Individual Differences in Face Processing



Mirta Stantić

St Catherine's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Individual Differences in Face Processing

Mirta Stantić

St Catherine's College, University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Supervisor: Geoffrey Bird, PhD

Acknowledgments

When this journey got hard and I couldn't see the path ahead, I was very lucky to have many people remind me that what we do boils down to thinking and reading and having coffee with clever young people and that, despite the occasional hardship, that is an incredible privilege.

Geoff Bird is the supervisor I needed. At every step of this journey, he encouraged me to ask difficult questions and bite off more than it seemed I should chew. Our supervisory meetings were neither scheduled nor structured. Always without a set agenda they yielded, over many hours and lots of coffee and cake, some of the most interesting questions I've ever asked. But mostly, Geoff supported my career wholeheartedly even when things were messy. When life clashed with work in ways that required flexibility and adjustments, he clearly and loudly advocated life taking a front seat, for which I am forever grateful.

Caroline Catmur demonstrated by example that you can think hard about statistical methods and teach with humanity, juggle too many balls and know which are allowed to drop and be a supportive supervisor even when life is busy. I am very lucky that I got to watch her in action and learn from her.

I have had the privilege of calling a phenomenal lab mine over the past four years. From day one ("the day without computers"), I've shared this journey with Eri Ichijo and Hélio Cuve. While collaborating on interesting projects, they also debugged code, shared funny tweets, contributed to our not-so-secret stash of chocolate biscuits, ordered late-night

Chinese food, and sipped tequila on Fridays. I am only sorry the pandemic cut our camaraderie short.

Rebecca Brewer and Jenny Murphy held me up as a scientist and a person, always equally happy to talk experiments, house renovations, counterbalancing conditions, and babies. Leora Sevi afforded me the phenomenal joy of watching an advisee become a colleague.

I am so grateful to collaborators who helped sharpen my work and my thinking. Brad Duchaine, in particular, has supported me in ways small and large since we first met. Kindness like his can seem rare in academia, so I hope to grow into a scientist who lives with the same grace. I have benefited greatly from the expertise and collegiality of Tirta Susilo and Sarah Bate, who generously shared their patient databases with me.

I have enjoyed working with so many smart undergraduate students over the years. Tinkering with problems alongside them made my own thinking clearer. I am grateful to Samuel Day, Amani Krayem, Rebecca Williams, Wangjingyi Liao, Charlotte Peart, Bethan Hearne, Katie Brown, Isabel de Castro, Iva Galić, Vito Bošnjak and Ana (Miličević) Bošnjak for help with recruitment, data collection, and analysis. Thank you for your assistance on my projects, and for trusting me to help shape your own.

Funding from the Economic Social Research Council, St Catherine's College, and the Frankopan Fund has been instrumental to my time at Oxford. Thank you for encouraging me to focuse my time on research and helping many collaborations thrive.

My research would not be possible without the generous contribution of time from our participants. I am grateful to all of them for their help, but particularly to our autistic participants and their families, as well as to all developmental prosopagnosics and super recognizers who took part in my studies. I have enjoyed our interactions tremendously.

Ivana and Gea were there for many wonderful moments, and for all the hardest ones. They know that to get here took some ugly days – and on every single one of them they cheered me on, genuinely, relentlessly, loudly. Finishing big chapters in life can sometimes resemble magic, but those who have seen the mountains of work behind the curtains know that it is anything but.

My parents, Anita and Goran, ever grateful that this move brought me thousands of miles closer to them, unfailingly offered up their time so we could vent and brainstorm and troubleshoot, even when they did not understand why I was stressed. Despite my stubbornness to admit it, their advice nearly always helped. Being here is a tribute to growing up in our house, always overflowing with books, and people who took the time to answer endless questions from curious little minds. I have loved getting to hang out with the little minds who grew up with me – Ljudevit and Vedrana – so much more often during these years; and our gang is endlessly better for having Lucija, a steady and supportive presence on every journey.

But choosing to share this life with Igor remains the single best decision I will ever make.

Words will forever fall short of expressing my grace – so let me say simply that this would not have happened without him. I owe a mountain of gratitude – for helping with the technical parts of my work so I could focus on asking the interesting questions, for every midnight math discussion, for model sketches on napkins and mechanisms scribbled on backs of coffee shop receipts, for unwavering belief in my mind's ability to push, for reminding me of why I started when I ran out of fuel, for genuinely wondering about questions I found fascinating, for doing whatever needed doing so I could do what I do best, and for loving me even when I was tired and sleep-deprived and overwhelmed and unkind. Despite all the curveballs of the last few years, you lit up a path to this place that I

could not have reached alone. You reminded me continually that there are no shortcuts to any places worth going. Thank you for choosing this madness – me, two dozen boxes on a ship across the Atlantic, and our growing family – every day.

And to Alina, the light of our lives. Without you, this dissertation would have been finished much sooner, but in a world filled with so much less joy. Your face is the centre of my whole universe.

Declaration

This thesis is submitted to University of Oxford in support of Mirta Stantić's application for the degree of Doctor of Philosophy. It has been composed by Mirta Stantić and has not been submitted in any previous application for any degree.

This thesis is composed of partial copies of published papers and papers currently under review at peer-reviewed journals. Mirta Stantić is the lead author on all of these papers and authored the figures and text reproduced here. Mirta Stantić led the design, conceptualisation, recruitment, data collection, analyses, data visualisation, interpretation and writing of all the work that contributed to this thesis. This work was formatted specifically for inclusion in this thesis. Authors' contributions are included below for each individual chapter. All co-authors on any published papers have given approval for this material to be submitted as part of this dissertation.

As well as the introduction and discussion sections that are specific to each article, this thesis includes a general introduction, providing an overview of the literature relevant to the experimental chapters, and a general discussion of the main findings, interpretations, limitations and outstanding questions.

Authors' contributions

Chapter 2 (Development and validation of the OFMT):

Study design for experiments in both papers was conducted by Geoffrey Bird and Mirta Stantić. Mirta Stantić created the stimuli and programmed the experimental paradigm, collected the data (for Experiments 1, 2, and 3, Mirta Stantić collected the data; for Experiments 4 and 5 data collection was done in collaboration with undergraduate students Akhina Gaches, Jacob Knyspel and Yinning Liu). Mirta Stantić ran the analyses and led interpretation and visualization of results. Mirta Stantić drafted the manuscript, which was then edited by co-authors Sarah Bate, Rebecca Brewer, Tirta Susilo, Bradley Duchaine, Michael Banissy and Caroline Catmur. The final manuscript was edited by Geoffrey Bird and Mirta Stantić.

Chapter 3 (Individual differences in face perception and memory in the typical population):

Study design was conducted by Geoffrey Bird and Mirta Stantić. Mirta Stantić created the stimuli and programmed the experimental paradigm, collected the data (for Experiment 2 reported in this chapter) and ran the analyses. Mirta Stantić interpreted and visualized the results and drafted the manuscript, which was then edited by co-author Caroline Catmur. The final manuscript was edited by Geoffrey Bird and Mirta Stantić.

Chapter 4 (Face perception and memory in developmental prosopagnosia):

Study design was conducted by Geoffrey Bird and Mirta Stantić. Mirta Stantić created the stimuli and programmed the experimental paradigm, collected the data and ran the analyses, interpreted and visualized the results. Mirta Stantić drafted the manuscript, which

was then edited by co-authors Zoe Pounder and Caroline Catmur. Co-authors Sarah Bate and Tirta Susilo advised on recruitment and shared participants databases, as well as edited the manuscript draft. The final manuscript was edited by Geoffrey Bird and Mirta Stantić.

Chapter 5 (Face perception and memory in super recognition):

Study design was conducted by Geoffrey Bird and Mirta Stantić. Mirta Stantić created the stimuli and programmed the experimental paradigm. Mirta Stantić collected and pre-processed the data and ran the analyses with help from Zoe Pounder. Mirta Stantić interpreted and visualized the results and drafted the manuscript, which was then edited by co-authors Zoe Pounder and Caroline Catmur. Co-author Sarah Bate shared a participant database and reviewed the manuscript draft. The final manuscript was edited by Geoffrey Bird and Mirta Stantić.

Chapter 6 (Face perception and memory in autism):

Study design was conducted by Geoffrey Bird and Mirta Stantić. Mirta Stantić created the stimuli and programmed the experimental paradigm, collected the data for Experiment 1 reported here in collaboration with an undergraduate student (Katie Brown) and ran the analyses. Mirta Stantić drafted the manuscript, which was then edited by co-authors Eri Ichijo, Katie Browne, and Caroline Catmur. The final manuscript was edited by Geoffrey Bird and Mirta Stantić.

Chapter 7 (Effects of age and gender on face perception and memory):

Study design was conducted by Geoffrey Bird and Mirta Stantić. Mirta Stantić created the stimuli and programmed the experimental paradigm, collected the data in collaboration with an undergraduate student (Bethan Hearne) and ran the analyses. Mirta Stantić

interpreted and visualized the data, drafted the manuscript, which was then edited by co-authors Caroline Catmur and Bethan Hearne. The final manuscript was edited by Geoffrey Bird and Mirta Stantić.

List of chapters and associated papers

Chapter 2: Development and validation of the OFMT

- Stantić, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., Catmur, C. & Bird, G. (2021). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavior research methods*, 1-16. https://doi.org/10.3758/s13428-021-01609-2
- Stantić, M., Knyspel, J., Gaches, A., Liu, Y., Bird, G.P., & Catmur, C. (under review) The Oxford Face Matching Test: Short Form Alternative

Chapter 3: Individual differences in face perception and memory in the typical population

- Stantić, M., Brown, K. Catmur, C. & Bird, G. (under review). Independent Measurement of Face Perception, Face Matching, and Face Memory Reveals Impairments in Face Perception and Memory, but Not Matching, in Autism.

Chapter 4: Face perception and memory in developmental prosopagnosia

- Stantić, M., Pounder, Z., Bate, S., Susilo, T., Catmur, C., Bird, G.P. Individuals with developmental prosopagnosia show independent impairments in face perception, face memory and face matching. *Cortex 157*, 266-273. https://doi.org/10.1016/j.cortex.2022.09.012

- Stantić, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., Catmur, C., & Bird, G. (2022). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavior research methods*, 54(1), 158-173. https://doi.org/10.3758/s13428-021-01609-2

Chapter 5: Face perception and memory in super recognition

- Stantić, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., Catmur, C. & Bird, G. (2021). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavior research methods*, 1-16. https://doi.org/10.3758/s13428-021-01609-2
- Stantić, M., Pounder, Z., Bate, S., Catmur, C., & Bird, G. (in prep) Individuals who are 'super-recognisers' show superior performance on independent measures of face perception, face memory and face matching.

Chapter 6: Face perception and memory in autism

- Stantić, M., Ichijo, E., Catmur, C., & Bird, G. (2021). Face memory and face perception in autism. *Autism*, *26*(1), 276–280. https://doi.org/10.1177/13623613211027685
- Stantić, M., Brown, K., Ichijo, E., Catmur, C. & Bird, G. (under review). Independent Measurement of Face Perception, Face Matching, and Face Memory Reveals Impairments in Face Perception and Memory, but Not Matching, in Autism.

Chapter 7: Effects of age and gender on face perception and memory

- Stantić, M., Hearne, B., Catmur, C., & Bird, G. (2021). Use of the Oxford face matching test reveals an effect of ageing on face perception but not face memory. *Cortex*, 145, 226-235. https://doi.org/10.1016/j.cortex.2021.08.016

Abbreviations

ADOS – Autism Diagnostic Observation Schedule.

AQ-50 or AQ-28 – Autism Spectrum Quotient Questionnaire (50 or 28 items)

ASD – Autism Spectrum Disorder

WASI – Weschler Abbreviated Scale of Intelligence

VOTC – ventral occipito temporal cortex

AP – acquired prosopagnosia or acquired prosopagnosic

DP – developmental prosopagnosia or developmental prosopagnosic

SR – super recognition or super recogniser

CFMT – Cambridge Face Memory Test

CFPT – Cambridge Face Perception Test

OFMT – Oxford Face Matching Test

GFMT – Glasgow Face Matching Test

FFT – Famous Faces Test

MMT – Models Memory Test

KFMT – Kent Face Matching Test

PI-20 – 20-Item Prosopagnosia Index

Table of Contents

Acknowledgments ***2***

Declaration ***6***

Authors' contributions ***7***

List of chapters and associated papers ***9***

Abbreviations ***11***

List of Figures ***15***

List of Tables ***16***

Abstract ***17***

Chapter 1: Introduction ***19***

1.1. Background **19**

Importance of face perception 19

Theoretical frameworks of face recognition 20

A wide range of face recognition abilities 24

Prosopagnosia 25

Autism 29

Elderly adults and dissociation between memory and perception 32

Super recognition 33

Individual differences in the 'normal' range 35

Issues with existing tests 38

1.2. Conclusion/objectives **44**

1.3. Guiding research questions **45**

Chapter 2: The development and validation of the Oxford Face Matching Test ***48***

2.1. Abstract **49**

2.2. Introduction **50**

2.3. Common methodology **54**

2.3.1. Tasks 54

2.3.2. Exclusion criteria 56

2.4. Experiment 1 ..... **57**

2.4.1 Methods 57

2.4.2. Results 62

2.4.3. Discussion 64

2.5. Experiment 2 ..... **65**

2.5.1. Methods 65

2.5.2. Results 65

2.5.3. Analyses across Experiment 1 and Experiment 2 67

2.5.4. Discussion 68

2.6. Experiment 3 ..... **69**

2.6.1. Methods 70

2.6.2. Results 70

2.6.3. Discussion 72

2.7. Experiment 4 ..... **73**

2.7.1. Methods 73
2.7.2. Results 74

2.8. Experiment 5 75
2.8.1. Methods 75
2.8.2. Results 76

2.9. General Discussion 77

Chapter 3: Individual Differences in Face Matching and Memory in Neurotypical Population 83

3.1. Abstract 84

3. 2. Introduction 85

3.3. Methods 87
3.3.1. Procedure 87
3.3.2. Participants 87
3.3.3. Analysis strategy 88

3.4. Results 89
3.4.1. Group differences 89
3.4.2. Effect of Face Perception on Face Matching 91
3.4.2. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory .. 91

3.5. Discussion 93

Chapter 4: Face perception and memory in developmental prosopagnosia 96

4.1. Abstract 97

4.2. Introduction 98

4.3. Methods 102
4.3.1. Participants 102
4.3.2. Procedure 102
4.3.3. Analysis strategy 103

4.4. Results 103
4.4.1. Group differences 103
4.4.2. Effect of Face Perception on Face Matching 108
4.4.3. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory 109

4.5. Discussion 110

Chapter 5: Face perception and memory in super recognition 115

5.1. Abstract 116

5.2. Introduction 117

5.3. Methods 119
5.3.1. Participants 119
5.3.2. Procedure 120
5.3.3. Analysis Strategy 120

5.4. Results 120
5.4.1. Group differences 120
5.4.2. Effect of Face Perception on Face Matching 125
5.4.3. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory 126

5.5. Discussion 127

Chapter 6: Face perception and memory in Autism Spectrum Disorder 130

6.1. Abstract **131**

6.2. Introduction **132**

6.3. Methods **133**
6.3.1. Procedure 133
6.3.2. Analysis Strategy 133
6.3.3. Participants 134

6.4. Results **135**
6.4.1. Group Differences 135
6.4.2. Effect of Face Perception on Face Matching 138
6.4.3. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory 139

6.5. Discussion **140**

Chapter 7: Effects of age and gender on face perception and memory ***143***

7.1. Abstract **144**

7.2. Introduction **145**

7.3. Methods **147**
7.3.1. Participants 147
7.3.2. Tasks 148
7.3.3. Procedure 149
7.3.4. Analysis Strategy 149

7.4. Results **149**
7.4.1. Effect of Age on Face Perception 150
7.4.2. Effect of Age on Face Memory 151
7.4.3. Effect of Age on Self-Reported Difficulties with Face Recognition 152
7.4.5. Effect of Face Perception on Face Matching 154
7.4.6. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory 154
7.4.7. Effect of Gender – All Measures 155
7.4.5. Relationships Between OFMT and Other Face Processing Measures 157

7.5. Discussion **159**

Chapter 8: General Discussion ***164***

8.1. Development of better tests **165**
8.1.1. Findings and implications 166
8.1.2. Limitations and future directions 167

8.2. Towards understanding individual differences in typical groups **170**
8.2.1. Findings and implications 170
8.2.2. Limitations and future directions 174

8.3. Towards understanding individual differences in atypical groups **175**
8.3.1. Findings and implications 175
8.3.2. Future direction and theoretical updates 179

8.4. General limitations **183**
8.4.1. Terminology and diagnostic criteria 183
8.4.2. Sample 185
8.4.3. Stimuli 186

8.5. Future directions **187**

References ***195***

List of Figures

Figure 2.1. Sample trials of the face processing tasks. 56
Figure 2.2. A sample trial from the Oxford Face Matching Task – Long Form. 60
Figure 2.3. Relationship between participants' similarity judgments and average algorithmic similarity values. 63
Figure 2.4. Distribution of scores on the Short Form OFMT and their relationship to Long Form OFMT. 64
Figure 2.5. Relationship between Long-Form and Short-Form OFMT accuracy. 66
Figure 2.6. Relationship between participants' similarity judgments and average algorithmic similarity values. 67
Figure 2.7. Variance in performance by bin quintile. 68
Figure 2.8. Performance on the two versions of the test (with and without similarity judgements). 75
Figure 3.9. Correlations between all measures of face processing, and distributions of scores for individual tasks. 90
Figure 3.10. Linear regressions showing relationships between the deviation of individual participants' similarity judgments and the face matching and memory tasks. 93
Figure 4.11. Group comparisons for all face processing tests. 107
Figure 5.12. Group comparisons for all processing tasks. 124
Figure 6.13. Group differences for all face processing tasks. 137
Figure 7.14. Performance on all face perception measures shown as a function of age with the best fit line for the model that explains most variance. 153
Figure 8.15. A suggested update to the existing models of face recognition subprocesses. 173
Figure 8.16. Theoretical comparison of 3-D face space representations in typical and atypical participants. 182

List of Tables

Table 2.1. Relationships between all face processing measures included in Experiment 3. 72
Table 2.2. Performance on face processing measures in Experiment 5. 76
Table 3.3. Descriptive statistics for all three face perception measures. 90
Table 4.4. Relationship between all face measures across all participants. 108
Table 4.5. Relationships between all tasks separated by group (Developmental Prosopagnosia or Control). 108
Table 5.6. Relationship between all face measures across all participants. 125
Table 5.7. Relationships between all tasks separated by group (Super recognizers or Control). 125
Table 6.8. Relationships between all tasks separated by group (Autism or Control). 138
Table 7.9. Descriptive statistics for all participants across all face perception measures. 156
Table 7.10. Relationships between all face perception measures for all participants. 157
Table 7.11. Distribution of OFMT scores by age and gender. 158

Abstract

Human ability to recognize identities from faces is a fundamental cognitive skill that underlies our functioning in the social world. This ability varies widely in the general population, with the ends of the spectrum of abilities characterized by 'atypical' populations (people with developmental prosopagnosia, autism, and super recognition). Different theoretical models have been suggested to explain this variability, suggesting atypical visual perception (e.g., scanning patterns) or information integration (e.g., holistic processing). However, experimental evidence for these theories has been mixed, with most models inadequately accounting for individual differences. This thesis seeks to improve our approach to measuring face perception to advance our understanding of face recognition subprocesses.

Chapter 1 reviews the current conceptualization of atypical groups and outlines the theoretical framework for my research. Key hypotheses and mechanisms are outlined to improve measurement of identity processing across abilities. In particular, I focus on tests of face perception and memory and their shortcomings in the context of measuring individual differences, and outlined a novel way of developing face processing tests relying on facial recognition algorithms.

Chapter 2 outlined the process of developing and validating a novel measure of face perception, the Oxford Face Matching Test (OFMT). It revealed that the novel measure captures a wide range of performance in typical groups and is stable over time. **Chapter 3** explored how the OFMT can be utilized in person and online to measure face recognition processes in a large sample of diverse neurotypical participants. Results confirmed the hypothesized contributions of face perception and face matching to face memory.

Chapters 4, 5, and 6 explored whether administration of the OFMT gives sufficiently sensitive scores at the ends of the spectrum in large groups of developmental prosopagnosics (**Chapter 4**), super recognisers (**Chapter 5**) and people with autism (**Chapter 6**). These chapter investigated the relationship between independently measured subprocesses necessary for successful face recognition. Findings suggested different patterns of atypicality, with super recognizers showing superior performance on all subprocesses, and people with autism and developmental prosopagnosia showing different patterns of impairments with face perception, matching and memory.

Chapter 7 looked at ageing-related changes to face perception and memory abilities in a large sample of healthy participants across a range of ages. Results revealed effects of ageing on face matching tests, but importantly no ageing-related effects on face memory when face perception was measured precisely using the more sensitive OFMT.

In **Chapter 8**, I discuss the implications of this body of work for updating theoretical models of face recognition and applying novel methods to the development of future face processing measures, as well as their application in real-life scenarios of training and automation of security processes. This body of work overhauls the methodological approach to measuring individual differences in face perception and offers a novel approach to independently capturing subprocesses necessary for face recognition. These methods hold promise for answering long-standing questions of the relationships between these subprocesses and the resulting explanatory mechanisms of atypical perception.

