceived femininity/masculinity of faces predicted participants' choices. This effect was replicated with the original set of face stimuli used by Stirrat and Perrett (2010). When participants were asked to select between two similar faces differing on WHR, they generally preferred the narrower faces (just like in Stirrat & Perrett, 2010). However, they also judged the narrower faces as more feminine-looking and we found a large correlation ($r = .90$) between choices based on the trustworthiness dimension and those based on the perceived femininity dimension. The results of these two experiments suggest face trustworthiness judgments are driven by perceived femininity/masculinity rather than WHR.

Further studies are needed to confirm this suggestion. For example, showing that wider faces with more feminine features are perceived to be more trustworthy than narrower faces with more masculine features would be compelling evidence. One way to make faces appear more feminine/masculine is to modify their eyebrows (Bruce et al., 1993).

5.6. How accurate are face trustworthiness impressions in real-life settings and do dynamic images improve accuracy? Does response accuracy correlate with confidence?

Previous studies furnished mixed evidence on the accuracy of face trustworthiness judgments. Their criterion was behaviour in laboratory games (i.e. cooperative or not) or group membership (i.e. criminals versus non-criminals). Using data from a television programme, where contestants played economic games for real and substantial amounts of money, I investigated if deceptive behaviour can be inferred from faces. The stimuli were either static images (face stills) or dynamic images (five second face clips, no sound) recorded at the time when deception was likely to occur. In addition and in line with previous studies, I examined whether deceptive behaviour in laboratory trust games can be detected from static and dynamic face stimuli.

Overall, the results from three experiments suggest that people can distinguish between cooperators and cheaters from faces. For laboratory stimuli, discrimination was successful only for female stills, and at chance for male stills and female and male clips. For real-world stimuli, response accuracy did not vary between male and female stimuli. Performance with clips was above chance in two out of two experiments, while performance with stills was better than chance in two out of three studies. There was no clear advantage noted for clips compared to stills.

These results suggest that people are capable of deriving useful information from faces about deceptive behaviour. Notably, the performance is only slightly higher than chance. To maximise the benefits provided by the ability to “read” faces, an insight with respect to when face impressions are accurate or not is required. However, the null correlations between response accuracy and confidence reveal that people do not have this insight. The implication is that individual decisions to trust or not another party should not be guided by face impressions.

An interesting question is whether trustworthiness detection can be improved. For example, implicit evaluations were found to be more accurate than explicit judgments of trustworthiness (Bonnefon et al., 2012). Similarly, it is conceivable more analytical evaluations (such as asking participants to reflect on their responses by providing confidence ratings, like in our study) may decrease accuracy. Perhaps trust-related consequential decisions made on a set of faces would be more successful when there is no attention drawn to what is actually measured.

5.7. Does reputation abolish the effect of face trustworthiness in economic interactions?

Previous studies demonstrated that face trustworthiness influences economic interactions when agents do not have any information about each other (e.g. Stirrat & Perrett, 2010; van't Wout & Sanfey, 2008). However, economic partners typically have a reputation built on previous actions, and reputation is a stronger predictor of future behaviour than face configuration. Therefore, when reputation is known, it is conceivable that the effect of face trustworthiness will disappear. Contrary to this possibility, I showed that face trustworthiness still matters, although the impact is reduced and much smaller than the impact of information about past behaviour. Interestingly, face trustworthiness and reputation did not interact.

For research on face impressions, the results demonstrate the importance of replicating potentially interesting findings obtained in strictly controlled laboratory conditions in more realistic settings. For example, economic interactions almost never occur between partners who do not know anything about each other. There is a large face trustworthiness premium in these artificial conditions (around 42%, see Rezlescu, Duchaine et al., 2012), but in situations which better simulate real interactions the premium turns out to be far more modest (6%), although still significant. In real life, the face trustworthiness premium is likely to be even more reduced, considering the plethora of factors affecting social and economic interactions.

On the other hand, an interesting question to be answered by future studies is whether the face trustworthiness premium varies with the degree of ambiguity in reputation. In the present study, potential economic partners had markedly positive or negative reputation, which might have decreased reliance on facial cues. Crossmodal integration in sensory studies suggest a channel's dominance depends on signal reliability.

5.8. How do faces and voices contribute to an integrated person impression?

In Section 4.4 I showed that voices, like faces, determine trait impressions that are highly reliable (i.e. consistent across listeners). Voice trustworthiness correlated with voice attractiveness (these two traits correlate for faces as well), but not with voice dominance (in contrast with face impressions correlations). There was no significant correlation between voice and face impressions, with the exception of a negative correlation for perceived dominance in men. This might be a compensatory strategy, by which men with less dominant looks compensate through their voices.

When participants were asked to provide a person impression based on face *and* voice, different patterns of information integration emerged. For all measured traits (trustworthiness, attractiveness, dominance), the main effects of face and voice impressions were significant, but the effect sizes differed. For perceived trustworthiness, the effect sizes were comparable and face and voice impressions also showed an interaction effect. For attractiveness, the effect size of face impressions was four times larger than the effect size of voice impressions. For dominance, the reverse was true: the effect size of voice impressions was approximately three times larger than the effect size of face impressions. Faces and voices did not show an interaction effect for either attractiveness or dominance.

The results for attractiveness and dominance were explained in terms of perceived signal reliability for face and voice information. On one hand, because attractiveness is predominantly a visual concept, visual information is likely to be considered most reliable and thus given higher weight in integrated judgments of attractiveness. On the other hand, dominance is related to perception of masculinity (Oosterhof & Todorov, 2008) and voice is highly sexually dimorphic. Furthermore, testosterone level, which correlates with dominant/aggressive behaviour (Mazur & Booth, 1998), was linked to deep voices (Dabbs & Mallinger, 1999). Therefore, impressions based on voices are likely to be judged as more reliable for male dominance and given more weight than faces in integrated perception of dominance.

Future studies should extend the number of traits measured and examine whether variance in various voice impressions can be explained by a reduced number of dimensions. A similar approach in face impressions has produced a two-dimensional model of facial trait perception (Oosterhof & Todorov, 2008), with trustworthiness and dominance the principal components. It would be interesting to learn if this two-dimensional model is specific to faces or is amodal, extending to trait impressions from voices. Furthermore, it would be interesting to verify if the proposed dependency of face trustworthiness/dominance judgments on facial expression/sex recognition mechanisms replicates with voices.

Spectral analysis of voices has the potential to reveal important physical attributes associated with particular trait impressions. Furthermore, it can reveal potential valid cues to actual personality traits or behavioural propensities.

GENERAL SUMMARY

In summary, this thesis has investigated several key topics related to social perception of faces. In Chapters 2 and 3, acquired prosopagnosia was used as a tool to: i) test (and reject) two alternative hypotheses to face specificity; ii) confirm dissociations between face identity and trait perception, and test the emotion and sex overgeneralisation theories in facial trait perception. In Chapter 4 behavioural studies with healthy participants examined facial cues to perceived trustworthiness, accuracy of face trustworthiness judgments and interactions of face impressions with reputation and voice impressions. In addition to answering important questions, some of these studies have paved the way for future explorations into the cognitive neuroscience and social cognition of faces.

References

Abrosoft Co. (2002). FantaMorph (Version 3.0) [computer software]. Beijing: Abrosoft. Available from http://www.facegen.com.

Adolphs, R., & Tranel, D. (1999). Intact recognition of emotional prosody following amygdala damage. *Neuropsychologia*, 37(11), 1285–92.

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature*, 393, 470-474.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, 372(6507), 669-672.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. R. (1995). Fear and the human amygdala. *Journal of Neuroscience*, 15(9), 5879-5891.