Chapter 1: Introduction

1.1. Background

Importance of face perception

The ability to recognize people around us from their faces is a fundamental cognitive capacity that underlies our easy, seamless functioning in the social world. It allows us to tell strangers from friends, lets us determine the context of their familiarity, and recall semantic information about them. Healthy humans extract a great deal of information about a face at a glance, from recognizing sex, ethnicity and age, to quickly perceiving emotion and determining the attractiveness of a person. A great deal of consequential decisions are made based on these inferences, including who we vote for (Todorov, Mandisodza, Goren, & Hall, 2005; Olivola & Todorov, 2010, Little, Burris, Jones, & Roberts, 2007; Ballew & Todorov, 2007; Sussman, Petkova, & Todorov, 2013), who we convict of crimes (Funk & Todorov, 2013; Blair, Judd & Clapleau, 2004), or who we promote to leadership positions (Rule & Ambady, 2008; Rule & Ambady, 2009; Mueller & Mazur, 1996), amongst others. This ability to extract socially relevant information from faces in a very short period of time, therefore, has a profound impact on our behaviours in daily life. It is for this reason that attempts to understand the mechanisms that underlie face processing and face recognition have remained core aims in social psychology in the last few decades.

Of fundamental importance to this thesis is that, even in healthy adults, the ability to recognise faces varies substantially, even in the presence of broad trends characterising typical face recognition. While everyone is better able to recognize familiar faces than

unfamiliar ones (Hancock, Bruce & Burton, 2000; Bruce et al., 2001; Megreya & Burton, 2006; Megreya & Burton, 2007), and faces of own ethnicity and age than those of other ethnicities and ages (Lindsay, Jack, & Christian, 1991; Valentine & Bruce, 1986; Ellis & Deregowski, 1981), differences between healthy people at each of these abilities are vast. Attempts to understand the individual differences in face recognition performance have produced some theories about the mechanisms that underlie this ability, and the aspects of those mechanisms that could be driving the wide individual differences.

Theoretical frameworks of face recognition

Given that face recognition has been one of the most heavily studied abilities in social psychology, and researchers consistently observe wide variance between people in the general population, it is not surprising that a number of theories have been proposed to explain the cognitive processing that underlies it.

It is widely accepted that there are multiple steps to face recognition, from detecting a face in a visually crowded environment to forming a stable cognitive representation of its identity. Even the simplest consideration of the problem recognizes that a number of processes have to be involved - such as recognizing that something is a face (compared to a non-face object), recognizing the unstable or changeable features of the face (e.g., hairstyle or emotional expression), recognizing the stable or unchangeable features of the face (e.g., ethnicity, gender, eye-colour), and finally recognizing the identity of a face. Failure at any of these steps would impair an individual's normal social functioning and would result in a variety of impairment profiles, from an inability to detect that something is a face, an inability to recognize stable features of a face (such as ethnicity or gender),

failure to recognise facially-signalled states such as emotions, or an inability to recognize the identity of a person in front of you.

Consensus on what steps are actually involved in face recognition has been reached by converging cognitive (Bruce & Young, 1986) and neural evidence (Haxby, Hoffman, & Gobbini, 2000) in a hierarchical model. These suggest that every step of the face recognition process is related to activity in a particular dedicated brain region, mostly in the ventral occipito-temporal cortex (VOTC). Following face detection, early perception of face-like features occurs in the inferior occipital gyri, and forms a percept of features that is view-and expression-dependent. Two separate streams of processing then extract further information. Lateral fusiform gyrus encodes unchangeable information that can be used for recognition of a specific identity (e.g., eye colour), while superior temporal sulcus encodes information about the changeable aspects of the face (e.g., expression-related movement of the eyes or the lips). Finally, an extended set of brain areas conducts a separate set of cognitive steps, generally thought to involve semantic recall (memory including information about the identity of a person) or social cognition (emotional response towards the person) – these occur in the anterior temporal cortex, intraparietal sulcus, and the amygdala (Gobbini & Haxby, 2007).

More recently, an alternate theoretical framework of the division of cognitive and neural steps involved in building a representation of facial identity been suggested (Duchaine & Yovel, 2015). This model proposes that the core areas in traditional models represent a 'ventral pathway', which is instead responsible for processing structural information about faces and allows for representation of identity, as well as basic features like emotional expression, age, sex, and ethnicity. An extended set of regions in the 'dorsal pathway',

including anterior and posterior superior temporal sulcus, fusiform area, and inferior frontal gyrus, in a separate set of cognitive steps, are suggested to encode those aspects of facial structure that can change – things like expression, eye-gaze, movements of the mouth, etc.

These models explain the consistent differences in ability to recognize familiar and unfamiliar faces, as the mechanism of recognizing a familiar face is thought to be different than that of recognizing unfamiliar faces. This difference in processing is ascribed to the difference in how the information is sampled from various regions of the face, relating specifically to the encoding of changeable and unchangeable aspects of the face in the fusiform gyrus and superior temporal sulcus. Namely, familiar face recognition seems to benefit from information being gathered from the internal facial features (e.g., eyes, nose, eyebrows, mouth), while processing of unfamiliar faces seems to rely more heavily on sampling of external features (e.g., shape of the face, hairstyle; (Ellis, Shepherd & Davies, 1979). The qualitative difference between processing of familiar and unfamiliar faces necessarily orients researchers to focus on one or the other when examining the overall mechanisms of face recognition and developing tests to measure this ability. Thus, while some debate remains on the actual steps involved in facial recognition, in the order they occur, and in the neural areas involved, researchers agree that recognition of an identity from a face rests on one's ability to form a stable representation of another person's face.

One of the most widely accepted theoretical frameworks of how we organize these representations of faces is norm-based coding. This idea posits that humans use a multidimensional map of faces, known as the face space, within which each face is represented as a point (Valentine, 1991; Webster & MacLeod, 2011). For sake of

illustration, one might imagine that a two-dimensional face space could be defined by two axes, each representing a facial feature distance (e.g., the distance between a person's eyes on the x-axis and the distance between their nose and mouth on the y-axis). Each face could then be represented in this two-dimensional space as a single point, with faces that are more similar to each other located closer to each other in space than faces that are more dissimilar from each other.

In reality, the visual system uses more dimensions to represent faces and the nature of these dimensions remains somewhat elusive. However, the dimensions of this multidimensional face space need not necessarily be known, or even universal, as long as the centre of the space is defined by an average of all faces that are encountered by an individual. In this multidimensional space defined by unknown dimensions, each face can then be represented as a point defined by its distance to the perceiver's average face. The distinctiveness of each face then defines its distance from the average alongside any number of dimensions that define the space. Faces that are very distinct from the average would, therefore, be placed far from the average and judged as atypical, whereas faces that are less distinct would be placed closer to the average and judged as more typical. This type of organization allows not only for the judgment of distinctiveness/typicality at the level of each face, but also for any two faces to be judged in similarity to each other. The closer the faces are in this multidimensional space, the more similar they should appear to the perceiver. Crucially, the way that the face space is shaped and organized is unique to the perceiver, though some research has pointed out that even people with severe difficulties in face perception show representations of faces that are similar to those of neurotypical people (Nishimura, Doyle, Humphreys, & Behrmann, 2010; Mondloch, LeGrand, & Maurer, 2002). This suggests perhaps that salient cues about distinctiveness of

a face are used by all in building face representations, but difficulties in other stages of processing prevent those with difficulties to process more standard-looking faces or those that appear more alike.

A wide range of face recognition abilities

A wide range of abilities in face recognition has been observed, from those who struggle with face perception, like developmental or acquired prosopagnosics (Kress & Daum, 2003; Behrman & Avidan, 2005; Duchaine & Nakayama, 2006a; Susilo & Duchaine, 2013; Bodamer, 1974; Hecaen & Angelergues, 1962; de Gelder & Rouw, 2000; Schiltz et al., 2006; Rossion et al., 2003; Sorger, Goebel, Schiltz, & Rossion, 2007; Barton, 2008) people with autism (Dawson et al., 2005; Schultz, 2005; Golerai, Grill-Spector, & Reiss, 2006; Boucher & Lewis, 1992; Davies et al., 1994; Faja et al., 2009), and elderly adults (Duchaine, Germine & Nakayama, 2007; Shah et al., 2015; Megreya & Bindemann, 2015; Habak, Wilkinson, & Wilson, 2008), to super recognisers, who show superior face perception (Russell et al., 2009; Ramon, 2021; Nador et al., 2021; Bate et al., 2019; de Haas, 2022; Cohan, Nakayama, & Duchaine, 2016). While most research has focused on studying the extremes of the performance spectrum, such as prosopagnosics and, in the last decade, super recognizers, recent work has suggested the need to focus on individual differences in the entire population (for review, see Wilmer, 2017). Next, I review the published literature examining the performance at the ends of the spectrum and testing batteries that underlie this work before looking into existing findings in the typical population.

Prosopagnosia

Developmental prosopagnosia (DP) is defined as a neurodevelopmental disorder associated with severe difficulties in recognition of familiar faces and learning of novel identities from unfamiliar faces. First defined in the second part of the 19th century (McConachie, 1976), it occurs in the absence of any other issues with intelligence, vision, or memory (Kress & Daum, 2003; Behrman & Avidan, 2005; Duchaine & Nakayama, 2006a; Susilo & Duchaine, 2013). It also appears in the absence of any known neurological lesions, which is a characteristic of acquired prosopagnosia (AP), a condition which is defined by severe impairments in face perception arising following neurological damage, most commonly to the VOTC (Bodamer, 1974; Hecaen & Angelergues, 1962).

Regardless of the type of prosopagnosia an individual has, the diagnosis is related to a number of problems in sociocognitive functioning. Prosopagnosics tend to avoid social situations as their inability to recognize people they have previously met before makes interactions difficult and oftentimes embarrassing, and as a result of this they often suffer from social anxiety or depression (Yardley et al., 2008; Duchaine & Nakayama, 2006b; Dalrymple et al., 2014). A variety of compensatory strategies are deployed by prosopagnosics to lessen this deficit, such as relying on information about voice, gait, or individual distinctive facial features to recognize a person (Cook & Biotti, 2016; Shah et al., 2015), though these are often insufficient to make up for their difficulty in everyday social interactions, which tend to require very quick reactions. Attempts to 'train' developmental prosopagnosics to perform better on various tasks or improve face perception in everyday life have so far shown limited efficacy (DeGutis, Cohan, & Nakayama, 2014), in line with other research showing that improvements in face recognition performance are time-constrained and limited in scope (lasting for a short

period of time and not transferring to non-trained identities) even in the typical population (Woodhead, Baddeley & Simmonds, 1979; Towler, White, Kemp, 2014; Alenezi & Bindemann, 2013).

Reaching a diagnosis of DP is not straightforward, as different research laboratories use different criteria and cut-offs since DP is not defined as a mental health disorder in the latest Diagnostic and Statistical Manual for Mental Disorders, DSM-V (American Psychological Association, 2013). While the agreement in the field on what measures should be used for diagnosing DP has not yet been reached, the issue has been debated in detail (Dalrymple & Palermo, 2016; Barton & Corrow, 2016) and recent studies tend to use more consistent criteria compared to older work. Most research laboratories nowadays use a combination of self-report measures or clinical interviews and objective computer-based tests of face recognition to diagnose prosopagnosia.

Two types of objective computer-based measures tend to be used in prosopagnosia testing – those that test the ability to recognize unfamiliar (unknown) faces, as well as those that test the ability to recognize familiar (known, most commonly famous) faces. As mentioned previously, these two types of face processing have been argued to be distinct (see Johnson & Edmonds, 2009 for review), with familiar face processing relying more heavily on internal facial features and occurring without detriment even in poor viewing conditions, and unfamiliar face processing requiring more of a reliance on the external facial features (Ellis, Shepherd & Davies, 1979 though see Abudardham, Shkiller & Yovel, 2018). Therefore, separate tests have been developed to measure the two abilities.

Two tests utilizing unfamiliar faces are used most commonly in diagnosis of DP as gold-standard measures. By far the most commonly used test to diagnose DP is the Cambridge Face Memory Test (CFMT, Duchaine & Nakayama, 2006a), which was designed specifically with this purpose in mind. While completing the CFMT, participants learn six target faces at the beginning of the test, after which they are tested on three-alternative forced-choice trials. On each trial, two images are distractors and one is an image of a learned target identity with participants being required to identify the learned identity. The test is divided into three stages of increasing difficulty, involving 18 test trials with no change of viewpoint or lighting, 30 trials with viewpoint and lighting changes, and 24 trials with viewpoint and lighting changes along with the addition of visual noise. Therefore, a total of 72 points can be obtained on the test if one answers correctly on every trial. The original paper publishing the test reported an average accuracy score of 57.9 (Duchaine & Nakayama, 2006a), and other papers have since found similar scores (55.3-59.6; Garrido, Duchaine & Nakayama, 2008; Duchaine, Yovel, & Nakayama, 2007; Bowles, et al., 2009) though it is worth noting that inclusion of older participants tends to lower scores slightly (mean 50.4 with inclusion of 60+ year-olds; Stantić et al., 2021a; Bowles et al., 2009).

Commonly, participants undergoing a diagnostic evaluation for DP also complete the Cambridge Face Perception Test (CFPT, Duchaine, Germine & Nakayama, 2007). The CFPT consists of 8 trials, on each of which participants are required to sort faces based on resemblance to a test face. Unbeknownst to them, participants are provided with six morphed images, each of which is created by varying the percentage of the test identity (e.g., one image is a 10% morph of the target identity and another is a 50% morph). The test is scored by calculating a number of errors made from the objectively correct order of

images (and the final score calculated as 1 - % total errors, with chance accuracy at 35.6%; Rezlescu, Pitcher & Duchaine, 2012; Rezlescu et al., 2017). Instructions and execution of the test can be difficult to understand for atypical or elderly participants, and its reliability has been shown to be poor (Bowles et al., 2009).

A semi-structured clinical interview is often used to determine the extent of difficulties with face recognition in everyday life (Grueter et al., 2007; Murray, Hills, Bennetts, & Bate, 2018), though in recent years some research laboratories utilizing larger samples have made use of the 20-Item Prosopagnosia Index (PI-20, Shah et al., 2015) or the Hereditary Prosopagnosia Questionnaire (Kennerknecht, et al., 2006) instead to more objectively score this self-reported difficulty. These tools are, importantly, deployable online and can be used as part of screening procedures to test large number of participants before conducting a more comprehensive diagnostic battery in the laboratory. It is important to note that the PI-20 has been shown to correlate with objective performance reasonably well in participants who have very good and very poor perception of faces, but there appears to be little corelation with objective performance when tested in average participants (Palermo et al., 2017; Bobak, Mileva & Hancock, 2019; Arizpe et al., 2019). For this reason, researchers have rarely relied solely on scores from the self-report questionnaires or interviews to reach a DP diagnosis, but have taken these into account in combination with other validated computer-based tests.

Other tests utilize familiar faces (most commonly famous faces) in diagnosis of developmental prosopagnosia, though much less consistently, making comparisons across individuals and samples difficult. For example, the Famous Faces Test (FFT; Duchaine, 2000; Duchaine & Nakayama, 2005; Palermo, Rivolta, Wilson & Jeffery, 2011) requires

observers to recognize faces of celebrities. Performance on these tests is greatly dependent on participants' exposure to and interest in popular culture, as well as their age and cultural background. The test is generally normalized so that participants are only scored on their ability to recognize celebrities that they semantically report being familiar with. The Local Heroes Test (LHT; White, Rivolta, Burton, Al-Janabi, & Palermo, 2017) requires participants to determine whether pairs of presented images depict the same person or different people. Unbeknownst to participants, pairs of images are constructed such that either famous or unfamiliar people are presented, and the argument is that familiarity should aid perception in those who do not have an impairment in face recognition.

Due to a number of reasons, most notably the heterogeneity of the prosopagnosic population (Susilo & Duchaine, 2013; Duchaine & Nakayama, 2006a; Behrmann, Avidan, Marotta, & Kimchi, 2005), most researchers seek convergence from both qualitative and quantitative evidence. In this thesis I have done the same, only including participants who both reported difficulties in face perception in everyday life, as well as showed severe deficits on objective assessments of face perception and/or memory. Details of the inclusion criteria for prosopagnosic participants are outlined in Chapter 4.

Autism

While difficulties with face perception are not part of the formal diagnosis of autism as defined by the APA (APA, 2013), they are a commonly observed symptom in people with autism, though literature on the type of face recognition deficits observed in autism leaves many questions about the nature of this difference unanswered (see Weigelt, Koldewyn and Kanwisher, 2012 for review). Furthering our understanding of any autistic face

recognition deficit is the reason that studies outlined in this dissertation also include a sample of people with autism.

Autism is defined as a neurodevelopmental disorder (APA, 2013) with two groups of symptoms: 1) persistent deficits in social communication and interaction, and 2) restricted, repetitive patterns of behaviour, interests, or activities. A diagnosed individual must exhibit at least two symptoms of autism in early development (normally defined as the first two years of life) so diagnosis often relies on notes and observations from parents and carers (Pruette, 2013; Rutter, Couteur, & Lord, 2003). In addition, a standardized clinical interview (Autism Diagnosis Observation Schedule, ADOS, Lord et al., 2000) is regularly conducted, as well as self-report questionnaires measuring autistic traits (AQ-50, or a shortened version AQ-28, Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), which include components measuring social skills, communication, imagination, attention to detail, and attention switching.

It is widely accepted that emotion perception and recognition is affected in autism (Adolphs, Sears, & Piven, 2001; Hobson, 1986; Pelphrey et al., 2002; Boraston, Blakemore, Chilvers & Skuse, 2007; Ashwin, Chapman, Colle & Baron-Cohen, 2006; Humphreys, Minshew, Leonard & Behrmann, 2007; Wallace, Coleman & Bailey, 2008). A number of explanations for this phenomenon have been proposed, including suggestions that these differences are driven by differences in perception of positive and negative emotions (Pelphrey et al., 2002; Boraston et al., 2007; Ashwin et al., 2006; Humphreys et al., 2007; Wallace, Coleman & Bailey, 2008), simple and complex emotions (Adolphs, Sears, & Piven, 2001; Ashwin et al., 2006), or issues with integration of emotional signal across the face parts (Boraston et al., 2008; Shah, Bird & Cook, 2016). A delay in

integration of emotional information across face parts seems to be driving issues both in laboratory-based tasks as well as everyday life, both of which require quick and appropriate behavioural responses. Most commonly, differences in gaze modulation during interactions with others, notably a lack of focus on the eye region of the face, have been named as the underlying reason for difficulties in perceiving emotion in autism (Joseph & Tanaka, 2003; Rutherford, Clements, & Sekuler, 2007; see Cuve, Gao & Fuse, 2018 for a recent review). More recently, alexithymia has been found to largely explain atypicalities in eye-gaze patterns and recognition of emotional facial expressions in autism (Cuve et al., 2021; Gaigg, Cornell & Bird, 2018; Shah, Catmur & Bird, 2016; Bird & Cook, 2013, Cook et al., 2013; Bird, Press & Richardson, 2011).

However, there is less clarity on whether identity recognition is also impaired in autism. The large number of studies conducted so far offer mixed results, with some reporting typical, and others impaired, identity recognition in autism. A recent comprehensive review argued that these results can be explained by a dissociation between memory and perception tasks, and more notably the inclusion of face memory or face perception measures in research protocols (Wiegelt, Koldewyn, Kanwisher, 2012). Specifically, this review suggests that autistic participants present a deficit in face memory, but show intact face perception, and that the mixed findings are a result of any reliance of short- or long-term memory for faces in tasks included in the batteries. Specifically, the findings of the review indicate that the memory component to any tests makes the tasks more difficult to participants with autism and that, as a result, differences are found in batteries that include tests that rely on face memory and less commonly in batteries that focus solely on face perception measures. It is, importantly, worth highlighting that the literature reporting data on face perception and face memory in the same participants with autism remains very

scarce, with only a single developmental study examining these abilities in the same sample to date (Gepner, Gelder, & Schonen, 1996). This is problematic due to the large degree of heterogeneity within the autistic population; it could be the case that one sample of autistic individuals tested in one study may have face impairments and happen to be tested on a face memory task, while another study tests a different sample of autistics who do not have face impairments and they happen to be tested using a face perception task.

Elderly adults and dissociation between memory and perception

Finally, a lot of reports of face recognition differences concern elderly adults, who seemingly struggle with face recognition. Though research on the topic has addressed processes from very low-level vision (Roberts and Allen, 2016) to higher-level recognition processes, there is little explanation for the consistently-reported findings of substantial decline in performance on face memory (Maylor, 1990; Maylor & Valentine, 1993; Crooke & Larrabee, 1992; Duchaine, Germine & Nakayama, 2007) and face perception tasks (Duchaine, Germine & Nakayama, 2007; Shah et al., 2015; Megreya & Bindemann, 2015; Habak, Wilkinson, & Wilson, 2008) with age. A review of this literature, however, makes it obvious that the methodology of studies that examined the issue has varied widely – from the tasks that have been applied to this population to the selection of the population itself. Most importantly, there seem to be no existing studies that examine this age-related decline in parallel – looking at face memory while controlling for face perception ability. This is crucial because perception is a prerequisite of memory, in that every task involving any degree of face memory will inevitably rely on face perception. Reported findings of a decline in face memory related to age may, therefore, be driven by decline in face perception, but current literature gives no insight into this. Some work in

this dissertation sets out to differentiate whether any of the reported differences stem from differences in perception, or if the issue does, in fact, stem from a decline in face memory with age.

Super recognition

In the last decade, an increasing number of studies have looked into the phenomenon of super recognizers (SRs), individuals who recognize faces with extraordinary accuracy (Russell et al., 2009), even other-ethnicity faces which typically are recognized worse than own-ethnicity faces (Robertson et al., 2020). These studies have attempted to better understand the nature of their superior face recognition mechanisms, though the overall size of the research body addressing these questions remains small in relation to the number of studies assessing face processing in developmental prosopagnosia.

Preliminary evidence has shown that some SRs exhibit heightened holistic processing (Bobak et al., 2016; Russell et al., 2009), while individuals with developmental prosopagnosia exhibit impaired holistic face processing (Avidan, Tanzer & Behrmann, 2011; Towler, Fisher & Eimer, 2018). Further, some SRs report qualitatively and quantitatively enhanced differences in face perception (Bobak et al., 2016, 2017), while impairments in face perception have been documented in individuals with developmental prosopagnosia (Biotti, Gray & Cook, 2019; Dalrymple, Garrido & Duchaine, 2014, although see also Jackson, Counter & Tree, 2017) though it is important to note that these claims have been made on the basis of imperfect tests. Accordingly, it is of theoretical importance to compare whether the processes that underpin the impairments in developmental prosopagnosia are the same processes in which SRs perform exceedingly

well. This would provide evidence for developmental prosopagnosia and SRs populations being the opposite end of the normal population.

Like developmental prosopagnosia, super recognition is not a difference defined in the DSM-V and has simply been defined as an extraordinary ability at the top end of the face recognition spectrum, leaving a great deal open to interpretation of individual research groups as to who counts as a SR. As a result, different research laboratories use different procedures to diagnose super recognizers. Originally, Russell et al. (2009) proposed that SRs ought to be defined as individuals presenting with an extraordinary ability in both face perception and face recognition, though it is worth noting that their own examination only focused on 4 individuals with suspected SR and not all performed out of the ordinary ranges on the perception measure that was included. Other researchers have since largely departed from this guidance and now classify SRs as individuals with either an extraordinary ability in perception or memory (Bobak et al., 2016; Noyes et al., 2017; Phillips et al., 2018).

Possibly resulting from the inconsistency in the choice of tests deployed to classify SRs, the resulting literature has been very heterogeneous. There has, accordingly, been considerable discussion over this issue in the past years, with debate centring over whether super recognition ought to be defined by its applicability in everyday settings (e.g., whether recognisers perform particularly well in policing and national security tasks) or by its deviation from performance norms defined psychometrically using previous research (Bate, Portch, Mestry, 2021; Ramon, Bobak, White, 2019; Ramon, 2021; Young & Noyes, 2019). Regardless of the specific cut-offs for identification of a participant as a SR, there is consensus in the field that instead of using a single test, as has largely been done in recent

years, a multi-test assessment would be preferable (Bate et al., 2018; Stacchi et al., 2020; Fysh et al., 2020; Ramon, 2021). Instead of relying on a single measure, there is therefore a movement to merge participant pools across laboratories and identify true SRs by classifying their relative performance on a wider battery of validated and more difficult tests of face perception and memory.

As far as current screening tools are concerned, the vast majority of papers utilizing samples of super recognizers define their participants based on their performance solely on the Cambridge Face Memory Test, CFMT+ (Russell et al., 2009). This is an extended version of the CFMT which includes 30 additional, more difficult, trials designed to distinguish super recognizers from the superior end of typical performers. While the cut-off for diagnosis differs between research groups, most have followed the requirement that participants score 2 standard deviations above the neurotypical norms. In practical terms, participants have generally needed to score above 90 or 95 on the test to be included. For the purpose of studies included in this dissertation, a consistent criterion is applied for inclusion of participants such that all participants reported having a superior face recognition ability, as well as scored above 90 on the CFMT+ (details about the inclusion criteria can be found in Chapter 5).

Individual differences in the 'normal' range

Studies of the typical population have shown face recognition ability to be normally distributed, with wide individual differences in the ability to recognize identity from faces (Bate, Parris, Haslam, & Kay, 2010; Bowles et al., 2009; Russell et al., 2009; Megreya & Bindemann, 2013; Megreya & Burton, 2006). These individual differences have not been

widely studied, which may be in no small part due to the nature of the face perception tests commonly used in research. Most tests outlined above have been designed specifically with the purpose of diagnosing participants on either end of the spectrum (see CFMT, CFMT+, FFT, CFPT, PI-20) and do not have good discriminability within the typical population. Further, only small correlations have been reported between subjective assessments (such as self-report questionnaires) and objective computer-based tasks (such as the CFMT) in the typical population (Bindemann, Attard, & Johnston, 2014; McGuin, Richler, Herzmann, Speegle, & Gauthier, 2012; Bobak, Mileva & Hancock, 2019). This further makes it difficult to reliably study large samples of typical participants and map the relationships between various aspects of face processing ability in that population.

Other tests, designed to be deployed in typical populations, have run into challenges as well. These issues can be grouped into three themes: 1) tests that discriminate poorly; 2) tests that are too difficult or uncorrelated with face recognition abilities; and 3) tests that are too easy.

Some tests have been deployed in typical populations but found to discriminate poorly between participants. Most notably an issue with the Benton Facial Recognition Test (Benton et al., 1983) and the Warrington Recognition Memory Test (Warrington, 1984), research has shown over time that prosopagnosic participants regularly score in the normal range on these (Duchaine & Nakayama, 2004; Duchaine, 2000), as well as the fact that typical participants rely on compensatory strategies such as feature matching and use of non-facial information for discrimination judgments (Duchaine & Weidenfeld, 2003). As a result, both of these tests are rarely deployed in research.

Attempts to make more difficult tests with better discriminability have been moderately successful. In general, tests have been made more difficult by increasing ecological validity and seeking to mirror everyday face perception more closely (examples include the Facial Identity Card Sorting Task (Jenkins et al., 2011) and the Yearbook Test (Bruck et al., 1991)). These tasks resemble certain everyday life situations more closely, asking participants to sort face photos into groups of perceived identities or match images across a long period of time, like one would do when viewing a yearbook. Both have been shown to be very difficult, with most participants making a great number of errors (Stacchi et al., 2019). More importantly, as could perhaps be expected, one's ability to account for facial effects of ageing or sort images into piles does not correlate with individual differences in other face recognition tasks. As a result, these tasks are deployed in very specific situations when study aims require measurement of these very specific abilities.

The converse has also been true, with some tests simply being too easy to usefully discriminate between participants. For instance, a standardly deployed test of face matching, the Glasgow Face Matching Test (GFMT, Burton et al., 2010) has become less useful over time. Even when first published, the test was reasonably easy, with an average participant scoring 82% accurately. Over time, average accuracy scores have increased to 90% in typical populations (Towler et al., 2019) and ceiling effects are more consistently reported in the typical population (Dowsett & Burton, 2015; Stantić et al., 2021a; Stantić et al., 2022a; Dunn et al., 2020; Verhallen et al., 2016). Much of this issue could be driven by unlimited viewing time which allows participants to utilize featural matching strategies, like with the Benton Facial Recognition Test and Warrington Recognition Memory Test. This made the GFMT impractical in settings investigating individual differences (Fysh & Bindemann, 2018). An updated version of the GFMT was made available to the research

community following the completion of experimental work described in this dissertation (GFMT2, White et al., 2022). Featuring several improvements to the original version, the new test might be more suitable for investigations of individual differences in typical groups. Most importantly, the long version of test was specifically developed to be more difficult and improve the range of test scores, therefore yielding more discriminability in large groups.

Two tests have shown promise in measuring performance in typical groups – the Kent Facial Matching Test (Fysh & Bindenmann, 2018) and the Models Face Matching Test (Dowsett & Burton, 2015). Both are applied measures that increase the difficulty levels by incorporating more natural within-person variation, and both show an increased range in typical population accuracy. They have accordingly been deployed more frequently in recent work despite lacking some validation measures (Fysh, Stacchi & Ramon, 2020; Stacchi et al., 2020). Although able to measure individual differences in typical groups, these two tests have not been able to discriminate performers in lower and upper parts of the distribution, which is likely why they have not been deployed in research more widely.

Issues with existing tests

It is obvious from this overview that a single test sensitive to individual differences across the full range of facial recognition performance spectrum – from those scoring as poorly as DPs to those scoring as well as SRs, including everyone in between – does not currently exist. Such a test is difficult to design, even in principle. Tests designed to identify DPs are inevitably focused on providing the precision in the lower range of the performance spectrum and are therefore likely insensitive to the upper echelon of performers, unable to

differentiate between SRs and those who have good recognition but fall within the normal range. The converse is also obviously true of tests that are designed to identify SRs. Finally, tests that are designed to measure individual differences in the typical population are typically insensitive to performance at the edges of the spectrum because of being finely tuned to typical performance in the middle of the spectrum.

The field is, therefore, in a need of a novel test that is capable of measuring differences in the typical range of performance, as well as in the top and bottom ranges characterized most frequently by SRs and DPs. A test that can fulfil this purpose must understandably contain items of varying degrees of difficulty. A sufficient number of easier items should be included, allowing individuals with DP to be differentiated from those who score in the lower range of typical performance; and a sufficient number of more challenging items must be included to allow individuals with SR to be differentiated from those who score in the upper range of typical performance. Finally, for this test to meaningfully measure individual differences in the typical range of performance, a sufficient number of items of 'medium' difficulty need to be included.

Even if one had an unlimited number of task item options available, establishing their difficulty is not straightforward, mostly due to the variety of strategies that can be deployed to do this. In practice, authors of existing tests were not choosing from a virtually unlimited array of task item options – in fact, the pool of options was always reasonably small – so it was always possible to simply examine available options while choosing. For instance, one could administer each possible test item to a large sample of typical participants, and then calculate its difficulty as [1 - % of performers who make an error]. With such use of typical performance as a proxy for difficulty, items can be selected to

span a range of difficulties, from those on which few participants make a mistake to those on which performance is closer to chance. Though this has been the most commonly-used method in the existing literature, there are a number of issues with using this technique.

Firstly, it is now widely recognized that large-scale online testing can be unreliable, proving noisy data due to low-effort responses that are difficult to identify and remove (Fleischer, Mead & Huang, 2015; Pyo & Maxfield, 2021; Zorowitz, Niv & Bennetts, 2021), an issue that was not as widely known when the tests were first designed. A selection that relies on such a sample to determine difficulty of individual items is also vulnerable to issues of presentation order (trials presented earlier might yield higher-effort responses than those presented later), variation in effort (with long testing procedures, participants might focus on the task at one point but not another) and distinguishing difficulty from random answering (an item might have lower accuracy due to its true difficulty or due to a large proportion of participants answering at random). Furthering the issue is the well-established finding that humans recognize same-ethnicity faces much better than different-ethnicity faces (Lindsay, Jack, & Christian, 1991; Valentine & Bruce, 1986; Ellis & Deregowski, 1981). As a result of this, it is crucial that the demographic makeup of the participant group selecting the images and the demographic makeup of the images included in the items being tested be similar – and this has often not been the case thus far. Finally, some studies have reported differences in face perception performance as early as age fifty when studying healthy ageing (Crook & Larrabee, 1992; Duchaine, Germine, & Nakayama, 2007). However, the vast majority of item-selection based on neurotypical performance has utilized young adult populations, often undergraduate students. Given the wide range of testing environments in which face perception batteries

are deployed, including testing older adults, this raises further questions about the appropriateness of using measures based on student populations alone.

The most important issue, however, is a much more nuanced methodological problem. There are known reliable differences between strategies deployed by different groups that make a selection strategy that relies solely on neurotypical performers problematic. For instance, typical observers have been argued to process faces holistically, while those with developmental prosopagnosia or ASD process faces in piecemeal fashion, part by part (Gauthier et al., 2009; Joseph & Tanaka, 2003; Avidan et al., 2011; DeGutis et al., 2012; Palermo et al., 2011; though see Biotti et al., 2017; Brewer et al., 2019; Faja et al., 2009). Similar arguments have been made about the processing in healthy elderly adults (Konar, Bennett, & Sekuler, 2013; Schwarzer, Kretzer, Wimmer, & Jovanovic, 2010; Wiese, Kachel, & Schweinberger, 2013; Meinhardt-Injac, Persike, & Meinhardt, 2014; Shwarzer, Kretzer, Wimmer, & Jovanovic, 2010). Therefore, using typical performance as a proxy for difficulty might offer poor calibration for atypical populations and therefore not result in the chosen items truly spanning a full range of difficulty space. This could systematically bias findings in any atypical group. More specifically, it is in theory possible that there is no difference between a neurotypical person and an autistic person when performing on an unlimited number of task items. However, for practical reasons tests are not infinitely long and item numbers are limited in scope. If a selection of items is made based on neurotypical performance, it is possible that differences in strategy, shape of face space or preference would result in a sub-selection of items such that the neurotypical performance seems different from autistic performance, despite not in reality being different. This issue can be translated to any test for which items are chosen by typical performers only. It is worth noting that the converse strategy is not tenable either -

calibrating difficulty based on performance of atypical groups would result in the same limitation when using the test in typical samples.

I suggest in this dissertation that the solution to this problem is moving away from typical samples altogether. Instead, I suggest relying on an objective, non-biased method of determining difficulty of test items. One way of doing this is by using the performance of validated, rigorously-tested face recognition algorithms. If you use a face matching task in which participants are asked to determine whether two presented faces are of the same person or different people, then an important determinant of difficulty is similarity of those two faces. Similar faces then constitute easy trials when the faces are of the same identity and difficult trials when the faces are of different identities. Conversely, when faces are dissimilar, they constitute easy trials when the faces are of different identities, and difficult trials when the faces are of the same identity. Thus, algorithmically-derived facial similarity can be used to estimate difficulty of trial instead of relying on judgments of typical participant samples.

It is important that multiple algorithms are used to prevent overweighting the performance of a single DNN. This is particularly crucial given that these types of convolutional neural networks do not reveal what information is used by each layer of the network to make a determination about the stimuli it is presented with. We, therefore, need to be sure that multiple measurements of difficulty (which are indexed here as [1-algorithmically-provided similarity]) are taken into account for each item that is considered for the test. Importantly for the logic of the selection process, all that is required is that the algorithmic judgment about the two faces is accurate. This aspect of the algorithm's determination is always verifiable given that there is an absolute truth about the two presented faces (they

either are, or are not of the same person, and the researcher always knows which is true). Usefully, it is not necessary to know the peculiarities of the strategy the algorithm is relying on to make this judgment; in fact, different algorithms using different strategies to make these judgments decrease the likelihood that selection of algorithms has itself biased the results. A crucial technicality is also that the same strategy is deployed by the same algorithm for each pair of faces that is considered as a test item. Given extensive documentation of these stable, commercially available algorithms, this should only become an issue if the research is assessing items at different times and an algorithm's strategy has been altered in the meantime.

Using algorithmically derived similarities presents a novel approach in the face perception literature that allows for objective judgments of item difficulty. It removes bias from the judgments of difficulty by relying on multi-layered learning models instead of large groups of typical observers and therefore reduces or eliminates bias to neurotypicality. Most importantly, it allows for an expansion of the difficulty span in an objective manner. Expanding difficulty in this way should allow the entirety of the population to be meaningfully assessed using the same set of test items.

Further to developing the first perception measure to rely on algorithmic similarities of faces, I use a novel approach to test what people's ability to judge the similarity of faces tells us about their face matching ability (ability tell faces apart or 'together'). Specifically, I query whether I can use the objective performance of the selected combination of algorithms to derive each participant's ability to determine face pair similarity (by calculating how much they deviate from the algorithmically derived similarity). Further, I investigate what this 'precision' tells us about performance on other validated face

perception and face memory measures. I can answer this question in a large sample of typical participants, as well as sizeable samples of prosopagnosics, super recognizers, people with autism and healthy elderly adults. This approach allows me to examine the nature of any face processing differences in these atypical groups.

1.2. Conclusion/objectives

Taken together, the tests of face perception currently used in the reviewed literature have so far yielded mixed results in differentiating between groups of different abilities. Our understanding of the mechanisms underlying healthy and disordered face recognition has been furthered by decades of study on the topic, but we still have a limited understanding of individual differences in typical groups, and limited tools for comparing performance across populations of different abilities. Whether mixed findings can mostly be attributed to different strategies for choosing participant samples across studies or insufficiently precise tests remains to be seen. Tests that are most used in research today are limited in the scope of conclusions they offer researchers for a number of reasons. Most importantly, they were developed with diagnosis of a particular population in mind and regularly seek to allow discrimination between typical and atypical performance, either at the lowest or topmost end of the spectrum. Other tests are poor at discriminating typical performers or affected by training effects over time, with ceiling effects observed even in typical performers.

However, the most prominent problem is that these tests were developed after validation by human observers only, often solely by neurotypical students. With most tests using items that were selected based on inputs from a small number of non-representative human

observers, conclusions of these studies remain open to questions of bias that can consistently skew the results when testing a different population.

1.3. Guiding research questions

As described through the topics introduced above, this dissertation seeks to clarify our understanding of the mechanisms involved in human face recognition by attempting to answer five research questions, as follows:

1. Can use of computer vision algorithms inform development of more objective and less biased tests of face perception, and do these tests provide new insight into individual differences in face perception?
2. How do individual differences in face perception explain performance on other tasks and how does this ability change with age?
3. Can use of tests that are free of neurotypicality bias expand our understanding of the performance of atypical groups (developmental prosopagnosics, people with autism and super recognizers)?
4. How does the ability to separately measure matching and perception performance help us gain insight into face recognition mechanisms?
5. Does the relationship between human and algorithmic assessments of similarity tell us anything about an individual's face recognition ability?

This dissertation seeks to systematically assess these questions by developing and validating novel sensitive psychophysical tasks, and then deploying those tasks in sizeable

samples of individuals with developmental prosopagnosia, autism, and super recognition, as well as healthy elderly adults, on top of a large sample of typical participants. All these participants completed a comprehensive set of validated and novel face perception and memory measures in order to elucidate face recognition mechanisms and further our understanding of the qualitative, as well as quantitative, nature of any differences between these groups.

This dissertation furthers existing work by asserting that useful formal comparisons of human performance on face recognition tasks must be preceded by a change in frameworks to design, test and validate new measures. Importantly, novel measures must feature items that vary in difficulty based on objective measures of difficulty rather than, as was previously standard, a small number of human difficulty judgments. In Chapters 2 and 3, I propose a new matching measure designed and validated by objectively measuring difficulty through use of a number of state-of-the-art face recognition algorithms. Further, I investigate use of this novel measure in atypical populations on both ends of the spectrum. Unlike previously available tests, deploying this measure allows me to independently assess performance in matching, perception, and memory domains in typical and atypical populations. Finally, I am able to derive a measure of an individual's face similarity space relative to that of the algorithms. This ability to independently assess perception, matching and memory allows me to examine individual differences in performance on a variety of other face recognition tasks.

Chapter 2 will outline the process of developing and validating the novel measure of face perception, the Oxford Face Matching Test (OFMT).

Chapter 3 looks into individual differences in face perception ability as measured by the OFMT, and its relationship to other face recognition processes (perception, meta-cognition, memory), in a large sample of neurotypical participants both in-person and online.

Chapter 4 seeks to explore whether administration of the OFMT in a large sample of developmental prosopagnosics gives sufficiently sensitive scores at the lower end of the performance spectrum.

Chapter 5 examines whether the OFMT shows sufficient sensitivity to the upper end of the performance spectrum by deploying it in a large group of super recognizers, who have traditionally scored at ceiling on any tests designed for typical populations.

Chapter 6 seeks to explain the mixed results in existing literature about face perception and memory in autism by exploring face matching and face memory performance in the same group of people with autism.

Chapter 7 examines ageing-related changes to face perception and memory abilities in a large sample of healthy elderly participants by deploying an extensive battery of face recognition measures.

Chapters 3, 4, 5, and 6 also investigate the relationship between explicit judgments of face pair similarity in humans and algorithms in their respective typical and atypical populations and explore how the precision of this ability relates to other face recognition processes.

Chapter 2: The development and validation of the Oxford Face Matching Test

An edited or partial version of this chapter has been published as the following journal articles:

- Stantić, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., Catmur, C. & Bird, G. (2021). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavior research methods*, 1-16.

- Stantić, M., Knyspel, J., Gaches, A., Liu, Y., Bird, G.P., & Catmur, C. (under review) The Oxford Face Matching Test: Short Form Alternative

2.1. Abstract

Tests of face processing are typically designed to identify individuals performing outside of the typical range; either prosopagnosic individuals who exhibit poor face processing ability, or super recognisers, who have superior face processing abilities. Here we describe the development of the Oxford Face Matching Test (OFMT), designed to identify individual differences in face processing across the full range of performance, from prosopagnosia, through the range of typical performance, to super recognisers. Such a test requires items of varying difficulty, but establishing difficulty is problematic when particular populations (e.g., prosopagnosics, individuals with Autism Spectrum Disorder) may use atypical strategies to process faces. If item difficulty is calibrated on neurotypical individuals, then the test may be poorly calibrated for atypical groups, and vice versa. To obtain items of varying difficulty we used facial recognition algorithms to obtain face pair similarity ratings that are not biased towards specific populations. These face pairs were used as stimuli in the OFMT, and participants required to judge if the face images depicted the same individual or different individuals. Across five studies the OFMT was shown to be sensitive to individual differences in the typical population. The test-retest reliability of the task was at least equivalent to the Cambridge Face Memory Test and the Glasgow Face Matching Test. Short versions of the task are also validated.

2.2. Introduction

The ability to recognise our conspecifics accurately and rapidly is critical in supporting our social interactions. Impaired recognition of others is thought to contribute to social anxiety (Yardley et al., 2008; Dalrymple et al., 2014), and to trigger a cascade of impairments including in the inference of mental states (Conway, Catmur, & Bird, 2019), attachment issues (Happé, Cook, & Bird, 2017), and in the establishment and recognition of group affiliations (Verosky, Turk-Browne, & Todorov, 2013). Impairments in facial recognition, potentially the primary way in which we recognise others (Bruce & Young, 1986), are also thought to contribute to the symptoms observed in Autism Spectrum Disorder (Boucher & Lewis, 1992; Wallace, Colleman, & Bailey, 2008; Hedley, Brewer, & Young, 2011).