Aharon, I., Etcoff, N., Ariely, D., Chabris, C. F., O'Connor, E., & Breiter, H. C. (2001). Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron*, 32(3), 537-551.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in cognitive sciences*, *4*(7), 267–278.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 256–274.

Albright, L., Kenny, D., & Malloy, T. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, 55, 387-395.

Albright, L., Malloy, T. E., Dong, Q., Kenny, D. A., Fang, X., Winquist, L., et al. (1997). Cross-cultural consensus in personality judgments. *Journal of Personality and Social Psychology*, 72, 558–569.

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, 69(3), 518–529.

Arrow, K. (1973). *The limits of organization*. New York: Norton & Company.

Ashworth, A. R. S., Vuong, Q. C., Rossion, B., & Tarr, M. J. (2008). Recognizing rotated faces and Greebles: What properties drive the face inversion effect? *Visual Cognition*, *16*(6), 754–784. doi:10.1080/13506280701381741

Assal, G., Favre, C., & Anderes, J.P. (1984). Nonrecognition of familiar animals by a farmer: zooagnosia or prosopagnosia for animals. *Revue Neurologique* (Paris), 140(10), 580– 584.

Banissy, M. J., Sauter, D. A., Ward, J., Warren, J. E., Walsh, V., & Scott, S. K. (2010). Suppressing sensorimotor activity modulates the discrimination of auditory emotions but not speaker identity. *Journal of Neuroscience*, 30(41), 13552–7.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, 6(2), 269–78. doi:10.1037/1528-3542.6.2.269

Barclay, P., & Lalumiere, M. L. (2006). Do people differentially remember cheaters? *Human Nature*, 17, 98–113.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42, 241–251.

Barton, J. J. S. (2008). Structure and function in acquired prosopagnosia: Lessons from a series of 10 patients with brain damage. *Journal of Neuropsychology*, 2, 197–225.

Barton, J. J. S., Cherkasova, M. V., Press, D. Z., Intriligator, J. M., & O'Connor, M. (2004). Perceptual functions in prosopagnosia. *Perception*, 33, 939–956.

Bauer, R. M., & Trobe, J. D. (1984). Visual memory and perceptual impairments in prosopagnosia. *Journal of Clinical Neuro-ophthalmology*, 4, 39–46.

Behrmann, M., Avidan, G., Marotta, J. J., & Kimchi, R. (2005). Detailed exploration of face-related processing in congenital prosopagnosia: 1. Behavioral findings. *Journal of Cognitive Neuroscience*, 17, 1130–1149.

Behrmann, M., & Kimchi, R. (2003). What does visual agnosia tell us about perceptual organization and its relationship to object perception? *Journal of Experimental Psychology: Human Perception and Performance*, 29, 19–42.

Behrmann, M., Marotta, J., Gauthier, I., Tarr, M. J., & McKeeff, T. J. (2005). Behavioral change and its neural correlates in visual agnosia after expertise training. *Journal of Cognitive Neuroscience*, 17(4), 554-68.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3), 129–35.

Belot, M., Bhaskar, V., & Van De Ven, J. (2012). Can observers predict trustworthiness? *The Review of Economics and Statistics*, 94(1), 246–259.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, *8*(6), 551–565. doi:10.1162/jocn.1996.8.6.551

Bentin, S., Taylor, M. J., Rousselet, G. A., Itier, R. J., Caldara, R., Schyns, P. G., Jacques, C., et al. (2007). Controlling interstimulus perceptual variance does not abolish N170 face sensitivity. *Nature Neuroscience*, *10*(7), 801–802. doi: 10.1093/cercor/bhl100

Benton, A. L., & van Allen, M. W. (1972). Prosopagnosia and facial discrimination. *Journal of the Neurological Sciences*, 15, 167–172.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior*, 10, 122–142.

Berinsky, A.J., Huber, G.A., & Lenz, G.S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3), 351–368.

Berry, D. S. (1990). Taking people at face value: Evidence for the kernel of truth hypothesis. *Social Cognition*, 8, 343–361.

Berry, D. S. (1991). Accuracy in social perception: contributions of facial and vocal information. *Journal of Personality and Social Psychology*, 61(2), 298–307.

Bodamer, J. (1947). Die Prosop-Agnosie. *Archiv für Psychiatrie und Nervenkrankheiten*, *179*(1-2), 6–53. doi:10.1007/BF00352849

Bond, C. F., Berry, D. S., & Omar, A. (1994). The kernel of truth in judgments of deceptiveness. *Basic and Applied Social Psychology*, 15, 523–534.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3), 214–234. doi:10.1207/s15327957pspr1003

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2012). The Modular Nature of Trustworthiness Detection. *Journal of experimental psychology. General*, Advance online publication. doi:10.1037/a0028930

Boutet, I., & Faubert, J. (2006). Recognition of faces and complex objects in younger and older adults. *Memory & Cognition*, *34*(4), 854–64.

Boutsen, L., & Humphreys, G. W. (2002). Face context interferes with local part processing in a prosopagnosic patient. *Neuropsychologia*, *40*(13), 2305–13.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., . . . & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant – stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, 26, 423–455.

Brants, M., Wagemans, J., & Op de Beeck, H. P. (2011). Activation of fusiform face area by Greebles is related to face similarity but not expertise. *Journal of Cognitive Neuroscience*, 23(12), 3949–58.

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., et al. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, 17, 875–887.

Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., & Linney, A. (1993). Sex discrimination: how do we tell the difference between male and female faces? *Perception*, 22, 131-152.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327.

Bruyer, R., & Crispeels, G. (1992). Expertise in person recognition. *Bulletin of the Psychonomic Society*, *30*, 501–504.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on *Psychological Science*, 6(1), 3–5.

Bukach, C. M., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: the power of an expertise framework. *Trends in Cognitive Sciences*, *10*(4), 159–66. doi:10.1016/j.tics.2006.02.004

Bukach, C. M., Gauthier, I., Tarr, M. J., Kadlec, H., Barth, S., Ryan, E., Turpin, J., et al. (2012). Does Acquisition of Greeble Expertise in Prosopagnosia Rule Out a Domain-General Deficit? *Neuropsychologia*, 50(2), 289-304.

Bukach, C. M., Bud, D. N., Gauthier, I., & Tarr, M. J. (2006). Perceptual expertise effects are not all or none: Spatially limited perceptual expertise for faces in a case of prosopagnosia. *Journal of Cognitive Neuroscience*, 18, 48–63.

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., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, 42(1), 286-291.

Busey, T. A, & Vanderkolk, J. R. (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research*, *45*(4), 431–48. doi:10.1016/j.visres.2004.08.021

Busigny, T., Graf, M., Mayer, E., & Rossion, B. (2010). Acquired prosopagnosia as a face-specific disorder: ruling out the general visual similarity account. *Neuropsychologia*, *48*(7), 2051–67. doi:10.1016/j.neuropsychologia.2010.03.026

Busigny, T., Joubert, S., Felician, O., Ceccaldi, M., & Rossion, B. (2010). Holistic perception of the individual face is specific and necessary: Evidence from an extensive case study of acquired prosopagnosia. *Neuropsychologia*, *48*(14), 4057–4092. doi:10.1016/j.neuropsychologia.2010.09.017

Busigny, T., & Rossion, B. (2010). Acquired prosopagnosia is not due to a general impairment in fine-grained recognition of exemplars of a visually homogeneous category. *Behavioural Neurology*, *23*(4), 229–31. doi:10.3233/BEN-2010-0302

Busigny, T., & Rossion, B. (2011). Holistic processingimpairment can be restricted to faces in acquired prosopagnosia: Evidence from the global/local Navon effect. *Journal of Neuropsychology*, 5, 1–14.

Buxbaum, L. J., Glosser, G., & Coslett, H. B. (1996). Relative sparing of object recognition in alexia-prosopagnosia. *Brain and Cognition*, 32, 202–205.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641–651.