Within the general population (i.e., those without a clinical diagnosis), individuals at the extremes of face recognition ability are labelled as either 'prosopagnosic' (if their face recognition is very poor) or as a 'super recogniser' (SR; if they exhibit superior face recognition). The majority of research interest thus far has focussed on prosopagnosic individuals (e.g., Duchaine & Nakayama, 2005; Duchaine & Nakayama, 2006b; Susilo & Duchaine, 2013; Bate & Tree, 2017; Towler, Fisher, & Eimer, 2017; see Corrow, Dalrymple, & Barton, 2016; Cook & Biotti, 2016; Geskin & Behrmann, 2018 for recent reviews), particularly those who have suffered from face recognition impairments throughout life rather than as a result of brain trauma (developmental prosopagnosics; DPs). However, more recent work with SRs has probed the specificity and nature of their enhanced face recognition ability (e.g., Russell, Duchaine, & Nakayama, 2009; Bobak et al., 2016; Bate et al., 2019a; Bate et al., 2019b).

A single test that is sensitive to individual differences across the full range of facial recognition performance is hard to design – tests designed to identify DPs are likely insensitive to the range of ability needed to distinguish SRs from those with good recognition within the normal range, and the converse is true for tests designed to identify SRs. This paper therefore presents the development and validation of a test that is able to measure individual differences in the typical range of performance, as well as in the range characterised by DPs and SRs.

A test suitable for the full range of performance should clearly consist of items of varying degrees of difficulty: with easier items allowing those with DP to be distinguished from those at the lower end of typical performance, and more challenging items allowing SRs to be distinguished from those at the higher end of typical performance, and with individual differences in ability within DPs and SRs also detectable. Establishing the difficulty of items in a task that can be completed successfully using a variety of strategies is not straightforward, however. Difficulty could be established by administering items to a large sample of participants and recording average performance. If performance is taken as reflecting difficulty, then items can be selected that span a wide range of difficulty. However, if there are reliable differences between the strategies or processing styles of different groups (for example if typical observers process faces holistically while those with developmental prosopagnosia or Autism Spectrum Disorder process faces in a piecemeal fashion; Gauthier, Klaiman, & Schultz, 2009; Joseph & Tanaka, 2003; Avidan, Tanzer & Behrmann, 2011; DeGutis et al., 2010; Palermo et al., 2011; though see Biotti et al., 2017; Brewer et al.,, 2019; Faja et al., 2009), the items may be poorly calibrated for those populations with atypical strategies, with relative difficulty of items varying as a function of observer group. The sensitivity of such tests would therefore be biased towards the typical

population. Items calibrated based on performance of atypical groups would suffer the same limitation when used with typical samples.

To overcome this problem, the current test of face processing determined item difficulty in a non-biased, objective[1] manner, using the performance of facial recognition algorithms. Despite ethical concerns concerning the applied use of such algorithms (Mazura et al., 2012; Mogapatra, 2015; Klare et al., 2012), and findings suggesting that performance of some algorithms may be affected by ethnicity in the same way as seen in some human recognisers (e.g., Irebhude, Celestine & Aget, 2019), it is clear that, on average, performance of the leading face recognition algorithms is at least equal to, and regularly better than, that of human recognisers (Phillips & O'Toole, 2014; Phillips et al., 2018). Note that several of the leading algorithms are based on deep neural networks, meaning that they are 'black boxes' in which the strategies / rules determining performance are difficult to determine. Importantly, however, it is not essential that such algorithms use similar strategies to human observers; indeed, if there are reliable differences in the strategies used by different human sub-populations in recognising faces, then it is preferable that algorithms do not utilise the same strategy as (typical) human observers – all that is required is that identity recognition is accurate. Accordingly, three leading facial recognition algorithms (FaceSoft, Azure Face Recognition and AWS Rekognize) were used to determine the similarity of pairs of face images. Pairs of faces were selected as potential stimuli if similarity judgments were consistent across algorithms.

[1] Here 'objective' is used as defined in the Oxford English Dictionary "not influenced by personal feelings or opinions in considering and representing facts; impartial", without implying accuracy or status as 'ground truth'.

These face pair stimuli, and their objective degree of similarity, were used to derive a stimulus set that varied in difficulty when used in a face matching task requiring participants to determine whether two images of a face were of the same individual, or different individuals. Similarity is therefore related to difficulty in the following manner: very similar face images of the same person, and very different images of different people, are likely to be easier to judge correctly. In contrast, very different images of the same person, and very similar images of different people, are likely to be more difficult to judge correctly. Thus, sorting face pairs of the same individual and face pairs of different individuals by objective similarity allows the creation of items of increasing difficulty when used in a matching test. Note that such a test measures what has been referred to as face perception or face matching – determining whether instances of a face are exemplars of the same facial identity – not what has been referred to as face memory or identity recognition (labelling a face as a particular identity). Thus, it is possible for individuals to perform well on such a test, to be able to determine that Individual A is not Individual B, but not know that individual A is Derek Trotter and Individual B is Rodney Trotter. Face matching is necessary for identity recognition, however, as one must be able to determine that a particular face stimulus is a token of type 'Derek Trotter's face', in order to be able to recognise Derek Trotter (Bruce & Young, 1986).

Across three studies in this chapter, we report the development of the task (named the Oxford Face Matching Task, OFMT), the range of performance observed across individuals tested in laboratory and online, the relationship between performance on the OFMT and standard face recognition tasks, and test-retest reliability of the OFMT in comparison to standard measures. Further validation of the task is provided in Chapters 4 and 5 by testing sizeable samples of participants with DP and SR.

2.3. Common methodology

Throughout the rest of this dissertation, a number of tasks and procedures will be used consistently for purposes of comparability across participant groups. These are outlined here and any differences to the outlined procedure are referred to in individual experiments' methods sections henceforth.

2.3.1. Tasks

The development of the OFMT will be outlined below. Participants across all experiments reported in this dissertation completed the Long-Form of the OFMT. Participants across all experiments also completed some or all of the following standard measures of face perception or memory. In all studies, tests were completed in a randomized order.

Cambridge Face Memory Task (CFMT; Duchaine & Nakayama, 2006a)

The CFMT has been used extensively to measure face recognition, and was designed to measure unfamiliar face memory skills to distinguish those with prosopagnosia from typical perceivers. Participants learn six target faces at the beginning of the test, after which they are tested on three-alternative forced-choice trials. On each trial, two images are distractors and one is an image of a learned target identity. The test is divided into three stages of increasing difficulty, involving 18 test trials with no change of viewpoint or lighting, 30 trials with viewpoint and lighting changes and 24 trials with viewpoint and lighting changes along with the addition of visual noise.

Glasgow Face Matching Task (Burton, White & McNeill, 2010)

The GFMT is a standard test used to assess unfamiliar face matching ability. Participants are shown two faces of either the same individual (match trials) or different individuals (mismatch trials) for an unlimited amount of time and asked to determine if faces belong to the same person or different people. The short form used here consists of 20 match and 20 mismatch trials, presented in a random order.

20-Item Prosopagnosia Index (PI-20; Shah et al., 2015; Gray, Bird, & Cook, 2017)

The PI-20 is a self-report questionnaire used as a screening tool to identify people with difficulties in face recognition. It has previously been validated with the CFMT and has been shown to able to distinguish DPs from the neurotypical population (Gray, Bird & Cook, 2017). The survey consists of twenty items on which respondents can report face recognition difficulties in everyday life.

Autism Spectrum Quotient-50 (AQ-50, Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001)

The AQ-50 is a self-report diagnostic questionnaire used to determine the severity of ASD symptoms in adults with normal intelligence (Woodbury-Smith, Robinson, Wheelwright, & Baron-Cohen, 2005).

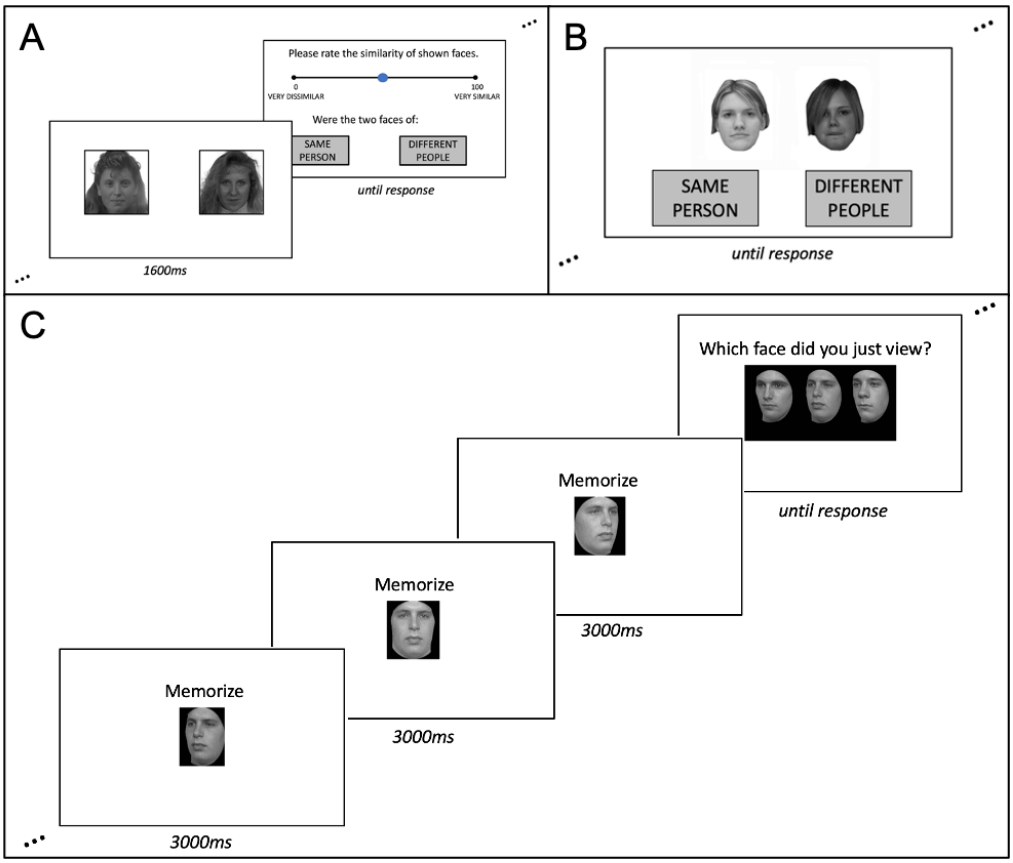


Figure 2.1. Sample trials of the face processing tasks.

A – the Oxford Face Matching Test, OFMT, a face matching task that presents faces for 1600ms before participants have to rate the similarity of two faces and decide whether the faces are of the same person or of different people (note that details of what an OFMT trial looks like are shown in Figure 2.2.); B – the Glasgow Face Matching Test, GFMT, a face matching task that presents faces for an unlimited amount of time while participants decide whether the faces are of the same person or different people; and C – the Cambridge Face Memory Test, CFMT, a face memory task in which participants learn faces from three viewpoints and subsequently select them from test displays with two distractors.

2.3.2. Exclusion criteria

Across all studies in this dissertation, any participants who met one or more of the following criteria were excluded: technical difficulties prevented completion of the task; did not meet recruitment criteria (over 18 years old, no current or past psychiatric or other neurodevelopmental diagnoses, no current use of psychotropic medications); failed to attend

to the task (determined by performance on attention check trials, outlined below). Additionally, unless recruited as part of a DP group, participants were excluded if they met any of the following criteria: scored more than 2 standard deviations below the group mean on the CFMT test (used to confirm prosopagnosia, Duchaine & Nakayama, 2006a); their PI-20 scores indicated the possibility of prosopagnosia (Shah et al., 2015, Gray, Bird, & Cook, 2017). Finally, any participant was excluded if their Autism Spectrum Quotient-50 questionnaire scores indicated high autistic traits (scores 32 or higher; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001).

2.4. Experiment 1

Experiment 1 aimed to develop the OFMT and to establish the range of performance on the test to assess its usefulness as a measure of individual differences. Participants completed both the Long-Form of the OFMT (which takes approximately 20 minutes), and a Short-Form version designed for testing situations requiring an accelerated procedure (which ranges from 4 to 10 minutes in length). In addition, scores on the OFMT were compared to standard measures of face processing such as the CFMT (Duchaine & Nakayama, 2006a) and the PI-20 (Shah et al., 2015; Gray, Bird, & Cook, 2017).

2.4.1 Methods

2.4.1.1. Selection of stimuli

Pairs of face stimuli were selected from databases, both publicly available (the FERET dataset provided by DARPA; Phillips, Moon, Rauss & Rizvi, 1997; Phillips, Wechsler, Huang & Rauss, 1998) and databases held by the authors. All databases contained ground

truth data as to whether the face pairs were images of the same person ('same pairs'), or of different people ('different pairs'). Images were naturalistic and varied in age and gender, but selected images were cropped to a 3:4 ratio to ensure that the face occupied most of the presented stimulus. Images of participants wearing glasses were excluded to avoid reliance on external features by both algorithms and humans. All images were of Caucasian people, frontal, presented with hair, but without background context. Images were both assessed by face recognition algorithms and presented to participants in greyscale, with no other low-level processing conducted.

Over three million face pairs were selected for similarity assessment by three algorithms: AWS Rekognition (https://aws.amazon.com/rekognition/), FaceSoft (retrieved from http://facesoft.io/) and Azure Face Recognition (https://azure.microsoft.com/en-us/services/cognitive-services/face/). The algorithms provide a similarity index ranging from 0 to 1 (or 0 to 100, scaled to 0 to 1 for ease of comparison). Face pairs were placed into 20 similarity 'bins' based on their similarity score, each 0.05 units wide (spanning the similarity scores from 0-1). Any face pair for which there was a difference of two or more bins on the similarity judgment between algorithms was excluded (for instance, if one algorithm rated a face pair as having 0.4 similarity and another as having 0.7 similarity, that pair was excluded for inconsistency). In practice, this was rare with only about a thousand of the three million pairs excluded for this reason.

For the remaining pairs, a mean similarity index was calculated across the three algorithms, and pairs were then assigned to one of the 20 bins based on their average similarity index. As a result, 20 bins of 'same' face pairs and 20 bins of 'different' face pairs were populated by possible test items to be considered. Ten image pairs were

randomly selected from each bin for use as stimuli in the OFMT, with the constraint that there were equal numbers of same and different pairs (resulting in a total of 200 stimulus pairs, 100 depicting the same identity and 100 depicting two different identities).

Stimulus similarity bins were then used to construct a range of difficulty bins. 5 same pairs and 5 different pairs from each of the 20 similarity bins were allocated to 20 new bins, arranged in order of 'difficulty', such that bin 1 contained the most similar same pairs and most different different pairs, while bin 20 contained the most different same pairs and most similar different pairs. Practically speaking, this meant that the ordering of the 'same' bins was mirror-inverted. Algorithmically provided 'same' bins spanned from 1 (very dissimilar same images, therefore difficult to recognize as the same person) to 20 (very similar same images, therefore easy to recognize as the same person). In the difficulty space, these were reversed. On the other hand, 'different' face pairs remained unchanged. Algorithmically provided 'different' bins spanned from 1 (very dissimilar different images, therefore easy to recognize as different people) to 20 (very similar different images, therefore difficult recognize as different people). A resulting set of bins, therefore, meant that same and different pairs within each bin ought to be comparably easy or difficult.

OFMT – Long Form

Participants were shown pairs of faces side-by-side for 1600ms and asked to make a judgment about the similarity of the faces on a scale of 1 (very dissimilar) to 100 (very similar), as well as to judge whether the face images were of the same person, or different people (see Figure 2.2.). A standard presentation time was used (rather than trials proceeding upon the participant's response) due to a concern that individuals with prosopagnosia may assume that they will perform poorly, and thus respond quickly without giving stimuli due

consideration, in order to finish quickly, or may compensate for their difficulties by studying the faces for longer time periods. Pilot testing with input from prosopagnosic individuals revealed that 1600ms was judged to be sufficient time for judgements as to the degree of similarity, and whether face images were of the same or different individuals, to be made**.** All 200 stimulus pairs were presented in a random order.

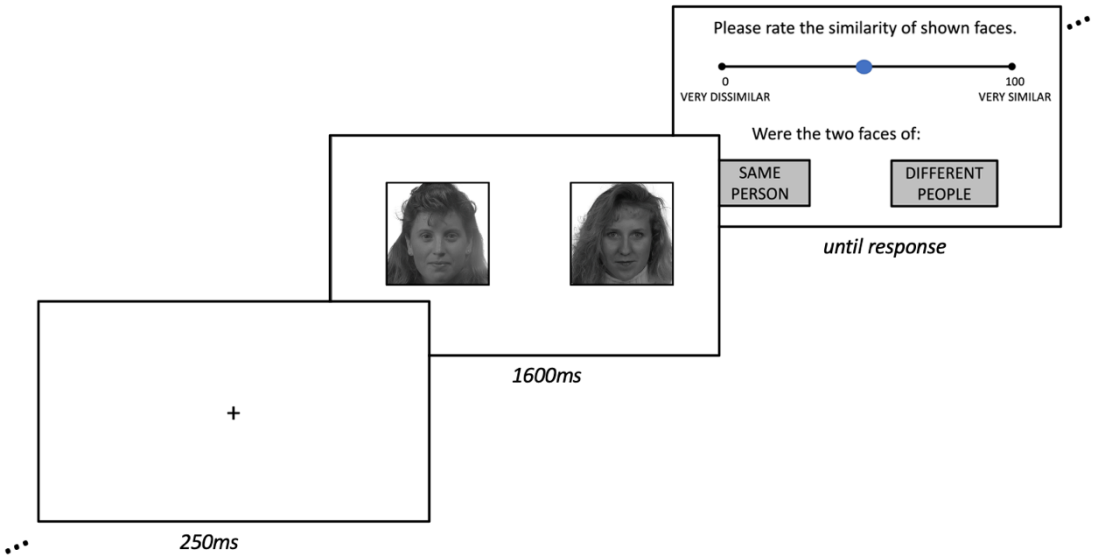


Figure 2.2. A sample trial from the Oxford Face Matching Task – Long Form.

Participants were shown two faces concurrently for 1600ms. They were asked to judge the similarity of the faces using a slider ranging from 1 (very dissimilar) to 100 (very similar), and to judge whether the face images were of the same person, or different people. Note: relative text size on the final frame enlarged to aid legibility.

Attention check trials

The long-form version of the OFMT included an additional 10 attention check trials. These trials were designed to be easy even for those with severe face processing impairments. Five same pairs were constructed using exactly the same images of a face, while five different pairs were constructed using faces of different genders. Note that pairs of faces of different genders consistently yielded similarity indices near zero, which is why they were selected

for attention check trials. There were no mixed-gender trials in the OFMT other than for these attention check trials. Any participant making two or more mistakes on these trials was excluded prior to analysis.

OFMT – Short Form

Trial structure was as detailed above for the Long Form of the task except that participants were not required to make a similarity judgement and instead only provided a judgment about whether the two faces presented were same or different. Rather than viewing all 200 face pairs, participants completed a reduced number of trials according to the logic of a staircase procedure, whereby difficulty increased following good performance, and decreased following poor performance.

Participants initially completed 50 trials of medium difficulty (10 trials from each bin in bins 8-12, in a randomized order). Based on their performance, three outcomes were possible: 1) Participants who did not achieve 80% accuracy on at least two consecutive bins (e.g., bins 8 and 9, bins 9 and 10, etc.) were presented with easier trials from consecutively lower bins until they achieved 80% accuracy on two consecutive bins or no more trials remained (i.e., they failed to achieve 80% accuracy on bins 1 and 2); 2) Participants who achieved 80% accuracy on the most difficult two bins (11 and 12) were presented with more difficult trials from consecutively harder bins until two consecutive failures to reach 80% accuracy per bin were observed, or no more trials remained (because participants achieved at least 80% accuracy on bins 19 and 20); 3) Participants who achieved 80% accuracy for any two consecutive bins (other than bins 11 and 12) completed no more trials. Participants always saw all trials from a bin before finishing the task or proceeding to the next bin. The threshold

was determined to be the average of the two highest consecutive bins completed with 80% accuracy. Note that an 80% accuracy threshold was used as this represents performance significantly above chance level (with an alpha of < .05).

2.4.1.4. Participants

For Experiment 1, 45 participants with normal or corrected-to-normal vision participated in the laboratory. Four participants were excluded for failing the attention checks, producing a final sample of 41 participants (24 female, mean age 26.6 years, SD = 3.9). Participants completed the PI-20, CFMT, and the OFMT Long- and Short- forms. The OFMT Short-Form was always completed first as it was deemed to be most susceptible to interference from prior completion of the other tasks, and the order of the other tasks was randomized for each participant. All tasks were completed using the Gorilla Experiment Builder platform (www.gorilla.sc).

2.4.2. Results

Accuracy on the OFMT Long Form ranged between 59.5% and 84% (raw scores: 119-168), with a mean of 74.1% (*SD* = 5.4%, raw scores: 148.2±10.8). A small-to-moderate relationship was observed between accuracy on the OFMT and CFMT ($r(39) = .32, p < .05$). A moderate relationship was observed between average algorithmic similarity values and judged similarity for same face pairs ($r(39) = .42$, $p < .05$) and a small-to-moderate relationship was for different face pairs ($r(39) = .32, p < .05$; Figure 2.3.). There was a small, non-significant relationship between PI-20 scores and performance on the Long-Form of the OFMT ($r(39) = -0.14, p = .38$).