Caramazza, A., & Shelton, J. R. (1998). Domain- specific knowledge systems in the brain: The animate–inanimate distinction. *Journal of Cognitive Neuroscience*, 10, 1–34.

Carey, S. (1992). Becoming a face expert. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 335, 95-103.

Carré, J. M., & McCormick, C. M. (2008). In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B: Biological Sciences*, 275(1651), 2651–6. doi:10.1098/rspb.2008.0873

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science*, 20(10), 1194–8. doi:10.1111/j.1467-9280.2009.02423.x

Carré, J. M., Morrissey, M. D., Mondloch, C. J., & McCormick, C. M. (2010). Estimating aggression from emotionally neutral faces: Which facial cues are diagnostic? *Perception*, 39(3), 356–377. doi:10.1068/p6543

Chang, L. J., Doll, B. B., Van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology*, 61(2), 87–105. doi:10.1016/j.cogpsych.2010.03.001

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, 2, 913–919.

Chatterjee, G., & Nakayama, K. (2013). Normal facial age and gender perception in developmental prosopagnosia. *Cognitive Neuropsychology*. In press.

Chiappe, D., & Brown, A. (2004). Cheaters are looked at longer and remembered better than cooperators in social exchange situations. *Evolutionary Psychology*, 2, 108-120.

Cloutier, J., Heatherton, T. F., Whalen, P. J., & Kelley, W. M. (2008). Are Attractive People Rewarding ? Sex Differences in the Neural Substrates of Facial Attractiveness. *Journal of Cognitive Neuroscience*, 20(6), 941–951.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, 1242, 126–35.

Collins, S. (2000). Men's voices and women's choices. *Animal Behaviour*, 60(6), 773–780. doi:10.1006/anbe.2000.1523

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, 65(5), 997–1004.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies in the Wason Selection Task. *Cognition*, 31, 187-276.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby, (Eds.), *The adapted mind: Evolutionary* *psychology and the generation of culture* (pp. 163–228) New York, NY: Oxford University Press.

Crawford, J.R., & Garthwaite, P.H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, 40, 1196-1208.

Crawford, J. R., & Garthwaite, P. H. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology*, 19, 318-331.

Crawford, J. R. & Garthwaite, P. H. (2007). Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach. *Cognitive Neuropsychology*, 24, 343-372

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *Clinical Neuropsychologist*, 12, 482–486.

Cunningham, M.R., Roberts, A.R., Barbee, A.P., Druen, P.B., & Wu, C-H. (1995). "Their ideas of beauty are, on the whole, the same as ours": consistency and variability in the cross- cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology*, 68, 261–279.

Damasio, A. R., Damasio, H., & Van Hoesen, G. W. (1982). Prosopagnosia: anatomic basis and behavioral mechanisms. *Neurology*, *32*(4), 331–41.

Dabbs Jr., J. M., & Mallinger, A. (1999). High testosterone levels predict low voice pitch among men. *Personality and Individual Differences*, 27, 801–804.

Dalrymple, K. A., Oruç, I., Duchaine, B., Pancaroglu, R., Fox, C. J., Iaria, G., ... & Barton, J. J. (2011). The anatomic basis of the right face-selective N170 IN acquired prosopagnosia: A combined ERP/fMRI study. *Neuropsychologia*, 49(9), 2553-2563.

DeBruine, L.M. (2002). Facial resemblance enhances trust. *Proceedings of the Royal Society B: Biological Sciences*, 269, 1307–1312.

DeBruine, L. M. (2005). Trustworthy but not lust-worthy: context-specific effects of facial resemblance. *Proceedings of the Royal Society B: Biological Sciences*, 272(1566), 919–22. doi:10.1098/rspb.2004.3003

De Gelder, B., Bachoud-Lévi, A. C., & Degos, J. D. (1998). Inversion superiority in visual agnosia may be common to a variety of orientation polarised objects besides faces. *Vision Research*, *38*(18), 2855–61.

Delvenne, J.-F., Seron, X., Coyette, F., & Rossion, B. (2004). Evidence for perceptual deficits in associative visual (prosop)agnosia: a single-case study. *Neuropsychologia*, *42*(5), 597–612. doi:10.1016/j.neuropsychologia. 2003.10.008

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., & Duchaine, B. (2012). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dis- sociations from face memory, and expertise effects. *Behavioral Research Methods*, 44, 587–605.

DePaulo, B. M., Stone, J. I., & Lassiter, G. D. (1985). Deceiving and detecting deceit. In B. R. Schlenker (Ed.), *The self and social life* (pp. 323-370). New York: McGraw-Hill.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review*, 1(4), 346–357.

De Renzi, E. (1986a). Current issues on prosopagnosia. In H. D. Ellis, M. A. Jeeves, F. Newcombe, & A. Young (Eds.), *Aspects of face processing* (pp. 243–252). Dordrecht, The Netherlands: Martinus Nijhoff.

De Renzi, E. (1986b). Prosopagnosia in two patients with CT scan evidence of damage confined to the right hemisphere. *Neuropsychologia*, 24, 385–389.

De Renzi, E., & di Pellegrino, G. (1998). Prosopagnosia and alexia without object agnosia. *Cortex*, 34, 403–415.

De Renzi, E., Faglioni, P., Grossi, D., & Nichelli, P. (1991). Apperceptive and associative forms of prosopagnosia. *Cortex*, 27, 213–221.

De Renzi, E., Faglioni, P., & Spinnler, H. (1968). The performance of patients with unilateral brain damage on face recognition tasks. *Cortex*, 4, 17–34.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of experimental psychology. General*, *115*(2), 107–17.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3), 285–290.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–3. doi:10.1126/science.1063414

Duarte, J., Siegel, S., & Young, L. A. (2010). Trust and credit. American Finance Association Annual Meeting, Atlanta, retrieved April 1, 2011, from SSRN: http://ssrn.com/ abstract=1343275

Duchaine, B. (2000). Developmental prosopagnosia with normal configural processing. *Neuroreport*, 11, 79–83.

Duchaine, B., Dingle, K., Butterworth, E., & Nakayama, K. (2004). Normal greeble learning in a severe case of developmental prosopagnosia. *Neuron*, 43(4), 469–73. doi:10.1016/j.neuron.2004.08.006

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*(4), 419–30. doi:10.1080/02643290701380491

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental proso- pagnosia. *Journal of Cognitive Neuroscience*, 17, 249–261.

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

Duchaine, B., Yovel, G., Butterworth, E., & Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology*, *23*(5), 714–747. doi:10.1080/02643290500441296

Duchaine, B., Yovel, G., & Nakayama, K. (2007). No global processing deficit in the Navon task in 14 developmental prosopagnosics. *Social Cognitive and Affective Neuroscience*, 2, 104–113.

Dzhelyova, M. P., Ellison, A., & Atkinson, A. P. (2011). Event-related repetitive TMS reveals distinct, critical roles for right OFA and bilateral posterior STS in judging the sex and trustworthiness of faces. *Journal of Cognitive Neuroscience*, 23(10), 2782–2796.

Ehrenkranz, J., Bliss, E., & Sheard, M. (1974). Plasma testosterone: Correlation with aggressive behavior and social dominance in men. *Psychosomatic Medicine*, 36, 469–75.

Ekstrom, R., French, J. W., & Harman, H. H. (1976). Manual for Kit of Factor-referenced Cognitive Tests. Princeton, NJ: Educational Testing Service.

Eimer, M. (1998). Does the face-specific N170 component reflect the activity of a specialized eye processor? *Neuroreport*, *9*(13), 2945–8.

Eimer, M. (2000). Effects of face inversion on the structural encoding and recognition of faces. Evidence from event-related brain potentials. *Brain research. Cognitive brain research*, *10*(1-2), 145–58.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, 32, 88–106.