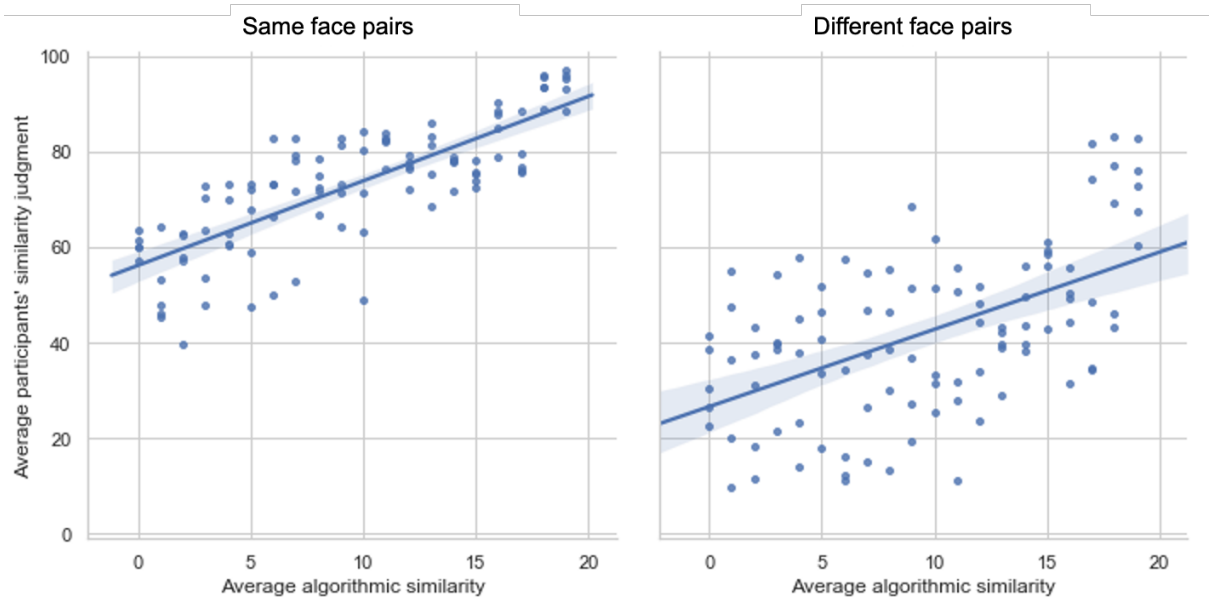


Figure 2.3. Relationship between participants' similarity judgments and algorithmic similarity values. Algorithmic similarity values are expressed here as bins 1-20. Each point shown on the graphs represents average group judgments for the similarity of a given face pair (5 same pairs and 5 different pairs are shown for each difficulty level).

A wide range of thresholds was observed using the Short Form of the OFMT (Figure 2.4., left panel). For participants for whom the threshold was found (one participant did not complete the task with >80% accuracy in bins 1 and 2), thresholds ranged from 1 to 15 (M = 9.7, SD = 3.4). In addition, a small-to-moderate statistically significant relationship ($r(39)$ = .34, $p < .05$) was observed between the threshold derived for each participant from the Short-Form of the OFMT and accuracy derived from the Long-Form of the OFMT (Figure 2.4., right panel).

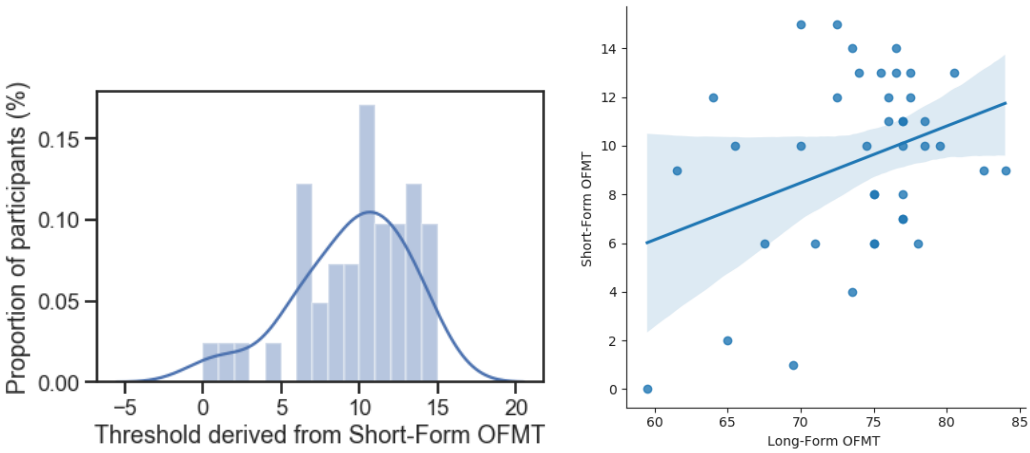


Figure 2.4. Distribution of scores on the Short Form OFMT and their relationship to Long Form OFMT. A distribution of thresholds derived from the Short Form of the Oxford Face Matching Task (OFMT; left panel). The right panel depicts the relationship between accuracy on the Long Form of the OFMT (200 trials) and the thresholds derived from the Short Form (between 50 and 110 trials, depending on performance). The shaded area in the right panel indicates the 95% confidence interval estimated using a bootstrapping method with replacement.

2.4.3. Discussion

Data from Experiment 1 indicate that both the Long- and Short-Forms of the OFMT can be completed and understood by participants and produce a wide range of performance. Performance on the OFMT Long-Form correlated significantly, but only moderately, with performance on the CFMT. This is to be expected as the difficulty of trials in the CFMT is increased, in part, through the addition of visual noise. This manipulation undoubtedly increases the difficulty of the task but may mean that individual differences in part reflect the ability to cope with visual noise, rather than pure face perception (it should be noted, however, that performance on the CFMT items with visual noise is significantly correlated with performance on items without visual noise; Duchaine & Nakayama, 2006a). Perhaps more importantly, it is also the case that performance on the CFMT is in part determined by individual differences in face learning/face memory, in contrast to the OFMT in which

performance is determined by face perception. Data also suggest that the Short-Form of the OFMT is, in its current form, a noisy proxy for the Long-Form test. The small-to-moderate correlation between the two forms of the OFMT is likely a result of the reduced number of items in the Short-Form, leading to a noisy estimate of an individual's ability.

2.5. Experiment 2

Experiment 2 constituted a direct replication of Experiment 1, except that participants completed it online rather than in the laboratory.

2.5.1. Methods

2.5.1.1. Participants

One hundred and six participants with normal or corrected-to-normal vision participated in the study online for payment. Nine participants were excluded for failing the attention check and/or failing to finish the task, after which the sample contained ninety-seven participants (49 female, M_{age} 27.2 years, SD = 7.71).

2.5.1.2. Procedure

The procedure was exactly the same as in Experiment 1, except that participants completed the tasks online, via the Prolific.co recruitment platform.

2.5.2. Results

Accuracy on the OFMT Long-Form ranged from 52% to 86% (raw scores: 104-186), with a mean performance of 72.9% (SD = 5.7%, raw scores: 145.8±11.4). As in Experiment 1, a

small-to-moderate statistically significant correlation was observed between accuracy on the OFMT Long-Form and CFMT scores ($r(95) = .34$, $p < .05$). A moderate correlation of $r(95) = .44$, $p < .05$) was observed between the Short-Form thresholds and Long-Form accuracy (Figure 2.5.). There was also a small relationship between PI-20 scores and performance on the Long-Form of the OFMT, which approached significance ($r(95) = -.20$, $p = .06$).

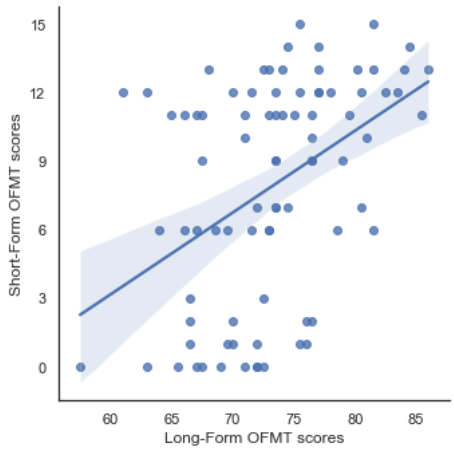


Figure 2.5. Relationship between Long-Form and Short-Form OFMT accuracy.

Note that the Long-Form OFMT consists of 200 trials and the thresholds from the Short-Form is derived from 50 to 100 trials, depending on performance.

In line with the findings of Experiment 1, a moderate relationship between algorithmic similarity and judged similarity was observed for same face pairs ($r(95) = .44$, $p < .05$) and a small-to-moderate relationship was observed different face pairs ($r(95) = .34$, $p < .05$) (Figure 2.6.).

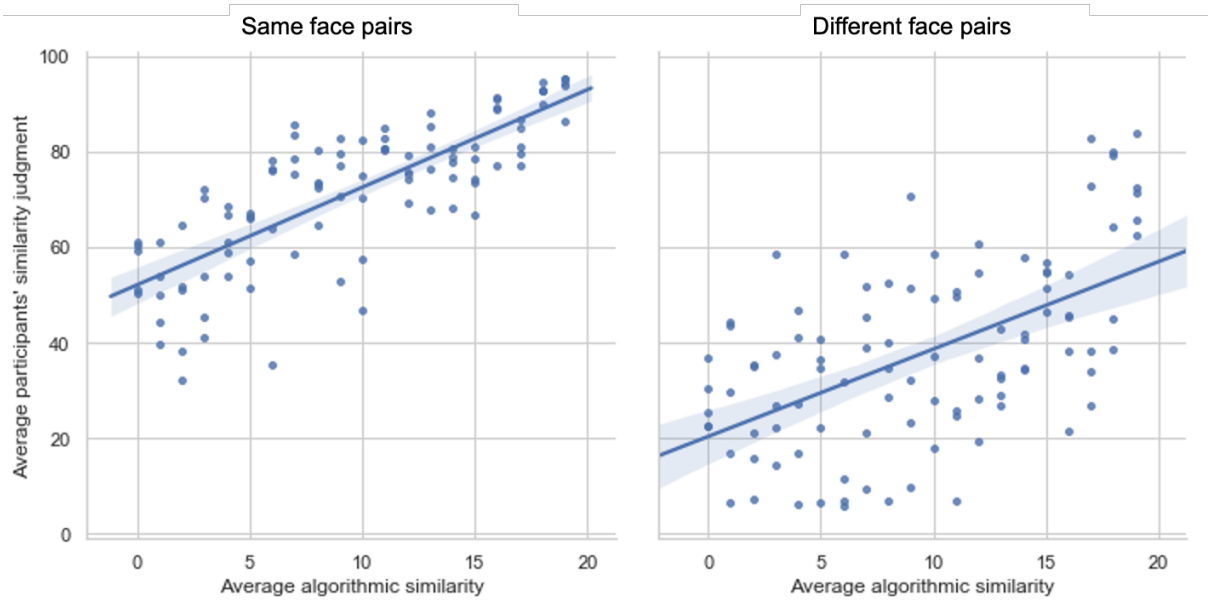


Figure 2.6. Relationship between participants' similarity judgments and average algorithmic similarity values. Here, the average algorithmic similarities are expressed in bins of 1-20 and shown split between same (match, left panel) and different (mismatch, right panel) trials.

2.5.3. Analyses across Experiment 1 and Experiment 2

Given that the design of Experiment 1 (conducted in-person) and Experiment 2 (conducted online) was identical, the following analyses were conducted on data from both studies.

The logic of the OFMT rests on the assumption that similarity will be related to difficulty, such that as objective similarity of face images increases, the difficulty of different pairs increases, and the difficulty of same pairs decreases. The 20 difficulty bins were constructed so that Bin 1 contains the most similar same identity pairs and most different different identity pairs, while Bin 20 contains the most different same pairs and the most similar different pairs. Thus, the trials in Bin 1 are hypothesized to be the easiest and the trials in Bin 20 the most difficult. Evidence for the hypothesised relationship between similarity and

difficulty was obtained by regressing bin number against performance on the data from Experiments 1 and 2. This analysis revealed that bin was a significant predictor of performance ($F(1, 18) = 95.99$, $p < .001$, $R^2 = 0.842$). Performance ranged from a mean of 91.2% (SD = 10.3%) at Bin 1, to a mean of 38.6% (SD = 14.5%) at Bin 20. To confirm the validity of the more difficult bins, we investigated the hypothesis that the variance in accuracy across participants should increase as trial difficulty increases. We analysed the distribution of variance at each of the quintiles of difficulty shown in Figure 2.7. and found that, as expected, Mauchly's test of sphericity indicated unequal variances across quintiles ($\chi2(5) = 13.66$, $p < .05$).

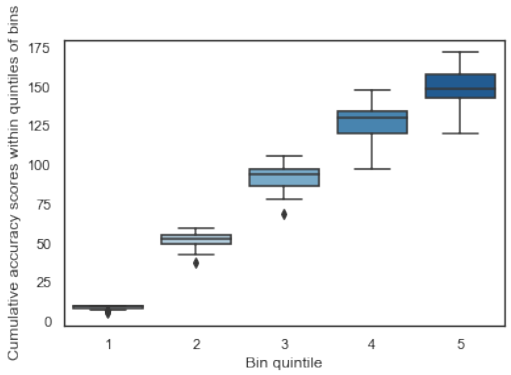


Figure 2.7. Variance in performance by bin quintile.

The increasing variance of neurotypical performance is shown as function of an increase in difficulty (indicated by bin quintile) across the test. The boxes represent the inter-quartile range of cumulative scores across participants for each bin quintile, the lines within boxes show median performance, and whisker-lines represent the extremes of the distribution with outliers indicated by diamonds.

2.5.4. Discussion

Experiment 2, an online replication of Experiment 1, replicated the results of Experiment 1. Both studies demonstrated a wide range of performance on both the Long- and Short-Form of the OFMT, and that although the Short-Form correlated significantly with the Long-

Form, there was only a moderate correlation between the long and short forms (r = .42) when data were collapsed across Studies 1 and 2. Correlations between the OFMT and CFMT were small-to-moderate in both studies. When data across studies were combined, the hypothesised relationship between similarity of face pairs, as assessed objectively by algorithms, and item difficulty was supported by the performance data across bins. Additionally, the inclusion of more difficult items increased variance in scores as would be expected. These data all support the validity of the OFMT.

2.6. Experiment 3

Experiment 3 assessed the test-retest reliability of the OFMT and compared it to three existing measures: the CFMT (Duchaine & Nakayama, 2006a) and the PI-20 (Shah et al., 2015), as widely used objective and subjective measures of face recognition, and the Glasgow Face Matching Task (GFMT; Burton, White & McNeill, 2010). The GFMT was introduced into the testing battery as it is a test of face matching, like the OFMT, without the face memory requirement of the CFMT. To our knowledge, there are only three published studies that have assessed the test-retest reliability of the CFMT. The first tested a large sample (n = 389) of participants with a delay of 6 months and found a correlation of .70 between testing sessions for identical versions of the CFMT and .76 between the original CFMT and a version with computer-generated face stimuli (Wilmer et al., 2010). The second study used an Australian version of the CFMT (a different stimulus set to the original version of the test) and found a correlation of .92 with a 24-hour delay between test sessions (McKone et al., 2011). Finally, a recently published paper found a test-retest reliability of .68 in a group of 70 developmental prosopagnosics tested in the laboratory and online (Murray & Bate, 2020).

2.6.1. Methods

2.6.1.1. Participants

Seventy-two participants with normal or corrected-to-normal vision participated in the study online. One was excluded for failing the attention check. Data for seventy-one participants are reported (43 female, M_{age} = 31.6, *SD* = 7.8 years).

2.6.1.2. Procedure and Tasks

The procedure was exactly the same as in Experiment 2, except that participants completed all tasks twice with 14 days between each completion. Task order was randomized at each testing time.

2.6.2. Results

Test-retest reliability ranged between acceptable and good for all measures: Cambridge Face Memory Test ($r(69) = .67$, $p < .05$), Glasgow Face Matching Task ($r(69) = .77$, $p < .05$), Oxford Face Matching Test ($r(69) = .75$, $p < .05$), and the 20-Item Prosopagnosia Index ($r(69) = .89$, $p < .05$). Across the three objective measures (OFMT, CFMT, and GFMT) test-retest reliability did not significantly differ (GFMT:OFMT: $z = -0.375$, $p > .05$; CFMT:GFMT: $z = -0.121$, $p > .05$; CFMT:OFMT: $z = -0.509$, $p > .05$). Test-retest reliability of the OFMT is therefore as high as the reliability of existing face processing measures.

A significant change in scores between testing days was not observed for the CFMT (T1: *M* = 56.7, *SD* = 10.2, T2: *M* = 55.5, *SD* = 10.6, $t(70) = 1.16$, $p = .25$), GFMT (T1: *M* = 32.5 (81.3%), *SD* = 4.5 (11.25%), T2: *M* = 32.7 (81.8%), *SD* = 5.3 (13.3%), $t(70) = -0.50$, $p = .62$) or the PI-20 (T1: *M* = 44.8, *SD* = 11.3, T2: *M* = 44.5, *SD* = 11.9, $t(70) = 0.41$, $p = .68$).

A small but significant increase in average scores was observed for the OFMT (T1: $M = 152$ (76.0%), $SD = 10.6$ (5.3%), T2: $M = 155.2$ (77.6%), $SD = 12.8$ (6.4%), $t(70) = -3.1$, $p < .05$).

Further, results from Experiment 3 corroborate findings reported in Studies 1 and 2. There was a small-to-moderate relationship between the scores on the CFMT and OFMT at time 1 of testing $r(69) = .41$, $p < .01$) as well as time 2 *(*$r(69) = .38$, $p < .01$). A stronger relationship, as expected, was observed between OFMT and GFMT, the two matching tasks, both at time 1 of testing ($r(69) = .46$, $p < .01$) and time 2 of testing ($r(69) = .59$, $p < .01$). A small, non-significant, relationship was observed between PI-20 and OFMT performance at time 1 ($r(69) = -.22$, $p = .07$) and time 2 *(*$r(69) = -.20$, $p = .11$). Relationships between all measures are shown in Table 2.1.

Finally, a small-to-moderate relationship was observed between average algorithmically-derived similarity values and participants' judgements for different face pairs for both time points (time 1: $r(69) = .29$, $p < .05$, time 2: $r(69) = .32$, $p < .05$) and a moderate relationship for same face pairs for both days (time 1: $r(69) = .49$, $p < .05$; time 2: $r(69) = .47$, $p < .05$).

Table 2.1. Relationships between all face processing measures included in Experiment 3.

Correlations significant at the 0.01 level are denoted with two asterisks (**) and correlations significant at the 0.05 level are denoted with a single asterisk (*). Testing time 1 is denoted by T1 and testing time 2 by T2. Testing times were 14 days apart. Given test-retest reliability data (reported in Experiment 3), correlations between all measures are high, except for correlations between the self-report PI-20 and the objective measures (OFMT, GFMT, and CFMT).

Variable	**CFMT (T1)**	**CFMT (T2)**	**PI-20 (T1)**	**PI-20 (T2)**	**GFMT (T1)**	**GFMT (T2)**	**OFMT (T1)**	**OFMT (T2)**
CFMT (T1)	-							
CFMT (T2)	.67**	-						
PI-20 (T1)	-.22	-.22	-					
PI-20 (T2)	-.16	-.16	.89**	-				
GFMT (T1)	.56**	.38**	-.25*	-.18	-			
GFMT (T2)	.58**	.39**	-.30*	-.27*	.77**	-		
OFMT (T1)	.41**	.41**	-.22	-.25*	.46**	.41**	-	
OFMT (T2)	.42**	.38**	-.18	-.20	.45**	.59**	.75**	-

2.6.3. Discussion

Results of Experiment 3 suggested that the test-retest reliability of the OFMT Long-Form was at least as good as that of the CFMT and GFMT. All tests had acceptable to good reliability, indicating their suitability as measures of individual differences in face processing (under the assumption that face processing ability is a trait that shows little variation across time). With test reliability established, Experiments 4 and 5 aimed to investigate alternative ways of shortening the OFMT.

2.7. Experiment 4

Compared to other face matching measures (GFMT, Burton, White & McNeill, 2010; Pairs Matching Test, Bate et al., 2018; Kent Face Matching Test, Fysh & Bindemann, 2018), the OFMT is longer to administer, taking about 20 minutes (compared to, for example, 3-4 minutes for the GFMT). This is in part due to a larger number of trials, and in part due to the necessity for two responses to be provided. Since face matching measures are often administered as part of a battery of face tests, this administration time may be problematic. A previous attempt to reduce administration time involved reducing the number of trials, but this resulted in relatively noisy estimates of performance (see Short-Form in Experiment 1). Providing researchers do not require the ability to distinguish between face perception and face matching, it might be possible to administer the OFMT without the perceptual similarity question to reduce administration time. However, whether the OFMT can be administered without the question about perceptual similarity, or if its exclusion would fundamentally alter test performance, is an empirical question. Here, therefore, we investigate the consistency of test performance when administered with and without similarity judgments (Experiment 4) and validate the resulting Short-Form OFMT against other face processing measures (Experiment 5).

2.7.1. Methods

2.7.1.1. Procedure

Participants completed the OFMT in two testing sessions 24-48 hours apart, such that one administration of the test required similarity judgments to be made, and the other did not (see Figure 2.2.). The order of testing sessions was counterbalanced across participants.

2.7.1.2. Participants

45 participants ($M \pm SD_{age}$ = 27 ± 9 years, 20 female, 24 male, one other) with normal or corrected-to-normal vision participated in the experiment online completed the 200-trial version of the OFMT online twice.

2.7.2. Results

Performance on the two versions of the test was broadly similar (matching accuracy with similarity judgements: $M \pm SD$ = 70.6 ± 7.5%; without similarity judgements: $M \pm SD$ = 70.8 ± 7.2%). Accuracy data were submitted to a 2 (version: with, without similarity judgements) x 2 (version order: with first, without first) mixed ANOVA. Neither the main effect of version nor the main effect of version order were significant, but an interaction was observed ($F(1,43) = 5.12$, $p = .029$, $n_p^2 = .106$) consistent with a slight practice effect (see Figure 2.8.; cf. Stantić et al., 2022a, Experiment 3). The simple effect of version was not significant for either version order.

The correlation between performance on the two versions of the test was strong, $r(44) = .760$, $p < .001$, and of a similar magnitude to previously reported test-retest reliability correlations for this test (Stantić et al., 2022a). Importantly, test duration was substantially shorter for the version without similarity judgements (*Mdn* = 14 minutes) compared to the version with similarity judgements (*Mdn* = 23 minutes, $Z_{Wilcoxon} = 3.74$, $p < .001$).

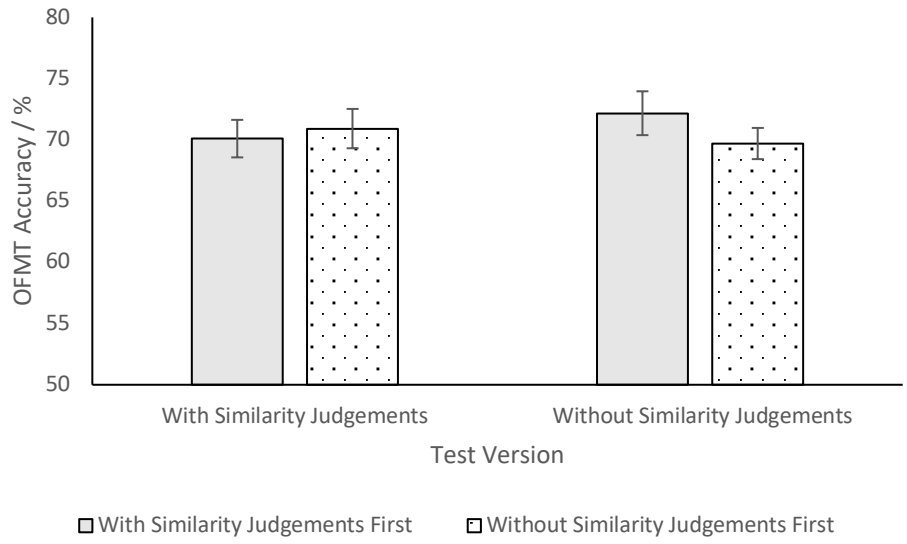


Figure 2.8. Performance on the two versions of the test (with and without similarity judgements). Performance is shown here across the two version orders (version with similarity judgements first, version without similarity judgements first). Error bars indicate standard error of the mean.

2.8. Experiment 5

Experiment 5 sought to validate the developed Short-Form version of the OFMT against other commonly used face processing measures, in a similar manner to the validation of the Long-Form OFMT (Stantić et al., 2022a, see Experiments 1-3).

2.8.1. Methods

2.8.1.1. Procedure

Procedure was as outlined in Experiment 2, but other face processing measures were included as well (GFMT, CFMT, PI-20, see Experiment 2 for task details).

2.8.1.2. Participants

146 participants (*M*±*SD* age = 19.8 ± 2.3 years, 134 female, 10 male, 2 other) with normal or corrected-to-normal vision completed all tasks in a counterbalanced order in one in-person testing session.

2.8.2. Results

Performance on the four measures is presented in Table 2.2.

Table 2.2. Performance on face processing measures in Experiment 5.
Note that scores on OFMT and GFMT represent percentage accuracy.

Measure	Mean	SD	Minimum	Maximum
OFMT Short Form	75.8	5.8	58.5	87.5
CFMT	52.9	9.4	30	70
GFMT	82.0	11.7	52.5	100.0
PI-20	40.1	9.7	21	71

Simple correlations were performed between scores on the OFMT Short Form and each of the other face processing measures. As measures other than the OFMT were not normally distributed, Spearman's ρ was used. Significant correlations were found between the OFMT Short Form and GFMT: $\rho_{144} = .263$, $p = .001$; CFMT: $\rho_{144} = .463$, $p < .001$; and PI-20: $\rho_{144} = -.305$, $p < .001$.

2.9. General Discussion

Tests of facial identity recognition are difficult to calibrate for the entire range of performance (from individuals with DP, through the normal range of performance, to individuals who are classified as SRs). In order to be sensitive to individual differences across a wide range of performance, test items need to be of varying difficulty – with easier items allowing sensitivity to individual differences within the DP and lower end of typical performance range, and more challenging items allowing sensitivity within the SR and higher end of typical performance range. A potential problem arises, however, when assessing difficulty. Given that face processing is a task that may be carried out using a number of different strategies – and the propensity to use different strategies may vary across groups of interest such as DPs, SRs, or individuals with Autism Spectrum Disorder – item difficulty should be established in a way that is not biased towards one particular population. The solution adopted here was to use algorithmic judgements of similarities of face pairs to construct items of a range of difficulties. These stimuli were used to develop the OFMT, a test of face perception.

The reliability and validity of the OFMT was assessed across five studies. The range of performance observed in neurotypical participants in Experiment 1 suggested that the OFMT is sensitive to individual differences within this population, with no participant performing at chance levels, or scoring 100% accuracy. The OFMT demonstrated a small-to-moderate correlation with the current gold-standard test of face recognition, the CFMT. This degree of correlation is encouraging, as the CFMT is thought to measure both face perception and memory, whereas the OFMT is thought to measure face perception only. One would therefore expect a modest degree of correlation between performance on the two measures, but not very large correlations. A correlation of similar size was observed between

the Long- and Short-Form of the OFMT, suggesting that the Short-Form currently provides only a very noisy indicator of Long-Form performance. The degree of correlation suggests further work is required to make the Short-Form a valid proxy of Long-Form performance, to which we suggest some solutions can be provided by omitting the question about face similarity per results of the Experiments 4 and 5. Finally, Experiment 1 established that algorithmic similarity judgements were in accordance with judgements provided by neurotypical participants, and that correlations were of greater magnitude for face images taken from the same person compared to face images of different individuals, replicating previous findings (Hancock, Somai, & Mileva, 2020). The size of these correlations was moderate, supporting the contention that the algorithms are not simply mimicking the strategy used by neurotypical participants when judging similarity. This latter result supports the basic premise behind using algorithmic ratings rather than ratings provided by a large group of neurotypical raters. The findings of Experiment 2 replicated those of Experiment 1, despite Experiment 2 being administered online rather than in the laboratory like Experiment 1. This replication suggests that the OFMT is suitable for remote administration and scalable to large numbers of participants. In addition, Experiment 2 supported the hypothesized relationship between the similarity of face images and difficulty of items.

It should be noted, however, that it is an oversimplification to describe the OFMT as a pure measure of face perception and the CFMT as a measure of face perception plus face memory. Both tests also require the ability to match two instances of an individual's face, or to recognise the mismatch between faces of different individuals. This face matching stage must occur following the construction of a perceptual representation of the face. In practice, face matching is required by both tests, but the face matching requirements of the OFMT are greater than those of the CFMT as the potential differences between images of the same

individual's face, and the similarity of different individuals' faces, are likely greater in the OFMT than CFMT. Here, face matching ability refers to the accuracy of representation 1) of the degree of difference, and 2) the kinds of difference permissible, between two facial images before they must be judged as faces of different individuals.

Finally, Experiment 3 assessed the reliability of the OFMT and compared it to two other standard tests of face processing, the CFMT and GFMT. Reliability data on these tests is surprisingly scarce, given the age of the tests and their extensive use, but data from Experiment 3 indicated that the reliability of the OFMT is good, and statistically indistinguishable from the CFMT and GFMT. Thus, Studies 1-3 suggest the OFMT is a reliable measure of face perception. Validity of the measure will further be explored in chapters 4, 5, and 6 through testing of atypical populations (groups of people with developmental prosopagnosia, autism and super recognition).

In Experiments 4 and 5, we sought to investigate whether the OFMT can be substantially shortened for ease of administration. The results indicate that matching performance remains consistent regardless of whether the similarity question is asked. Eliminating the similarity question substantially reduces the length of the full test, and the administration of the 200 trials of the OFMT takes, on average, less than 14 minutes without this question. Importantly, the relationships between performance on the short-form OFMT and performance on other face processing measures were consistent with those reported for the long-form OFMT in the initial validation study (Stantić et al., 2022a), with small to moderate relationships observed with all measures. This is also consistent with the magnitude of relationships reported between other face processing tests (e.g., Verhallen et al., 2017).

This finding has practical significance for all researchers interested in using the OFMT. It shows that face matching performance can be measured using the OFMT without explicitly asking participants to estimate the perceptual similarity of presented faces. It is therefore expected that indicative OFMT scores across ages and genders (Stantić et al., 2021a) can be used as guides for performance with or without similarity judgments. It should, however, be noted that we did not test neurodiverse populations (e.g., super-recognisers, developmental prosopagnosics, autistic participants) using the shortened version of the OFMT, and it is possible that such populations might show different patterns of results to the present sample as well as different patterns of results to those reported in Chapters 4, 5 and 6, which outline the results of deploying the Long-Form OFMT in those populations.

The key distinction between the OFMT and existing tests of face processing is the use of facial recognition algorithms to obtain an objective measure of face similarity. This has a number of advantages, most notably that similarity values are not biased against particular sub-populations, who may use different strategies when recognising faces to the neurotypical majority. A further advantage is practical: the use of algorithmic similarity judgements means that, provided adequate stimulus material exists, multiple parallel forms of the OFMT can be generated to allow repeated testing with independent stimuli. With appropriate selection of stimuli for algorithmic similarity rating, these parallel forms could be specific to processing of faces of certain ethnicities, ages, or genders.

At least two potential disadvantages exist, however. The first is that any particular algorithm may use a unique strategy to assess similarity, meaning that algorithmic similarity ratings

do not necessarily match those produced by any group of human raters. Data reported here show a moderate correlation between algorithmic ratings and human raters, potentially because three different algorithms were used to obtain algorithmic similarity ratings, which were then averaged, and stimulus pairs which produced discrepant ratings among the algorithms were not included. Therefore, while the algorithmic values used in the OFMT appear to have provided a valid, as well as objective, measure of similarity, separate processes are likely to contribute to these values across different algorithms and human perceivers. The second potential disadvantage is that algorithms may be systematically worse at judging the similarities of particular groups of faces (e.g., faces of particular gender, ethnicity, age, or their interaction). Despite numerous press assertions that this may be so, official testing by the US National Institute of Standards and Technology suggests that the 17 most accurate algorithms had similar levels of accuracy across demographic groups (McLaughlin & Castro, 2020). However, the assumption of non-biased performance by stimulus class is one that should be continuously scrutinised as facial recognition algorithms develop.

Another feature of the OFMT that distinguishes it from most other tests of face processing is that relatively naturalistic face images are used. While static, faces in the OFMT are shown with hair and without cropping any information about the face shape. Difficulty on more challenging trials is not introduced through introduction of artificial Gaussian noise or partial occlusion of facial information. Rather, difficulty is increased in ways in which it occurs in everyday life – by matching facial images of the same person when their appearance has changed substantially, or by distinguishing similar-looking people from one another. The OFMT is therefore arguably more ecologically valid than existing tasks of face recognition,

with clear applications to real-life situations such as identity verification and forensic contexts.

In summary, this paper describes the development and validation of the OFMT, a non-biased test of face perception, which is suitable for use across the full range of face perception ability. Additionally, while some researchers (for example, those interested in dissociating face perception and face matching) might find the deployment of the Long-Form OFMT with similarity judgments more suitable for their purposes, these results show that its use without similarity judgments significantly shortens the length of the test without affecting matching performance. The OFMT is freely available to researchers on the Gorilla Open Materials repository (https://gorilla.sc/openmaterials) for non-commercial use.

Chapter 3: Individual Differences in Face Matching and Memory in Neurotypical Population

An edited or partial version of this chapter is currently under review as the following article:

- Stantić, M., Brown, K. Catmur, C. & Bird, G. (under review). Independent Measurement of Face Perception, Face Matching, and Face Memory Reveals Impairments in Face Perception and Memory, but Not Matching, in Autism.

3.1. Abstract

There are multiple psychological processes required in order for a face to be recognised. However, when testing face memory using tasks such as the Cambridge Face Memory Task , it is rare for studies to attempt to account for individual differences in face perception and face matching in order to isolate variance in face memory specifically. Here the Oxford Face Matching Test was used to obtain an independent measure of face matching in addition to face perception in a large sample of participants (N=1,112). Results revealed independent contributions of face perception and matching to face memory, and these results replicated when the Glasgow Face Matching Test was used as a test of face matching.

3. 2. Introduction

As outlined in the introduction, psychological models of the processes by which we recognize our conspecifics from their faces acknowledge that face recognition can be broken down into a number of sub-processes (Bruce & Young, 1986; Haxby et al., 2000). For example, in the classic model by Bruce and Young (1986), recall of an individual's name upon viewing a photograph of their face requires a number of steps. Initially, a perceptual representation of a face is formed which consists both of view-dependent and more abstract perceptual representations of the face. These representations activate face recognition units – representations of known faces - according to the degree of similarity between the face in the photograph and the stored face representations. In turn, face recognition units activate person identity nodes which contain identity-specific semantic codes (information such as the person's occupation, where they live, etc.), which go on to activate the appropriate name code for the presented face. Bruce and Young (1986) also highlighted the importance of decision processes in their model, especially the ability to determine whether a specific instance of a face is an instance of a particular identity's face. This process is needed when deciding whether two photographs of a face depict the same person, and when deciding whether the face of a person matches any of the facial identities one has previously learned.

Turning from theory to empirical methods, measures of face processing can be split into three types based on the specific aspects of face processing that they aim to measure. The first category contains tasks assessing face detection or face attention – the ability (or propensity) to identify, and to attend to, faces among distractors. The second category contains tasks assessing face perception, which are generally described as assessing the degree to which a veridical 3-D perceptual representation of a face can be derived from

sometimes limited information (e.g., a 2-D image from one viewpoint only). In practice, such tasks are commonly (but not exclusively) tests of face matching, in which two facial images are presented concurrently or in quick succession, and participants are required to judge whether the facial images are of the same individual, or different individuals. The third category contains face recognition tests, with individuals required to recognize famous faces, or faces they have previously been exposed to.

With respect to the third category, although face recognition tests are of use when identifying individuals who exhibit severe difficulties with face recognition (individuals with prosopagnosia), they are not 'process pure'. If an individual fails to recognize famous faces, or faces they have previously seen, they may have difficulties with building a perceptual representation of the face, with determining whether a stimulus face matches a face stored in memory, or with face memory (storage and retrieval of face data) specifically. Face matching tasks are also not process pure; an individual may fail to recognize two facial images are of the same individual because they cannot form an accurate perceptual representation of the faces, or because they are poor at *deciding* whether the two faces are of the same individual or of different individuals, from accurate perceptual representations. Ideally, one would obtain independent measures of face perception, face matching, and face memory, and determine the degree to which individual differences in face perception and face matching explain variance in face matching. It is this that the present study aims to accomplish by making use of a new test of face matching, the Oxford Face Matching Test (OFMT; Stantić et al., 2022a).

The unique aspects of the OFMT have been outlined in Chapter 1 already. Most importantly to the question at hand, however, participants provide rating of face similarity

can be compared to the algorithmically-derived similarity rating for the same face pair. One could be able to judge face pair similarity consistently with algorithms but be inaccurate at determining whether the face pair is of the same person or different people; alternatively, they may have unusual perception and judge face pair similarity inconsistently with algorithms yet be accurate at determining whether face pairs are of the same person or different people. This unique feature of the OFMT allows for independent measurements of face perception and matching to be obtained, alongside face memory measures (CFMT; Duchaine & Nakayama, 2006a).

3.3. Methods

3.3.1. Procedure

The procedure was as outlined in Chapter 2 (2.3. Common Methodology).

3.3.2. Participants

1,189 adults took part in the experiment online. 77 were excluded based on predefined exclusion criteria (30 for failing attention checks, 23 for not finishing all tasks, 4 for data corruption, and 20 for scoring more than 3 standard deviations below the group mean on any task). Data are reported from the remaining sample of 1,112 participants (743 female, M_{age} = 37.60 years, SD = 18.98). Participants were recruited from a variety of sources, including physical and social media advertisements, and via Prolific.co.

3.3.3. Analysis strategy

In order to address the aims of the study, independent measures of face perception, face matching and face memory are required. To derive a measure of face perception, participants' ratings of the similarity of face pairs on the OFMT were compared to the algorithmically-derived similarity ratings. Specifically, for each participant, an average absolute deviation from algorithmic similarity was calculated. For each participant, this average absolute deviation from algorithmic similarity was calculated across 200 test trials.

An index of face matching independent of face perception was derived by regressing OFMT accuracy scores on average algorithmic deviations across participants. The residuals from this regression constitute face matching scores as they represent the variance in OFMT accuracy (i.e., in same/different judgements) that cannot be explained by face perceptual similarity judgements. The face matching scores are thus statistically independent from face perceptual similarity judgements due to the way they are derived. Note that face matching scores can also be derived from GFMT scores (Burton et al., 2010), if GFMT scores are regressed on average algorithmic deviations obtained from the OFMT. This double coding of face matching ability provides a data robustness check – allowing the replicability of results to be ascertained when using matching scores from the GFMT after controlling for OFMT perceptual similarity judgements.

Note that this robustness check stems from the independence of the two judgments made on each trial of the OFMT. The binary same/different decision about two faces, akin to the one made on each trial of the GFMT, constitutes an independent data point from the judgment of perceptual similarity of the two presented faces. The latter can therefore be

used as an index of participants' perceptual ability independent of their face matching ability.

In the current study, we use these residuals (i.e., the measure of face matching independent of face perception accuracy) in univariate analyses, e.g., in post-hoc tests. In multivariate analyses, entering OFMT or GFMT test scores and algorithmic deviations into the same analysis achieves the same aim (the variance explained by face matching performance can be identified independently of that accounted for by face perception - i.e., when face perception ability is held constant).

Finally, face *memory* scores that are independent of both face perception and face matching can be obtained by regressing CFMT accuracy scores on average algorithmic deviations from the OFMT and face matching test scores from either the OFMT or GFMT.

The exact same analysis strategy was replicated in Chapters 4, 5, and 6. In addition to the factors outlined above, Group was also included as a predictor in those chapters (e.g., DP or SR or ASD).

3.4. Results

3.4.1. Group differences

Descriptive statistics for all measures are shown in Table 3.3. The distribution of scores across all participants are shown in Figure 3.9. for all three tests, along with correlations between test scores.

Table 3.3. Descriptive statistics for all three face perception measures.

	n	*M*	*SD*	Range
OFMT	1112	155.44	13.36	107-190
CFMT	1112	52.34	10.44	15-72
GFMT	1112	33.22	4.98	6-40

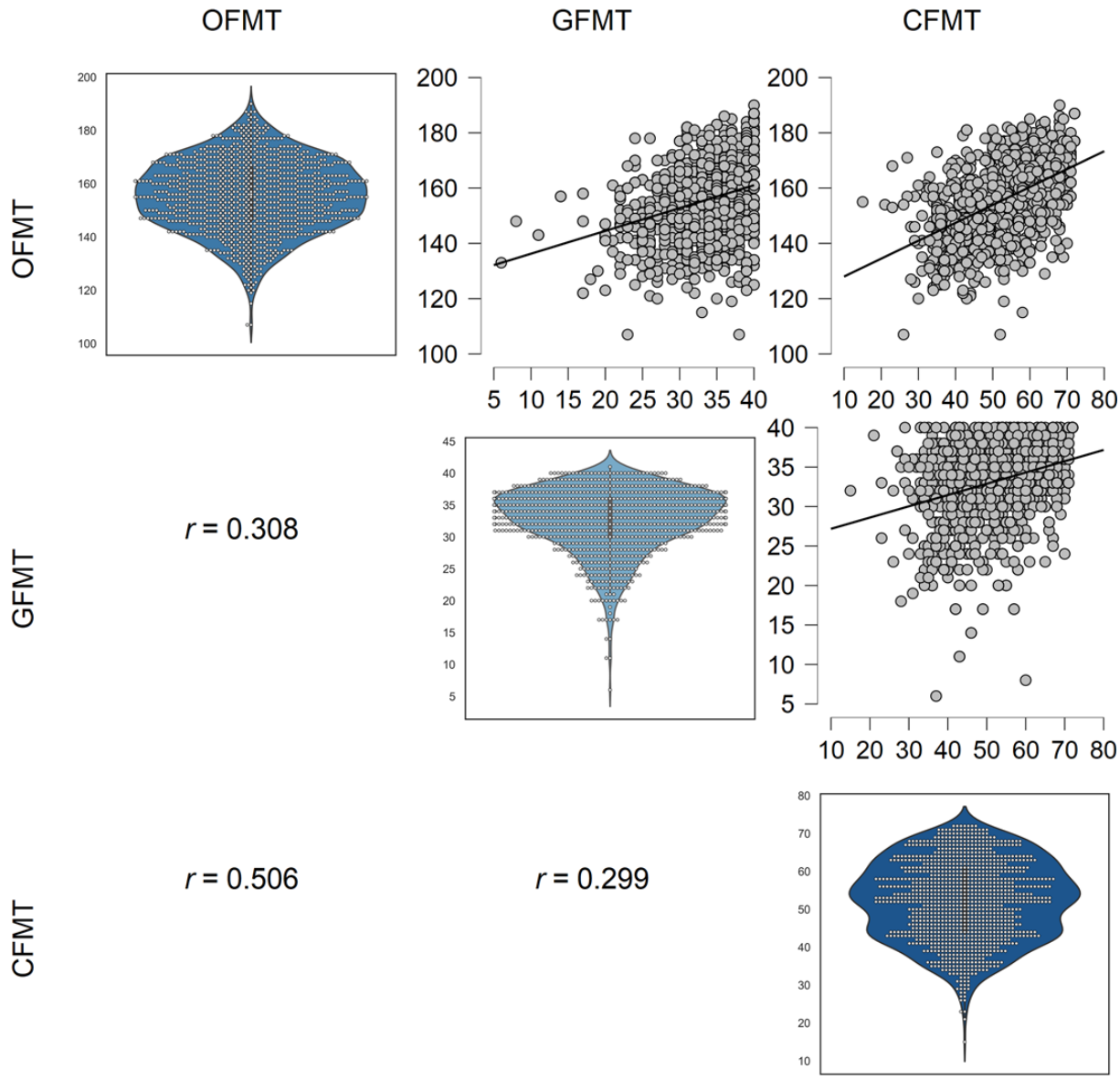


Figure 3.9. Correlations between all measures of face processing, and distributions of scores for individual tasks.