Ellis, A.W. (1989). Neuro-cognitive processing of faces and voices. In Young, A.W. & Ellis, H.D. (Eds.), *Handbook of Research on Face Processing* (pp. 207–215). Elsevier.

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in human amygdala. *Journal of Cognitive Neuroscience*, 19, 1508-1519.

Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Common neural mechanisms for the evaluation of facial trustworthiness and emotional expressions as revealed by behavioral adaptation. *Perception*, 39(7), 931–941. doi:10.1068/p6633

Fantz, R. L. (1963). Pattern vision in newborn infants. *Science*, *140*(3564), 296–297.

Farah, M. J. (1996). Is face recognition “special”? Evidence from neuropsychology. *Behavioural Brain Research*, 76, 181–189.

Farah, M. J. (2004). *Visual agnosia: Disorders of object rec- ognition and what they tell us about normal vision*. Cambridge, MA: MIT Press.

Farah, M. J., Levinson, K. L., & Klein, K. L. (1995). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33, 661–674.

Farah, M. J., McMullen, P. A., & Meyer, M. M. (1991). Can recognition of living things be selectively impaired? *Neuropsychologia*, 29, 185–193.

Feinberg, D. R. (2008). Are human faces and voices ornaments signaling common underlying cues to mate value? *Evolutionary Anthropology: Issues, News, and Reviews*, 17(2), 112–118. doi:10.1002/evan.20166

Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008). The role of femininity and averageness of voice pitch in aesthetic judgments of women’s voices. *Perception*, 37(4), 615–623. doi:10.1068/p5514

Feinberg, D. R., Jones, B., DeBruine, L., Moore, F., Law Smith, M., Cornwell, R., Tiddeman, B., Boothroyd, L., & Perrett, D. (2005). The voice and face of woman: one ornament that signals quality? *Evolution and Human Behavior*, 26, 398–408.

Fetchenhauer, D., Groothuis, T., & Pradel, J. (2010). Not only states but traits? Humans can identify permanent altruistic dispositions in 20 seconds. *Evolution and Human Behavior*, 31, 80-86.

Fink, B., Grammer, K., & Matts, P. (2006). Visible skin color distribution plays a role in the perception of age, attractiveness, and health in female faces. *Evolution and Human Behavior*, 27(6), 433–442. doi:10.1016/j.evolhumbehav.2006.08.007

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, 11(2), 77–83. doi:10.1016/j.tics.2006.11.005

Fitch,W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, 106, 1511–1522.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fox, C. J., Hanif, H. M., Iaria, G., Duchaine, B. C., & Barton, J. J. S. (2011). Perceptual and anatomic patterns of selective deficits in facial identity and expression processing. *Neuropsychologia*, 49, 3188–3200.

Fox, C. J., Moon, S. Y., Iaria, G., & Barton, J. J. S. (2009). The correlates of subjective perception of identity and expression in the face network: an fMRI adaptation study. *NeuroImage*, 44(2), 569–580. doi:10.1016/j.neuroimage.2008.09.011.The

Freiwald, W. A, Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. Nature Neuroscience, 12(9), 1187–96. doi: 10.1038/nn.2363

Fukuyama, F. (1995). *Trust: Social virtues and the creation of prosperity*. New York: Free Press.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., Schweinberger, S. R., et al. (2009). Developmental phonagnosia: a selective deficit of vocal identity recognition. *Neuropsychologia*, 47(1), 123–31.

Garrido, L., Furl, N., Draganski, B., Weiskopf, N., Stevens, J., Tan, G. C. Y., . . . & Duchaine, B. (2009). Voxel-based morphometry reveals reduced grey matter volume in the temporal cortex of developmental prosopagnosics. *Brain*, 132, 3443–3455.

Gauthier, I., Behrmann, M., & Tarr, M. J. (1999). Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience*, 11, 349–370.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, a W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191–7. doi:10.1038/72140

Gauthier, I., & Tarr, M. J. (1997). Becoming a “greeble” expert: Exploring mechanisms for face recognition. *Vision Research*, *37*(12), 1673–1682.

Gauthier, I., & Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition : Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2), 431- 446.

Gauthier, I., Tarr, M. J., Anderson, A W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform “face area” increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*(6), 568–73. doi: 10.1038/9224

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, a W. (2000). The fusiform “face area” is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience*, *12*(3), 495–504.

Gauthier, I, Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training “greeble” experts: a framework for studying expert object recognition processes. *Vision Research*, *38*(15-16), 2401–28.

Ge, L., Wang, Z., McCleery, J. P., & Lee, K. (2006). Activation of face expertise and the inversion effect. *Psychological Science*, *17*(1), 12–16. doi:10.1111/j. 1467-9280.2005.01658.x.Activation

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19(5), 847–57.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9, 416–429.

Gollwitzer, M., Rothmund, T., Alt, B., & Jekel, M. (2012). Victim Sensitivity and the Accuracy of Social Judgments. *Personality & Social Psychology Bulletin*. doi:10.1177/0146167212440887

Goren, C. C., Sarty, M., & Wu, P. Y. K. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, *56*, 544–549.

Gregory Jr., S. W., & Gallagher, T. J. (2002). Spectral Analysis of Candidates' Nonverbal Vocal Communication: Predicting U.S. Presidential Election Outcomes. *Social Psychology Quarterly*, 65(3), 298. doi:10.2307/3090125

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience*, *7*(5), 555–62. doi:10.1038/nn1224

Grueter, M., Grueter, T., Bell, V., Horst, J., Laskowski, W., Sperling, K., Halligan, P. W., et al. (2007). Hereditary prosopagnosia: the first case series. *Cortex*, *43*, 734–749.

Hamermesh, D. (2011). *Beauty pays: Why attractive people are more successful*. Princeton Press: Princeton.

Harley, E. M., Pope, W. B., Villablanca, J. P., Mumford, J., Suh, R., Mazziotta, J. C., ... & Engel, S. A. (2009). Engagement of fusiform cortex and disengagement of lateral occipital cortex in the acquisition of radiological expertise. *Cerebral Cortex*, 19(11), 2746-2754.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Science*, 1, 78–82.

Haselhuhn, M. P., & Wong, E. M. (2011). Bad to the bone: facial structure predicts unethical behaviour. *Proceedings of the Royal Society B: Biological Sciences*, (June), 6–11. doi:10.1098/rspb.2011.1193

Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, 78, 837–852.

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

Haxby, J.V., Hoffman, E.A. & Gobbini, M.I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry*, 51, 59–67.

Haxby, J.V., & Gobbini, M.I. (2011). Distributed neural systems for face perception. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V Haxby (Eds.), *The Oxford Handbook of Face Perception* (pp. 631-652). Oxford University Press.

Henke, K., Schweinberger, S. R., Grigo, A., Klos, T., & Sommer, W. (1998). Specificity of face recognition: Recognition of exemplars of non-face objects in prosopagnosia. *Cortex*, 34, 289–296.

Henson, R. N., Mattout, J., Singh, K. D., Barnes, G. R., Hillebrand, a, & Friston, K. (2007). Population-level inferences for distributed MEG source localization under multiple constraints: application to face-evoked fields. *NeuroImage*, *38*(3), 422–38. doi:10.1016/j.neuroimage.2007.07.026

Hoover, A. E. N., Démonet, J.-F., & Steeves, J. (2010). Superior voice recognition in a patient with acquired prosopagnosia and object agnosia. *Neuropsychologia*, 48(13), 3725–32.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3), 399–425.

Husk, J. S., Bennett, P. J., & Sekuler, A. B. (2007). Inverting houses and textures: investigating the characteristics of learned inversion effects. *Vision research*, *47*(27), 3350–9.

Itier, R. J., & Taylor, M. J. (2004). Source analysis of the N170 to faces and objects. *Neuroreport*, *15*(8), 1261–65.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17, 825–841.

Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2010). A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*, 79(1), 57-62.

Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2004). When facial attractiveness is only skin deep. *Perception*, 33(5), 569–576.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–11.

Kleisner, K., Kocnar, T., Rubesova, A., & Flegr, J. (2010). Eye color predicts but does not directly influence perceived dominance in men. *Personality and Individual Differences*, 49, 59–64. doi: 10.1016/j.paid.2010.03.011.

Kleisner, K., Priplatova, L., Frost, P., & Flegr, J. (2013). Trustworthy-Looking Face Meets Brown Eyes. *PLoS ONE*, 8(1), e53285. doi:10.1371/ journal.pone. 0053285.

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738), 2698–2704.

Kramer, R.S.S., Jones, A.L., & Ward, R. (2012). A Lack of Sexual Dimorphism in Width-to-Height Ratio in White European Faces Using 2D Photographs, 3D Scans, and Anthropometry. *PLoS ONE* 7(8): e42705. doi:10.1371/ journal.pone.0042705.

Kranz, F., & Ishai, A. (2006). Face perception is modulated by sexual preference. *Current Biology*, 16, 63–68.

Kung, C.-C., Peissig, J. J., & Tarr, M. J. (2007). Is region-of-interest overlap comparison a reliable measure of category specificity? *Journal of Cognitive Neuroscience*, 19(12), 2019–2034.

Lander, K. (2008). Relating visual and vocal attractiveness for moving and static faces. *Animal Behaviour*, 75(3), 817–822. doi:10.1016/j.anbehav.2007.07.001

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, 24(8), 1377-1388.

Lê, S., Cardebat, D., Boulanouar, K., Hénaff, M. A., Michel, F., Milner, D., ... & Démonet, J. F. (2002). Seeing, since childhood, without ventral stream: a behavioural study. *Brain*, 125(1), 58-74.

Leder, H., & Carbon, C.-C. (2006). Face-specific configural processing of relational information. *British Journal of Psychology (London, England : 1953)*, *97*(Pt 1), 19–29. doi:10.1348/000712605X54794

Levine, D. N., & Calvanio, R. (1989). Prosopagnosia: a defect in visual configural processing. *Brain and Cognition*, *10*(2), 149–70.

Levine, D. N., Calvanio, R., & Wolf, E. (1980). Disorders of visual behavior following bilateral posterior cerebral lesions. *Psychological Research*, 41, 217–234.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159–183.

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology*, 98(1), 111–126. doi:10.1348/000712606X109648

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska directed emotional faces [Database of standardized facial images]. Psychology Section, Department of Clinical Neuroscience, Karolinska Hospital, S-171 76 Stockholm, Sweden.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York, NY: Cambridge University Press.

Matts, P. J., Fink, B., Grammer, K., & Burquest, M. (2007). Color homogeneity and visual perception of age, health, and attractiveness of female facial skin. *Journal of the American Academy of Dermatology*, 57(6), 977–84. doi: 10.1016/j.jaad.2007.07.040

Mazur, A., & Booth, A. (1998). Testosterone and dominance in men. *Behavioral and Brain Sciences*, 21(3), 353–363.

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-Specific Processing in the Human Fusiform Gyrus. *Journal of Cognitive Neuroscience*, *9*(5), 605–610.

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences of the United States of America*, 109(42), 17063–8. doi:10.1073/pnas.1116333109

McKone, E., & Kanwisher, N. (2005). Does the human brain process objects of expertise like faces? A review of the evidence. In S. Dehaene, J. R. Duhamel, M. Hauser, & G. Rizzolatti (Eds.), *From Monkey Brain to Human Brain* (pp. 339–356). Cambridge, MA: MIT Press.

McKone, E., Kanwisher, N., & Duchaine, B. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11*(1), 8–15. doi: 10.1016/j.tics.2006.11.002

McKone, E., & Robbins, R. (2011). Are faces special? In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V Haxby (Eds.), *The Oxford Handbook of Face Perception* (Vol. 4, pp. 149–176). Oxford: OUP. doi:10.2307/302723

McNeil, J. E., & Warrington, E. K. (1993). Prosopagnosia: a face-specific disorder. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, *46*(1), 1–10.

Mealy, L., Daood, C., & Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethology and Sociobiology*, 17, 119-128.

Mehl, B., & Buchner, A. (2008). No enhanced memory for faces of cheaters. *Evolution and Human Behavior*, 29(1), 35–41. doi:10.1016/j.evolhumbehav.2007.08.001

Meyer, M., Baumann, S., Wildgruber, D., & Alter, K. (2007). How the brain laughs. Comparative evidence from behavioral, electrophysiological and neuroimaging studies in human and monkey. *Behavioural Brain Research*, 182(2), 245–60

Minnebusch, D. A, Suchan, B., & Daum, I. (2009). Losing your head: behavioral and electrophysiological effects of body inversion. *Journal of Cognitive Neuroscience*, *21*(5), 865–74. doi:10.1162/jocn.2009.21074

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology*, 11, 219–226.

Moore, C. D., Cohen, M. X., & Ranganath, C. (2006). Neural mechanisms of expert skills in visual working memory. *Journal of Neuroscience*, 26, 11187–11196.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychological Review*, *98*(2), 164–81.

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9*(5), 555–604.

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

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9, 353–383.

Noppeney, U., Price, C. J., Penny, W. D., & Friston, K. J. (2006). Two distinct neural mechanisms for category-selective responses. *Cerebral Cortex*, 16, 437–445.

Oda, R. (1997). Biased face recognition in the prisoner's dilemma games. *Evolution and Human Behavior*, 18, 309-317.

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia*, 41(2), 147-155.

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-Based Trait Inferences and Voting. *Journal of Nonverbal Behavior*. doi: 10.1007/s10919-009-0082-1

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, 46(2), 315–324. doi:10.1016/j.jesp.2009.12.002

Op de Beeck, H. P., Baker, C. I., DiCarlo, J. J., & Kanwisher, N. G. (2006). Discrimination training alters object representations in human extrastriate cortex. *Journal of Neuroscience* , 26(50), 13025–36.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–92. doi:10.1073/pnas.0805664105

Oosterhof, N.N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, 9, 128–133.

Özener, B. (2011). Facial width-to-height ratio in a Turkish population is not sexually dimorphic and is unrelated to aggressive behavior. *Evolution and Human Behavior*, 33(3), 169-173.

Peelen, M. V, & Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, 93(1), 603–8.

Penton-Voak, I.S., Pound, N., Little, A.C., & Perrett, D.I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition*, 24, 607-640.

Perrett, D.I., May, K.A., Yoshikawa, S. (1994). Facial shape and judgments of female attractiveness. *Nature*, 368, 239–242.

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., Henzi, S. P., et al. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, 394, 884–887.

Pillon, A., & d'Honincthun, P. (2011). A common processing system for the concepts of artifacts and actions? Evidence from a case of a disproportionate conceptual impairment for living things. *Cognitive Neuropsychology*, 28, 1–43.

Pitcher, D., Charles, L., Devlin, J. T., Walsh, V., & Duchaine, B. (2009). Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Current Biology*, *19*(4), 319–24. doi:10.1016/j.cub.2009.01.007

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage*, *56*(4), 2356–63. doi:10.1016/j.neuroimage.2011.03.067

Pitcher, D., Garrido, L., Walsh, V., & Duchaine, B. (2008). Transcranial magnetic stimulation disrupts the perception and embodiment of facial expressions. *Journal of Neuroscience*, 28(36), 8929–33.