All correlations are significant at p < .001.

3.4.2. Effect of Face Perception on Face Matching

Face perception ability, as measured by deviation scores from algorithmic similarity judgements, was found to be a significant predictor of OFMT scores (β = -.611, t = 25.71, p < .001), such that individuals whose similarity estimates deviated more from the algorithms' judgement were also worse at judging whether two images were of the same face, or different faces, on the OFMT. See Figure 3.10. (A, left panel).

As a robustness check, the same analysis was performed on GFMT scores and yielded the same pattern of results. Deviation scores were found to be a significant predictor of GFMT scores (β = -.284, t = 9.88, p < .001), such that individuals whose estimates deviated more from the algorithms' judgement on the OFMT were also worse at judging whether two images were of the same face, or different faces, on the GFMT. See Figure 3.10. (A, middle panel).

3.4.2. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory

CFMT scores (indexing face memory) were regressed onto face perception scores (deviation scores) and face matching scores derived from the OFMT as described in the 'Analysis Strategy' section. Note that this analysis assigns shared variance between face perception and face matching to face perception – consistent with a theoretical model in which face matching requires face perception but not vice versa – but the same pattern of significance was obtained when raw deviation and OFMT scores were entered into the model simultaneously (and thus shared variance assigned to neither predictor). A significant overall model accounted for 27.3% of the variance in CFMT scores, F(2, 1109)

= 208.5, $p < .001$. Deviation scores were found to be a significant predictor of CFMT scores ($\beta = -.413$, $t = 16.13$, $p < .001$), such that individuals whose estimates deviated more from the algorithms' judgement had worse face memory as measured by the CFMT (see Figure 3.10., A, right panel). Face matching was also a significant independent predictor of CFMT scores ($\beta = .321$, $t = 12.53$, $p < .001$), such that those with a better ability to determine whether two faces are the same or different had better face memory as measured by the CFMT (see Figure 3.10., B, left panel). Again, as a robustness check face matching scores derived from the OFMT were replaced with face matching scores derived from the GFMT and the same pattern of significance was observed (Deviations: ($\beta = -.413$, $t = 15.43$, $p < .001$; Face Matching: $\beta = .190$, $t = 7.01$, $p < .001$; Figure 3.10., B, middle panel).

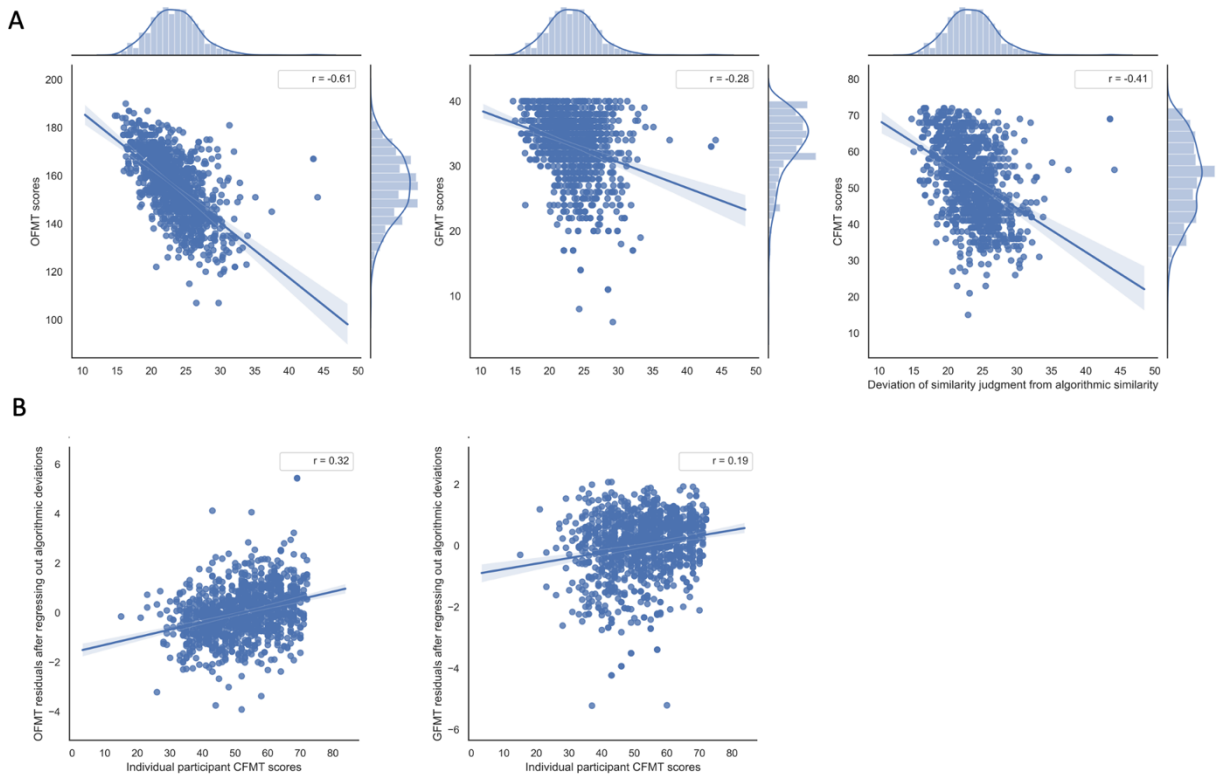


Figure 3.10. Linear regressions showing relationships between the deviation of individual participants' similarity judgments and the face matching and memory tasks.

Deviations of participants' similarity judgments are averaged across 200 trials of the OFMT from algorithmic similarity values. A: Left panel – relationship with the OFMT scores, middle panel – relationship with the GFMT scores, right panel – relationship with the CFMT scores. B: Left panel - Relationship between residual OFMT scores following regression of face perception scores and CFMT scores, middle panel - Relationship between residual GFMT scores following regression of face perception scores and CFMT scores.

3.5. Discussion

In order to successfully recognise a previously learnt face, one must retrieve a stored representation of that face, assess the degree of similarity between the stimulus face and the stored face representation, and then decide whether the degree of similarity is such that

the stimulus and stored faces match, requiring accurate perceptual representations of the faces (Bruce & Young, 1986; Haxby et al., 2000; Valentine, 1991). Here, we attempted to obtain independent measures of face perception and face matching, and test their hypothesised relationship to face memory.

Face perception was assessed by comparing participants' judgements of the similarity of two faces with a consensus decision derived from three leading face recognition algorithms. This measure of face perception predicted face matching performance (i.e., the accuracy of the decision as to whether two images were of the same face or different faces), and it did so on two independent tests of face matching, the OFMT and GFMT. Face perception also predicted performance on the CFMT, a test of face memory. Face matching, measured independently of face perception, also predicted face memory, irrespective of whether the OFMT or GFMT was used to derive the measure of face matching. These results are consistent with the hypothesised role of both face perception and face matching in face memory.

Use of the OFMT and the analysis strategy adopted allowed a measure of face matching ability independent of face perception to be obtained. In addition to face perception and face memory, this allows a further feature of individual differences in face processing to be measured. All three metrics can be used to assess the face processing abilities of different groups such as older adults (Stantić et al., 2021a) or individuals with atypical abilities, which is explored in the following three chapters. Specifically, in groups of participants with atypical performance, such as those with autism, developmental prosopagnosia or super recognition, implementation of this strategy could reveal information about their atypicalities that could help us better understand the mechanisms of intact face processing.

For instance, face memory has been hypothesized to be consistently affected in DP, largely because DP is diagnosed using a gold-standard measure in face memory, the CFMT. However, there has also been debate in the literature about the subtypes of prosopagnosia and suggestions that subtypes could be explained by different patterns of performance on perception and memory tasks (Dalyrmple, Garrido, & Duchaine, 2014).

Adopting a strategy as outlined in this paper in atypical populations could shed light on the hypothesized subtypes by highlighting whether there are subgroups of prosopagnosics who struggle at different stages of processing, all of which might present as differences in face memory. For instance, it is possible that either face perception or matching (or both) are affected in prosopagnosia, and that it is this difficulty that gives rise to a difficulty in face memory. Attempts to better integrate DPs through training programs (DeGutis, Cohan, & Nakayama, 2014) would benefit from a better understanding of cognitive clusters, if any.

Overall, these findings in a large sample of healthy neurotypical participants provide an interesting starting point for avenues of future research and more granular investigations of mechanisms of typical perception as well as understanding atypical groups. They also highlight the value of using facial recognition algorithms to construct face perception tests.

Chapter 4: Face perception and memory in developmental prosopagnosia

An edited or partial version of this chapter has been published as the following journal articles:

- Stantić, M., Pounder, Z., Bate, S., Susilo, T., Catmur, C., Bird, G.P. Individuals with developmental prosopagnosia show independent impairments in face perception, face memory and face matching. *Cortex*, *157*, 266-273.

- Stantić, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., Catmur, C., & Bird, G. (2022). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavior research methods*, 54(1), 158-173.

4.1. Abstract

Individuals with developmental prosopagnosia (DP) all exhibit impairments in face memory, but the specificity of these face memory impairments is debated. One problem is that standard behavioural tasks are not able to provide independent measurement of face perception, face memory, and face matching (the decision process required to judge whether two instances of a face are of the same individual or different individuals). The present study utilised a new test of face matching, the Oxford Face Matching Test (OFMT), and a novel analysis strategy to derive these independent indices. Twenty-nine individuals with DP and the same number of matched neurotypical controls completed the OFMT, the Glasgow Face Matching Test, and the Cambridge Face Memory Test. Results revealed individuals with DP exhibit impairments in face perception, face memory *and* face matching. Collectively, these results suggest that face processing impairments in DP are more comprehensive than has previously been suggested.

4.2. Introduction

Developmental prosopagnosia (DP) is a neurodevelopmental condition characterised by a lifelong inability to recognize faces (e.g., Behrmann & Avidan, 2005; Cook & Biotti, 2016; Susilo & Duchaine, 2013). While DP is, by definition, associated with problems remembering faces, the particular aspect of face processing responsible for these memory problems is unclear. Specifically, there is an ongoing debate as to whether individuals with DP are able to form intact perceptual representations of faces but have difficulty learning/recalling facial identities ('*memory hypothesis*'; Jackson, Counter & Tree, 2017; Stollhoff, Jost, Elze & Kennerknecht, 2011), or whether individuals with DP have difficulties with both forming perceptual representations of faces and face memory ('*perceptual hypothesis*'; Biotti, Gray, & Cook, 2019; Dalrymple, Garrido, & Duchaine, 2014, see also Towler, Fisher, & Eimer, 2018). This debate is ongoing partly due to the heterogeneity of DP, specifically the possibility that there may be subtypes of DP (e.g., Dalrymple et al., 2012) or variants (e.g., apperceptive or associative) that differ in terms of the nature of their face-processing impairment (Biotti et al., 2019; Corrow, Dalrymple, & Barton, 2016), raising the possibility of multiple subtypes of DP characterized by different types of impairments.

This chapter outlines an approach that could help differentiate between them. If face perception only is impaired in DP, then one would expect any impairment in performance on the CFMT in DPs to be fully explained by performance on the OFMT. If only face memory is impaired in DP, then one would expect DP to produce a selective deficit on the CFMT and not OFMT, while if both face perception and face memory are impaired then one would expect deficits on both tasks, but that the deficit on the CFMT would remain present after performance on the OFMT is accounted for.

So far, little attention has been paid to face matching in DP, which is also necessary for face recognition. It is worth noting that the term 'face matching' is used here in a very specific, and perhaps unusual, way. In contrast to the standard use of the term face matching, which refers to a set of tasks in which participants are required to judge whether two photographs of a face depict the same individual or different individuals, face matching in this context relates to the decision-making process necessary to determine whether two or more face images are of the same individual, or different individuals. Face matching is the psychological process necessary for successful performance on face matching tasks. Face matching is also required, however, when a participant is asked to determine whether a face stimulus matches that stored in memory, such as when deciding whether a photograph of a face matches a recently-learned identity. Although the former task is called a face matching task, and the latter called a face memory or face recognition task, the decision-making process (deciding whether two face instances are of the same individual or different individuals, regardless of whether one of those instances is stored in memory) is the same, and what we call face matching.

Problematically, existing tests are largely unable to isolate these distinct face processes. For example, whether individuals with DP are impaired at face memory is normally tested using the CFMT (Duchaine & Nakayama, 2006a), a task in which participants have to identify a pre-learned facial identity among distractors. Although face memory is required in order to perform accurately on this task, face perception and face matching are also required. Therefore, it is unclear whether poor performance on the CFMT reflects impaired face memory, face perception or face matching, or any combination of these processes. Similarly, face perception is often assessed using face matching tasks such as the GFMT

(Burton, White, & McNeill, 2010; see also Kent Facial Matching Test; Fysh & Bindemann, 2018), in which two facial images are presented and participants are required to judge whether the facial images are of the same person or different people. Matching tasks require intact face perception and face matching, and so impaired performance on matching tasks in DP may reflect either one, or both, of impaired face perception and face matching. When face matching tasks are not used to assess face perception, the Cambridge Face Perception Task (CFPT; Duchaine, Germine, & Nakayama, 2007) is often used. This test requires participants to order a set of test face stimuli in terms of their similarity to a target face, where those test stimuli are morphed faces containing varying proportions of the target face and a foil face. The contribution of face matching to performance on the CFPT is therefore hard to determine, as increasingly similar faces are achieved by the test face becoming increasingly the same face as the target face, due to the test face containing objectively more of the target face in the morph. Conversely, decreasing similarity is achieved by the test face containing objectively less of the target face. Thus, responses of a participant basing their judgement on objective face perceptual similarity only, and a participant basing their responses on the outcome of a face matching decision process, would be perfectly correlated. Ideally, perceptual similarity would be unconfounded from facial identity, so that facial images of the same person are sometimes less similar (due to changes in appearance due to ageing, weight gain, etc.) than facial images of different people. What has been lacking thus far, therefore, is a means to obtain independent measures of face perception, face matching, and face memory, in order to determine the degree to which these are affected in DP.

The Oxford Face Matching Test (OFMT; Stantić et al., 2022a) is a novel test designed so that it can be used to assess individual differences in face processing abilities in clinical

and non-clinical populations in a non-biased manner (Stantić et al., 2022b; Stantić, Brown, Catmur & Bird, *under review*). Of relevance to the current study is that independent measures of face perception and face matching can be obtained from the OFMT, and when these scores are used to partition variance in CFMT scores, an independent measure of face memory can also be derived. In Chapter 2, this approach was used to show that, in neurotypical individuals, face perception contributes to performance on face matching tests, and that face perception and face matching make independent contributions to CFMT performance. In addition, when this approach was used with volunteers with autism[2], results suggested that autism was associated with deficits in face perception and face memory, but not face matching (Stantić et al., *under review,* see Chapter 6).

Accordingly, the current study uses the OFMT, GFMT and CFMT to achieve three aims:

1. To validate the use of OFMT across the full performance spectrum and examine individual differences in a group of developmental prosopagnosics (DP)
2. To distinguish between different types of face processing deficits in DP, testing the perceptual vs memory hypothesis of prosopagnosia
3. To derive independent measures of face perception, face matching, and face memory in a group of DPs and a matched neurotypical control group so that the nature of face processing impairment(s) in DP can be identified.

[2] To respect the wishes of autistic individuals and report the study in line with scientific parlance, we use language preferred by clinical professionals (e.g., 'individuals with autism'), as well as the term 'autistic', a term endorsed by many individuals with ASD (see Kenny et al., 2016).

4.3. Methods

4.3.1. Participants

Twenty-nine developmental prosopagnosics were recruited to participate in the study (9 male; M_{age} = 42.31, *SD* = 11.31). Participants were selected from pre-existing author databases of individuals with DP. These participants met the criteria for impaired performance (defined as 2 SDs below the neurotypical mean score) on at least two of three face processing measures (CFMT, CFPT, and the Famous Face Test; (Bobak et al., 2017; Duchaine & Nakayama, 2006a; Duchaine, Germine, & Nakayama, 2007; following primary criteria outlined in Barton & Corrow, 2016). Two of the participants did not show impaired performance on the Famous Faces Test, but their PI-20 scores (79 and 88, respectively) indicated a high degree of issues with face recognition in everyday life. As these participants were not outliers on any measures, they were included in the analysis. All DP participants scored higher than 70 on the PI-20. No DP participants were excluded from the current study for failing the attention check trials on the OFMT (see below). An age- and gender-matched sample of 31 neurotypical participants was recruited via Prolific.co and the authors' database. Two participants were excluded for failing to pass attention checks on the OFMT, providing a final sample of 29 neurotypical participants (11 male; M_{age} = 42.41, *SD* = 9.81). All participants reported having normal or corrected-to-normal vision.

4.3.2. Procedure

The procedure was as outlined in Chapter 2 (2.3. Common Methodology).

4.3.3. Analysis strategy

The analysis strategy was as outlined in Chapter 3 (3.3.3. Analysis Strategy).

Further, we determined whether DP impacts face perception, face matching (independent of face perception) and face memory (independent of face perception and face matching) using a series of regression analyses in which group (DP vs Control), predictors (e.g., face perception), and their interaction, were used to predict test scores (e.g., CFMT test scores). Including the interaction term in regression models allows for the relationship between, e.g., face perception (i.e., algorithmic deviation) and CFMT scores, to vary across groups.

4.4. Results

4.4.1. Group differences

OFMT: Matching Performance

Nine participants with DP (31%) performed two standard deviations below the Control group's mean score, and an additional 13 participants with DP (45%) performed one standard deviation below the Control group mean. One DP had an OFMT score that was greater than the neurotypical mean score (but less than one standard deviation above the control mean). At the group level, an independent t-test showed a significant difference in OFMT accuracy [$t(56) = 5.90$, $p < .001$, $d = 1.55$] with DPs less accurate ($M = 66\%$, $SD = 6\%$, range: 55%-78%) than control participants ($M = 74\%$, $SD = 5\%$, range: 65% - 84%).

Signal Detection Theory (Green & Swets, 1966) was used to characterise performance on the matching component of the OFMT, providing a measure of sensitivity (d') and bias

(criterion). An independent t-test showed a significant difference in d' [$t(56) = 5.40$, $p < .001$, $d = 1.42$] and criterion [$t(56) = 3.21$, $p = .002$, $d = 0.84$] values between the DP and Control groups, with the DP group showing less sensitivity and an increased bias (DP d': $M = 0.91$, $SD = 0.31$; criterion: $M = -0.36$, $SD = 0.33$; Control d': $M = 1.38$, $SD = 0.35$; criterion: $M = -0.06$, $SD =$. 0.38). Thus, individuals with DP exhibited less sensitivity to signals relevant to face matching, and needed more evidence that the faces were the same before they made this response.

To address the perceptual and memory hypotheses, hierarchical multiple regression was used whereby OFMT scores were entered into step 1 of a model predicting CFMT scores, and then Group (DP vs Control) entered into Step 2. The change in R^2 from Step 1 to Step 2 (.16) was significant ($F(1,57) = 28.94$, $p < .001$), meaning that variance in CFMT scores could be accounted for by Group over and above that explained by OFMT scores.

OFMT: Algorithmic Deviation

Nine participants with DP (31%) performed two standard deviations above (indicating worse performance) the Control mean deviation score, while 11 participants with DP (38%) performed one standard deviation above the Control mean. Eight participants with DP (28%) had a deviation score that was greater than the neurotypical deviation mean (but less than one standard deviation). At group level, a Mann-Whitney test revealed a significant difference in deviation scores ($U = 105$, $p = < .001$, $r = .64$) with the DP group having worse deviation scores, i.e., higher deviation from algorithmic judgements (median = 26.69, range: 22.81 – 48.28), than the Control group (median = 23.57, range: 16.66 – 27.38).

GFMT

Nine participants with DP (31%) performed two standard deviations below the Control mean score, while 11 DPs (38%) performed one standard deviation below the Control mean. Five participants with DPs (17%) had a GFMT score that was greater than the neurotypical mean score (but less than one standard deviation above the control mean). At group level, an independent t-test showed a significant difference in GFMT accuracy [$t(56) = 5.14$, $p < .001$, $d = 1.35$], with DPs less accurate ($M = 26.59$, $SD = 4.78$, range: 19 – 35) than Control participants ($M = 32.83$, $SD = 4.46$, range: 23 – 39).