Pitcher, D., Walsh, V., Yovel, G., & Duchaine, B. (2007). TMS evidence for the involvement of the right occipital face area in early face processing. *Current Biology*, *17*(18), 1568–73. doi:10.1016/j.cub.2007.07.063

Polk, T. A, Park, J., Smith, M. R., & Park, D. C. (2007). Nature versus nurture in ventral visual cortex: a functional magnetic resonance imaging study of twins. *Journal of Neuroscience*, *27*(51), 13921–5. doi:10.1523/JNEUROSCI.4001-07.2007

Pound, N., Penton-Voak, I.S., & Brown, W.M. (2007). Facial symmetry is positively associated with self-reported extraversion. *Personality and Individual Differences,* 43, 1572-1582.

Porter, S., England, L., Juodis, M., Ten Brinke, L., & Wilson, K. (2008). Is the face a window to the soul? Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 40(3), 171–177. doi:10.1037/0008-400X.40.3.171

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, 27(4), 283-296.

Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, 28(5), 340-344.

Quadflieg, S., Todorov, A., Laguesse, R., & Rossion, B. (2012). Normal face-based judgements of social characteristics despite severely impaired holistic face processing. *Visual Cognition*, 20(8), 865–882.

Ravina, E. (2008). Beauty, personal characteristics and trust in credit markets. American Law & Economics Association Annual Meeting, Stanford, retrieved April 1, 2011, from http://law.bepress.com/alea/18th/art67)

Reed, C. L., Stone, V. E., Bozova, S., & Tanaka, J. (2003). The body-inversion effect. *Psychological Science*, *14*(4), 302–8.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable Facial Configurations Affect Strategic Choices in Trust Games with or without Information about Past Behavior. (A. Rustichini, Ed.)*PLoS ONE*, *7*(3), e34293. doi:10.1371/journal.pone.0034293

Rezlescu, C., Pitcher, D., & Duchaine, B. (2012). Acquired prosopagnosia with spared within- class object recognition but impaired recognition of degraded basic-level objects. *Cognitive Neuropsychology*, 29(4), 325–347.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual review of psychology*, 57, 199–226. doi:10.1146/annurev.psych.57.102904.190208

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, 46(18), 2977–87. doi:10.1016/j.visres.2006.03.002

Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., Akamatsu, S. (2001). Attractiveness of facial averageness and symmetry in non-Western cultures: in search of biologically based standards of beauty. *Perception*, 30, 611–625.

Riddoch, M. J., & Humphreys, G. W. (1993). BORB: Birmingham Object Recognition Battery. Hove, UK: Lawrence Erlbaum Associates.

Riddoch, M. J., Johnston, R. A., Bracewell, R. M., Boutsen, L., & Humphreys, G. W. (2008). Are faces special? A case of pure prosopagnosia. *Cognitive Neuropsychology*, 25, 3–26.

Rivest, J., Moscovitch, M., & Black, S. (2009). A com- parative case study of face recognition: The contri- bution of configural and part-based recognition systems, and their interaction. *Neuropsychologia*, 47, 2798–2811.

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, *103*(1), 34–79. doi:10.1016/j.cognition.2006.02.008

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382–439.

Rossion, B., Caldara, R., Seghier, M. L., Schuller, A.-M., Lazeyras, F., & Mayer, E. (2003). A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain*, *126*(Pt 11), 2381–95. doi:10.1093/brain/awg241

Rossion, B., Gauthier, I., Goffaux, V., Tarr, M. J., & Crommelinck, M. (2002). Expertise training with novel objects leads to left-lateralized facelike electrophysiological responses. *Psychological Science*, 13(3), 250-7.

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport*, *11*(1), 69–74.

Rossion, B., & Jacques, C. (2008). Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain? Ten lessons on the N170. *NeuroImage*, *39*(4), 1959–79. doi:10.1016/j.neuroimage.2007.10.011

Rossion, B., Joyce, C. a, Cottrell, G. W., & Tarr, M. J. (2003). Early lateralization and orientation tuning for face, word, and object processing in the visual cortex. *NeuroImage*, *20*(3), 1609–1624. doi:10.1016/j.neuroimage.2003.07.010

Rotshtein, P., Henson, R. N. a, Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, 8(1), 107–113. doi:10.1038/nn1370

Rule, N. O., & Ambady, N. (2008). Brief exposures: Male sexual orientation is accurately perceived at 50ms. *Journal of Experimental Social Psychology*, 44(4), 1100–1105. doi:10.1016/j.jesp.2007.12.001

Rule, N. O., & Ambady, N. (2010). Democrats and republicans can be differentiated from their faces. *PloS ONE,* 5(1), e8733. doi:10.1371/journal.pone.0008733

Rule, N. O., Ambady, N., & Hallett, K. C. (2009). Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *Journal of Experimental Social Psychology*, 45(6), 1245–1251. doi:10.1016/j.jesp.2009.07.010

Rule, N. O., Garrett, J. V, & Ambady, N. (2010). On the perception of religious group membership from faces. *PloS ONE*, 5(12), e14241. doi:10.1371/journal.pone.0014241

Rule, N. O., Macrae, C. N., & Ambady, N. (2009). Ambiguous group membership is extracted automatically from faces. *Psychological Science*, 20(4), 441–3.

Said, C. P., Baron, S., & Todorov, A. (2009). Nonlinear amygdala response to face trustworthiness: Contributions of high and low spatial frequency information. *Journal of Cognitive Neuroscience*, 21(3), 519-528.

Said, C. P., Dotsch, R., & Todorov, A. (2010). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia*, 48(12), 3596–3605. doi:10.1016/j.neuropsychologia.2011.02.028

Said, C. P., Haxby, J. V, & Todorov, A. (2011). Brain systems for assessing the affective value of faces. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 366(1571), 1660–70. doi:10.1098/rstb.2010.0351

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, 9, 260–264.

Sauter, D. A, Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6), 2408–12.

Saxton, T. K., Caryl, P. G., & Roberts, S. C. (2006). Vocal and facial attractiveness judgments of children, adolescents and adults: The ontogeny of mate choice. *Ethology*, 112, 1179–1185.

Scapinello, K. F., & Yarmey, A. D. (1970). The role of familiarity and orientation in immediate and delayed recognition of pictorial stimuli. *Psychonomic Science*, *21*, 329–331.

Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, 22, 617–640.

Schwarzlose, R. F., Baker, C. I., & Kanwisher, N. (2005). Separate face and body selectivity on the fusiform gyrus. *Journal Neuroscience*, *25*(47), 11055–9. doi: 10.1523/JNEUROSCI.2621-05.2005

Schweinberger, S. R., Klos, T., & Sommer, W. (1995). Covert face recognition in prosopagnosia: A dissociable function? *Cortex*, 31, 517–529.

Scott, S. K. (2008). Voice processing in monkey and human brains. *Trends in Cognitive Sciences*, 12(9), 323–5.

Sergent, J., & Signoret, J. L. (1992). Varieties of functional deficits in prosopagnosia. *Cerebral Cortex*, 2, 375–388.

Singular Inversions. (2008). FaceGen Modeller (Version 3.3) [computer software]. Toronto, ON: Singular Inversions.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23, S208.

Smith Micro Software. (2009). Poser (Version 8) [computer software]. Aliso Viejo, CA: Smith Micro Software.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology. Human Learning and Memory*, 6, 174–215.

Steeves, J., Culham, J. C., Duchaine, B., Pratesi, C. C., Valyear, K. F., Schindler, I., Humphrey, G. K., et al. (2006). The fusiform face area is not sufficient for face recognition: evidence from a patient with dense prosopagnosia and no occipital face area. *Neuropsychologia*, *44*(4), 594–609. doi:10.1016/j.neuropsychologia.2005.06.013

Stephen, I. D., Coetzee, V., & Perrett, D. I. (2011). Carotenoid and melanin pigment coloration affect perceived human health. *Evolution and Human Behavior*, 32(3), 216–227. doi:10.1016/j.evolhumbehav.2010.09.003

Stirrat, M., & Perrett, D. I. (2010). Valid Facial Cues to Cooperation and Trust: Male Facial Width and Trustworthiness. *Psychological Science*, 21(3), 349-354.

Stirrat, M., & Perrett, D. I. (2012). Face structure predicts cooperation: men with wider faces are more generous to their in-group when out-group competition is salient. *Psychological Science*, 23(7), 718–22. doi: 10.1177/0956797611435133

Street, R. F. (1931). *A Gestalt completion test*. New York, NY: Teachers College Contributions to Education, Columbia University.

Sturgis, P., Read, S., & Allum, N. (2010). Does intelligence foster gener- alized trust? An empirical test using the UK birth cohort studies. *Intelligence*, 38, 45–54.

Surawski, M. K., & Ossoff, E. P. (2006). The effects of physical and vocal attractiveness on impression formation of politicians. *Current Psychology*, 25(1), 15–27. doi:10.1007/s12144-006-1013-5

Susilo, T., McKone, E., Dennett, H., Darke, H., Palermo, R., Hall, A., . . . & Rhodes, G. (2010). Face recognition impairments despite normal holistic processing and face space coding: Evidence from a case of developmental prosopagnosia. *Cognitive Neuropsychology*, 27, 636–664.

Susilo, T., Yovel, G., Barton, J. J. S., & Duchaine, B. (2013). Face perception is category-specific: Evidence from normal body perception in acquired prosopagnosia. Manuscript submitted for publication.

Swami, V., Furnham, A., & Joshi, K. (2008). The influence of skin tone, hair length, and hair colour on ratings of women's physical attractiveness, health and fertility. *Scandinavian Journal of Psychology*, 49(5), 429-437.

Sydnor, J. R., & Pope, D. G. (2008). What's in a Picture? Evidence of Discrimination from Prosper.com.

Takahashi, N., Kawamura, M., Hirayama, K., Shiota, J., & Isono, O. (1995). Prosopagnosia: A clinical and anatomical study of four patients. *Cortex*, 31(2), 317–329.

Takahashi, C., Yamagishi, T., Tanida, S., Kiyonari, T., & Kanazawa, S. (2006). Sex, attractiveness and cooperation in social exchange. *Evolutionary Psychology*, 4, 315–329.

Tanaka, J. W., & Curran, T. (2001). A neural basis for expert object recognition. *Psychological Science*, 12(1), 43-47.

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

Tarr, M. J. (2003). Visual object recognition: Can a single mechanism suffice? In M. Peterson & G. Rhodes (Eds.), *Perception of Faces, Objects and Scenes: Analytic and Holistic Processes* (pp. 172-211). OUP: London.

Thierry, G., Martin, C. D., Downing, P., & Pegna, A. J. (2007). Controlling for interstimulus perceptual variance abolishes N170 face selectivity. *Nature Neuroscience*, *10*(4), 505–11. doi:10.1038/nn1864

Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, 33(3), 210–216.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social Cognitive and Affective Neuroscience*, 3(2), 119–27. doi:10.1093/scan/nsn009

Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, 25, 395-410.

Todorov, A., & Engell, A. D. (2008). The role of the amygdala in implicit evaluation of emotionally neutral faces. *Social Cognitive and Affective Neuroscience*, 3, 303–312.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308(5728), 1623–6.

Todorov, A., & Olson, I. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social Cognitive and Affective Neuroscience*, 3, 195–203.

Todorov, A., Said, C. P., & Verosky, S. C. (2011). Personality impressions from facial appearance. In A. Calder, J. V. Haxby, M. Johnson, & G. Rhodes (Eds.), *Handbook of Face Perception* (pp. 631-652). Oxford University Press.

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

Todorov, A., Said, C. P., Oosterhof, N. N., & Engell, A. D. (2011). Task-invariant brain responses to the social value of faces. *Journal of Cognitive Neuroscience*, 23(10), 2766–81. doi:10.1162/jocn.2011.21616

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response properties of the human fusiform face area. *Cognitive Neuropsychology*, *17*, 257–279.

Tranel, D., & Hyman, B. T. (1990). Neuropsychological correlates of bilateral amygdala damage. *Archives of Neurology*, 47(3), 349.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311(5761), 670–4. doi:10.1126/science.1119983

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology*, 43(2), 161–204.

Valla, J. M., Ceci, S. J., & Williams, W. M. (2011). The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology*, 5(1), 66–91.

van't Wout, M., & Sanfey, A.G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, 108, 796–803.

Verdonck, A., Gaethofs, M., Carels, C. & de Zegher, F. (1999). Effect of low-dose testosterone treatment on craniofacial growth in boys with delayed puberty. *European Journal of Orthodontics*, 21, 137–143.

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: The sequel. A kernel of truth in predicting cheating detection. *Evolution and Human Behavior*, 28, 260–271.

Viggiano, M. P., Costantini, A., Vannucci, M., & Righi, S. (2004). Hemispheric asymmetry for spatially filtered stimuli belonging to different semantic categories. *Cognitive Brain Research*, 20, 519–524.

Wada, Y., & Yamamoto, T. (2001). Selective impair- ment of facial recognition due to a haematoma restricted to the right fusiform and lateral occipital region. *Journal of Neurology, Neurosurgery and Psychiatry*, 71, 254–257.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. Brain, 107, 829–854. Whiteley, A. M., & Warrigton, E. K. (1977). Prosopagnosia: A clinical, psychological, and anatomical study of three patients. *Journal of Neurology, Neurosurgery & Psychiatry*, 40, 395–403.

Weston, E.M., Friday, A.E., & Liò, P. (2007) Biometric Evidence that Sexual Selection Has Shaped the Hominin Face. *PLoS ONE*, 2(8): e710. doi:10.1371/journal.pone.0000710.

Whiteley, A. M., & Warrigton, E. K. (1977). Prosopagnosia: A clinical, psychological, and ana- tomical study of three patients. *Journal of Neurology, Neurosurgery & Psychiatry*, 40, 395–403.

Wiggins, J. S., Philips, N., & Trapnell, P. (1989). Circular reasoning about interpersonal behavior: Evidence concerning some untested assumptions underlying diagnostic classification. *Journal of Personality and Social Psychology*, 56, 296–305.

Willis, J., & Todorov, A. (2006). First impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science*, 17, 592-598.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 5238–5241.

Wilson, R. K., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly*, 59, 189–202.

Winston, J. S., Henson, R. N. A, Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, 92(3), 1830–9. doi:10.1152/jn.00155.2004

Winston, J. S., O'Doherty, J., Kilner, J. M., Perrett, D. I., & Dolan, R. J. (2007). Brain systems for assessing facial attractiveness, *Neuropsychologia*, 45, 195-206.

Winston, J. S., O'Doherty, J. & Dolan, R.J. (2003) Common and distinct neural responses during direct and incidental processing of multiple facial emotions. *Neuroimage*, 20, 84–97.

Winston, J. S., Strange, B., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, 5(3), 277- 283.

Woolrich, M. W., Ripley, B. D., Brady, M.,&Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *NeuroImage*, 14, 1370–1386.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 102(20), 7398–401. doi:10.1073/pnas.0502399102

Xu, Y., Liu, J., & Kanwisher, N. (2005). The M170 is selective for faces, not for expertise. *Neuropsychologia*, *43*(4), 588–97. doi:10.1016/j.neuropsychologia.2004.07.016

Yang, T.T., Menon, V., Eliez, S., et al. (2002). Amygdalar activation associated with positive and negative facial expressions. *NeuroReport*, 13, 1737–41.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. doi:10.1037/h0027474.

Young, A. W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, 102, 959-974. doi:10.1111/j.2044-8295.2011.02045.x

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747–59.

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*(24), 2256–62. doi:10.1016/j.cub.2005.10.072

Yovel, G., Pelc, T., & Lubetzky, I. (2010). It's all in your head: why is the body inversion effect abolished for headless bodies? *Journal of experimental* *psychology. Human perception and performance*, *36*(3), 759–67. doi:10.1037/a0017451

Yue, X., Bosco, T. S., & Biederman, I. (2006). What makes faces special? *Vision Research*, 46(22), 3802–3811.

Zebrowitz, L. A., Andreoletti, C., Collins, M.A., Lee, S.Y., & Blumenthal, J. (1998). Bright, bad, babyfaced boys: Appearance stereotypes do not always yield self-fulfilling prophecy effects. *Journal of Personality and Social Psychology*, 75, 1300-1320.

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.

Zebrowitz, L.A., & McDonald, S.M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, 15(6), 603–623.

Zebrowitz, L.A., & Montepare, J.M. (2008). Social Psychological Face Perception: Why Appearance Matters. *Social and Personality Psychology Compass*, 2, 1497–1517.

Zebrowitz, L.A., Voinescu, L., & Collins, M.A. (1996). "Wide-eyed" and "crooked-faced": Determinants of perceived and real honesty across the life span. *Personality and Social Psychology Bulletin*, 22, 1258-1269

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (ed.), *Advances in experimental social psychology* (Vol. 14, pp 1-59). New York: Academic Press

Appendix

Questionnaire from Experiment 1, Section 4.1.

University College London
Gower Street
London WC1E 6SSO

Participant's name:

Thank you for completing our interactive investment game. Please answer the following questions.

A. Please indicate the degree to which your personally agree or disagree with each of the following statements by choosing a number from the scale below that most accurately reflects your opinion.

1=strongly disagree ....4=neither agree nor disagree.......................... 7=strongly agree

1. I tend to be accepting of others.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

2. My relationships with others are characterized by trust and acceptance.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

3. Basically I am a trusting person.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

4. It is better to trust people until they prove otherwise than to be suspicious of others until they prove otherwise.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

5. I accept others at 'face value'.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

6. Most people are trustworthy.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

7. It is better to be suspicious of people you have just met, until you know them better.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

8. I make friends easily.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

9. Only a fool would trust most people

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

10. I find it better to accept others for what they say and what they appear to be.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

11. I would admit to being more than a little paranoid about people I meet.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

12. I have few difficulties trusting people.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

University College London
Gower Street
London WC1E 6SSO

Participant's name:

1=strongly disagree ....4=neither agree nor disagree......................... 7=strongly agree

13. Basically I tend to be distrustful of others.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

14. Experience has taught me to be doubtful of others until I know they can be trusted.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

15. I have a lot of faith in people I know.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

16. Even during the 'bad times', I tend to think that things will work out in the end.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

17. I tend to take others at their word.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

18. When it comes to people I know, I am believing and accepting.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

19. I feel I can depend on most people I know.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

20. I almost always believe what people tell me.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

21. I retreat from others.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

22. I am filled with doubts about things.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

23. I feel short-changed in life.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

24. I avoid contacts with others.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

25. I believe that most people would lie to get ahead.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

Participant's name:

26. I find it hard to forgive others.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

27. I believe that people seldom tell you the whole story.

1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐

B. For each of the following statements, please indicate the likelihood that you would engage in the described activity or behavior if you were to find yourself in that situation. Provide a rating from Extremely Unlikely to Extremely Likely, using the given scale.

1. Betting a day's income at the horse races.

1☐	2☐	3☐	4☐	5☐	6☐	7☐
Extremely Unlikely	Moderately Unlikely	Somewhat Unlikely	Not Sure	Somewhat Likely	Moderately Likely	Extremely Likely

2. Investing 10% of your annual income in a moderate growth mutual fund.

1☐	2☐	3☐	4☐	5☐	6☐	7☐
Extremely Unlikely	Moderately Unlikely	Somewhat Unlikely	Not Sure	Somewhat Likely	Moderately Likely	Extremely Likely

3. Betting a day's income at a high-stake poker game.

1☐	2☐	3☐	4☐	5☐	6☐	7☐
Extremely Unlikely	Moderately Unlikely	Somewhat Unlikely	Not Sure	Somewhat Likely	Moderately Likely	Extremely Likely

4. Investing 10% of your annual income in a new business venture.

1☐	2☐	3☐	4☐	5☐	6☐	7☐
Extremely Unlikely	Moderately Unlikely	Somewhat Unlikely	Not Sure	Somewhat Likely	Moderately Likely	Extremely Likely

5. Investing 5% of your annual income in a very speculative stock.

1☐	2☐	3☐	4☐	5☐	6☐	7☐
Extremely Unlikely	Moderately Unlikely	Somewhat Unlikely	Not Sure	Somewhat Likely	Moderately Likely	Extremely Likely

6. Betting a day's income on the outcome of a sporting event.

1☐	2☐	3☐	4☐	5☐	6☐	7☐
Extremely Unlikely	Moderately Unlikely	Somewhat Unlikely	Not Sure	Somewhat Likely	Moderately Likely	Extremely Likely

University College London
Gower Street
London WC1E 6SSO

Participant's name:

C. You are to report how often you performed the following actions:

1. I agreed that I was wrong, even though I wasn't.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

2. I smoked marijuana when everyone else did, even though I didn't want to.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

3. I walked out of the store knowing that I'd been short-changed.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

4. I accepted verbal abuse without defending myself.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

5. I made love with my partner when I didn't want to.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

6. When I stood to speak and the others continued talking, I simply sat down.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

7. I continued to apologize for the minor mistake.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

8. I drank a lot at the party when the others gave me a hard time.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

Participant's name:

9. I listened quietly when my parents said that my hair was ugly.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

10. I was not able to tell my friend that I was angry with her.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

11. At the meeting, I let others monopolize the conversation.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

12. I wept when I couldn't solve the simple problem.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

13. I said "thank-you" enthusiastically and repeatedly when someone did me an insignificant favour.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

14. I avoided direct eye contact when the shop clerk spoke to me.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

15. I did not start a single conversation at the party.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

16. Although my friends thought my partner had humiliated me, I date him/her again.

1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐
Never	Once	2-3 times	From time to time	Regularly	Quite often	Very often

Participant's name:

17. I blushed when he stared at me.

1 ☐ Never | 2 ☐ Once | 3 ☐ 2-3 times | 4 ☐ From time to time | 5 ☐ Regularly | 6 ☐ Quite often | 7 ☐ Very often

18. I pretended I was ill when declining the invitation to the party.

1 ☐ Never | 2 ☐ Once | 3 ☐ 2-3 times | 4 ☐ From time to time | 5 ☐ Regularly | 6 ☐ Quite often | 7 ☐ Very often

D. Indicate the degree to which your personally agree or disagree with each of the following statements by choosing a number from the scale below that most accurately reflects your opinion.

1=strongly disagree .... 4=neither agree nor disagree......................... 7=strongly agree

1. I believe I am more attractive than the average person my gender.

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐

2. A socially dominant person tells other people what to do, is respected, influential, and often a leader; whereas a submissive person is not influential or assertive and is usually directed by others. According to this description, I am more dominant than an average person.

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐

3. If I got in a fistfight with an average person of my age and gender, I would probably win.

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐

Please tell us about yourself:

Age: ☐ Gender: M ☐ F ☐

Highest level of education completed:

primary school ☐ secondary or high school ☐ technical or vocational school ☐

other college ☐ graduate school ☐ postgraduate or professional degree ☐

Occupation: ☐

That was all. THANK YOU!