As with the OFMT, Signal Detection Theory (Green & Swets, 1966) was used to characterise performance on the matching component of the GFMT. An independent t-test showed a significant difference in d' [$t(56) = 4.69$, $p <.001$, $d = 1.23$] with individuals with DP less sensitive to face matching signals ($M = 0.98$, $SD = 0.83$) than Control participants ($M = 2.10$, $SD = 0.98$). A Mann-Whitney test revealed a significant difference in criterion scores ($U = 254$, $p = .01$, $r = .34$; DP: median: -0.14, range: -1.46 – 2.49; Control: median: 0, range: -0.49 – 1.24). Thus, as on the OFMT, individuals were less sensitive to signals suggesting the faces were the same, and needed more evidence to make this response on the GFMT.

CFMT

Thirteen participants with DP (45%) performed 2 SDs below the neurotypical mean score, with the remaining members of the DP group (55%) performing between one and two standard deviations below the Control mean. As expected (given it was used to select participants), a Mann-Whitney test revealed a significant difference in CFMT performance

between groups ($U = 40.50$, $p < .001$, $r = .78$), with DPs less accurate (median = 36, range: 23 - 43) than the Control group (median = 56, range: 33 - 70).

The results for both groups on all tests are shown in Figure 4.11. Relationships between all measures are reported in Table 4.4. and split across participant groups in Table 4.5.

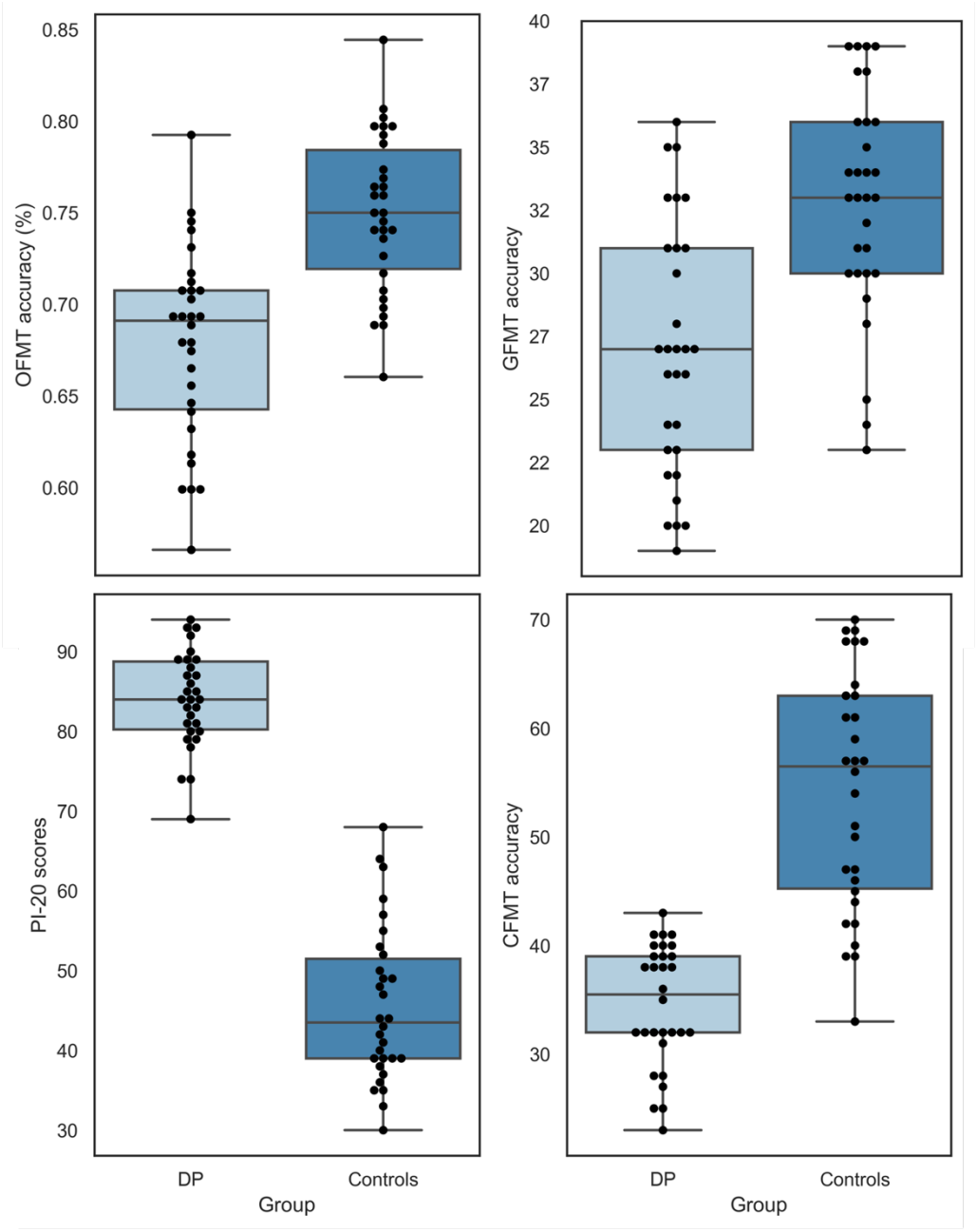


Figure 4.11. Group comparisons for all face processing tests.

Performance of a group of individuals with developmental prosopagnosia and an age- and gender-matched control group on the OFMT (left, top), the GFMT (right, top), the PI-20 (left, bottom) and the CFMT (right, bottom). The boxes represent the inter-quartile range, the lines within boxes show median performance, and whisker-lines represent the extremes of the distribution (excluding outliers).

Table 4.4. Relationship between all face measures across all participants.

Two asterisks (**) denote a relationship significant at p < .01 level.

	n	*M (SD)*	Cambridge Face Memory Test (CFMT)	20-Item Prosopagnosia Index (PI-20)	Glasgow Face Matching Task (GFMT)	Oxford Face Matching Test (OFMT)
CFMT	58	44.4 (13.2)	-			
PI-20	58	64.8 (21.0)	-.77**	-		
GFMT	58	29.9 (5.6)	.66**	-.58**	-	
OFMT	58	142.8 (12.0)	.73**	-.63**	.75**	-

Table 4.5. Relationships between all tasks separated by group (Developmental Prosopagnosia or Control).

Two asterisks (**) denote a relationship significant at p < .01 level

		n	**M (SD)**	**CFMT**	**PI-20**	**GFMT**	**OFMT**
CFMT	**DP**	29	34.6 (5.6)	-			
	Control	29	54.3 (10.9)	-			
PI-20	**DP**	29	84.1 (6.0)	-.02	-		
	Control	29	45.6 (9.8)	-.34	-		
GFMT	**DP**	29	26.9 (5.0)	.29	-0.05	-	
	Control	29	32.9 (4.4)	.61**	-.41	-	
OFMT	**DP**	29	135.6 (10.4)	.52**	-.32	.65**	-
	Control	29	150.0 (8.6)	.59**	-.21	.62**	-

4.4.2. Effect of Face Perception on Face Matching

Group (DP vs Control), Deviation scores, and their interaction were entered into two regressions, the first predicting OFMT matching accuracy and the second GFMT matching accuracy. For the OFMT analysis, Deviation scores were a significant predictor (β = -.70, t = -

6.58, p = <.001) of OFMT matching accuracy**,** suggesting that face perception abilities are related to OFMT performance. Group was also a significant predictor of OFMT matching accuracy (β = -.24, t = -2.46, p = .02), indicating that face matching was worse in the DP group even after accounting for face perception ability. The interaction between Group and Deviation scores was not a significant predictor (β = .09, t = 1, p = .32), suggesting that the relationship between face perception and OFMT matching performance did not vary as a function of group.

The same pattern of significance was observed in the analysis predicting GFMT matching accuracy - both Deviation scores (β = -.53, t = -3.95, p = <.001) and Group (β = -.28, t = -2.27, p = .03) were significant predictors of GFMT matching accuracy, while the interaction between Group and Deviation was not (β = .08, t = .72, p = .48).

4.4.3. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory

Group (DP vs control), Deviation scores, Face Matching scores (OFMT, and separately GFMT matching accuracy) and the interactions between Deviation scores and Group, and Face Matching and Group, were entered into regressions predicting CFMT scores. For the OFMT analysis, results showed significant independent contributions of face perception (as measured by Deviation scores; β = -.28, t = -2.08, p = .04) and face matching (β = .32, t = 2.52, p = .02). Group was a significant predictor (β = -.39, t = -4.20, p = <.001), indicating that face memory was worse in DP even after controlling for face perception and face matching. The two interactions were not significant predictors, indicating the relationships between face memory and face perception, and face memory and face matching, did not vary as a

function of group (Group x Deviation scores: β = .16, t = 1.38, p = .17; Group x Face Matching: β = -.10, t = -1.03, p = .31).

Similarly, for the GFMT analysis, results showed significant independent contributions of face perception (as measured by Deviation scores; β = -.36, t = -3.38, p = .001) and face matching (β = .26, t = 2.70, p = .009). Group was also a significant predictor (β = -.40, t = -4.46, p = < .001). The interactions between Group and Deviation scores and Group and Face Matching scores were not significant predictors (Group x Deviation scores: β = .16, t = 1.72, p = .09; Group x Face Matching: β = -.16, t = -2.00, p = .050). Given that the Group x Face Matching interaction approached significance, exploratory analyses assessed the correlation between GFMT residuals (which represent GFMT face matching independent of face perception, see 3.3.3. for Analysis Strategy section) and CFMT scores in each group. These revealed a significant positive correlation in the Control group (r_s = .46, p = .01) and a weaker, non-significant correlation (r_s = .02, p = .91) in the DP group, suggesting that the contribution of face matching to face memory performance, independent of face perception, was stronger for the control compared to the DP group.

4.5. Discussion

Results from this study support the validity of the OFMT. A group of individuals independently confirmed as DPs were significantly worse at the test than a matched neurotypical control group. Furthermore, OFMT scores correlated well with CFMT scores used, in part, to 'diagnose' individuals with DP. As expected, when standard test scores were compared, the group of individuals with DP performed worse on the other two face processing tasks (GFMT and CFMT) compared to the matched control group. It should be

noted that previous studies have not always found that prosopagnosic individuals perform poorly on the GFMT (e.g., Fysh & Ramon, 2021; White, Rivolta, Burton, Al-Janabi & Palermo, 2017), which may be due to issues with the sensitivity of the GFMT itself (White et al., 2022), as it tends to be the case that studies with larger sample sizes are more likely to find an impairment in DP. More interestingly, analysis of OFMT facial similarity judgements revealed that individuals with DP were worse at judging the perceptual similarity of two faces compared to the matched control group.

With respect to the debate concerning the perceptual and memory hypotheses of prosopagnosia, results supported both hypotheses. The DP group was significantly worse than the control group on the OFMT, a task with perceptual but not memory demands. The DP group was also significantly worse than the control group on the CFMT, a test with perceptual and memory demands, and this deficit in performance could not be fully accounted for by perceptual problems identified with the OFMT (as demonstrated using the hierarchical multiple regression analysis). Results therefore suggest that, as a group, individuals with DP have problems both with forming perceptual representations of faces and with learning/remembering facial identities. Indeed, when looking at individual cases, all but 2 of 30 DP participants scored below the control group's median score on the OFMT, confirming that the vast majority of individuals with DP have perceptual difficulties with faces.

The second analysis in this study examined the performance of individuals with DP on several face processing tasks, with the aim of deriving independent measures of face matching, face perception and face memory. Even after controlling for face perception performance, individuals with DP exhibited worse face matching performance (i.e., they

were worse at deciding whether two face images were from the same person, or different people). This result was robust: it was observed when matching was assessed using both the OFMT as well as the GFMT. This novel finding indicates that individuals with DP have difficulties with both face perception *and* face matching. Interestingly, this pattern is different from that seen in autistic individuals (Stantić et al., *under review*), whereby autistic individuals exhibit difficulties in face perception but not in face matching. Although it is usually claimed that there are high rates of prosopagnosia in the autistic population (e.g., Cook, Shah, Gaule, Brewer & Bird, 2015; Wilson, Palermo, Schmalzl & Brock, 2010), these data indicate that there may be subtle differences in face processing between autistic individuals and non-autistic individuals with DP.

Comparison of the results of autistic and prosopagnosic individuals reinforces the distinction between the psychological processes of face perception and face matching. Furthermore, it suggests that although face perception is likely necessary for accurate face matching, it is not sufficient. In addition to being able to form accurate perceptual representations of faces from memory or from a pictorial representation, one must have an accurate model of how, and how much, faces are allowed to vary before deciding they belong to different people. Results of the Signal Detection Theory analysis of the OFMT and GFMT matching task may be informative as to this point. In addition to a lower *d'*, participants with DP exhibited a more extreme bias towards 'different' responses. This is to say that they needed more evidence that the faces were the same before they responded that they were the same. This is consistent with personal reports from prosopagnosic individuals who report that they fail to recognise individuals – i.e., they fail to recognise instances of faces (whether stored in memory or available for visual inspection) of the same facial identity, and therefore are more likely to judge faces as different. These data

are also consistent with claims that individuals with DP show impaired performance on matching tasks specifically for trials in which the two faces depict the same person, compared to trials in which different individuals are depicted (White et al., 2017).

Finally, individuals with DP exhibited impaired face memory even after accounting for their difficulties with face perception and face matching, indicating problems with all three facial identity processes tested.

Despite the clear pattern of impairments seen in those with DP, it should be acknowledged that there were two main limitations of the current study which raise questions about the degree of generalisability that can be assumed from the current results. The first is the lack of any control stimulus class, as a result of which it is not clear whether the perceptual, matching and memory impairments seen for faces would extend to other stimulus classes. This problem is a general one for the field, and reflects the fact that appropriate control stimuli are difficult to identify (e.g., Fry et al., 2020; Susilo et al., 2010). The second limitation is that prosopagnosic individuals were selected on the basis that they showed impaired performance on two out of three diagnostic tests, one of which was the Cambridge Face Perception Task. As such, the sample of DP individuals identified may have been biased towards those that have perceptual difficulties in addition to problems with face memory. This possibility will only be able to be tested with further testing of the DP population with varying recruitment criteria, though it is worth noting that a fairly large degree of variation in test scores was observed within the group of individuals with DP. Unfortunately, the sample size was too small to allow formal testing for the presence of sub-groups (Dalmaijer, Nord, & Astle, 2022) but if future work adopts the same testing and analysis procedure then samples could be combined to enable this approach.

Collectively, these results suggest that individuals with DP exhibit impaired face perception, face memory *and* face matching. Thus, we suggest that neither the 'memory' nor the 'perceptual' hypothesis is a sufficiently comprehensive account of face processing difficulties in DP, and that decision-making processes involved in matching perceptual stimuli with stored face representations (e.g., Bruce & Young, 1986), and how they may be impaired, are also of importance in understanding DP.

Chapter 5: Face perception and memory in super recognition

An edited or partial version of this chapter is currently under review as a following article:

- Stantić, M., Pounder, Z., Bate, S., Catmur, C., Bird, G.P. (under review) Individuals who are 'super-recognisers' show superior performance on independent measures of face perception, face memory and face matching.

5.1. Abstract

Individuals who are atypically good at face recognition are described as '*super-recognisers'* (SRs). On standard face recognition tasks SRs outperform individuals who have typical face recognition ability. However, high accuracy on face recognition tasks may be driven by superior ability in one or a number of processes including face perception, face matching, and face memory. The present study utilised the Oxford Face Matching Test (OFMT), and a novel analysis strategy to derive independent measures of face perception, face matching and face memory. Thirty-two SRs and the same number of matched controls with typical face recognition ability undertook three face tasks: the OFMT, the Glasgow Face Matching Test (GFMT), the Cambridge Face Memory Test (CFMT) and one self-report measure, the 20-item Prosopagnosia Index (PI-20). Results showed that, when compared by group, SRs were more accurate than controls across all tasks and they reported greater face recognition ability. Of most importance, however, was the fact that SRs exhibited superior face perception, face matching and face memory. Collectively, these results suggest that SRs have superior ability across multiple independent face-related processes.

5.2. Introduction

Individuals who have an extraordinary ability to recognise faces are known as *'super-recognisers'* (SRs; Russell, Duchaine, & Nakayama, 2009). As would be expected, many studies have confirmed that SRs outperform typical individuals on face recognition tasks. Such a finding does not explain *why* SRs have such good face recognition ability, however, and so it is interesting that several studies report superior performance by SRs, or at least a subset of SRs, on other face processing tasks (e.g., Belanova et al., 2018; Bobak et al., 2016a; Bobak et al., 2016b; Smith et al., 2021). They excel on tests of face memory, and have been found in some studies to show superior performance on tests of face perception (assessing the ability to form an accurate representation of faces, often based on degraded input; Russell et al., 2009; Bobak et al., 2016a), and a subset of SRs have been found to perform at superior level at tasks of face matching (designed to test the ability to determine whether two facial images are of the same person or of different people; Bate et al., 2018; Bobak, Dowsett, & Bate, 2016).

While the pattern of performance of SRs across different tests of face processing is interesting, it still does not allow it to be ascertained why SRs have such good face recognition ability. This is because tests of face processing are not 'process-pure' (see Chapter 3, pg. 86). They do not allow performance on a single aspect of face processing to be tested independently of other processes.

Along with CFMT and GFMT which have previously been discussed, face perception tasks such as the Cambridge Face Perception Task (CFPT; Duchaine, Germine, & Nakayama, 2007) are also not process pure. In the CFPT, more perceptually similar images also contain more of the same identity vs a different identity, making it impossible

to determine whether face perception, face matching, or a combination of the two, is driving performance. Thus far, therefore, research has lacked the means to obtain independent measures of face perception, face matching, and face memory, to determine whether SRs show superior performance in each of these processes.

A strategy to obtain independent estimates of these processes has recently been developed, which relies on the use of the Oxford Face Matching Test (OFMT; Stantić et al., 2022a). The OFMT has been used to assess individual differences in face processing abilities in a range of non-clinical populations (Stantić et al., 2022a; Stantić et al., 2022b), as well as individuals with autism (Stantić, Brown, Catmur & Bird, *under review*) and developmental prosopagnosia (Stantić et al., 2022b). For instance, individuals with autism have frequently been demonstrated to have difficulties with face recognition (see Wiegelt, Koldewyn**,** and Kanwisher (2012) for review), and use of the OFMT and analysis technique described above revealed difficulties with both face memory and face perception, in the presence of intact face matching. In contrast, individuals with DP were shown to have impaired face memory, face perception, and face matching (Stantić et al., 2022b.

Super recognition ability and developmental prosopagnosia are often described as being at opposite ends of the typical face recognition spectrum despite there being little evidence to support this idea. What is needed are studies probing exactly why individuals with developmental prosopagnosia have such poor face recognition abilities, and whether it is those same processes which underlie the superior face recognition abilities seen in SRs. Given the work described above which demonstrated impaired face memory, perception and matching in individuals with developmental prosopagnosia, if SRs are truly at the opposite end of the typical face processing spectrum one would expect superior face

memory, perception and matching in SRs. It is this possibility that the current work was designed to test.

In summary, the current study uses the OFMT and the CFMT to derive independent measures of face perception, face matching, and face memory, in a group of SRs and a matched control group with typical face processing ability, to identify the specific processes that are superior in SRs.

5.3. Methods

5.3.1. Participants

Thirty-two SRs were recruited to participate in the study (16 female: M_{age} = 39.9, *SD* = 9.3). Super-recogniser participants were recruited from a pre-existing database of super recognisers who self-reported a superior ability to recognise faces and met the criteria for super recognition as assessed by clinical interview and performance on the Cambridge Face Memory Test Long Form (CFMT+; Russell et al., 2009; all SR participants scored above 90, *M* = 95.50, *SD* = 2.96, range 91 -102). An age and gender-matched sample of 32 neurotypical participants were recruited via Prolific. Six of these participants were excluded because they exhibited poor performance (*n* = 3, *M* = 31.67, *SD* = 5.13) or superior performance (*n* = 3, all scored 69) on the CFMT. These participants were replaced with 6 new neurotypical participants, who were recruited via Prolific, and a full sample is reported (*n* = 32; 16 female: M_{age} = 39.1, *SD* = 5.8). The SR and neurotypical controls did not differ significantly in terms of age ($t(62) = 0.36$, $p = .72$) or gender distribution ($X^2(1) = 0.62$, $p = .43$). No participants from either group were excluded for failing attention

check trials on the OFMT. All participants reported having normal or corrected-to-normal vision.

5.3.2. Procedure

The procedure was as outlined in Chapter 2 (2.3. Common Methodology).

Note that SRs did not complete the CFMT as part of the current testing session as it was completed during a diagnostic testing session.

5.3.3. Analysis Strategy

The analysis strategy was as outlined in Chapter 3 (3.3.3. Analysis Strategy).

In the current study, we explore SR impacts face perception, face matching (independent of face perception) and face memory (independent of face perception and face matching) using a series of regression analyses in which group (SR vs Control), predictors (e.g., face perception), and their interaction, were used to predict test scores (e.g., CFMT test scores). Including the interaction term in regression models allows for the relationship between, e.g., face perception (i.e., algorithmic deviation) and CFMT scores, to vary across groups.

5.4. Results

5.4.1. Group differences

Differences on group performance for all tasks are shown in Figure 5.12. Descriptive statistics for all participants are shown in Table 5.6. and split across participant groups (SR vs Control) in Table 5.7., with both tables showing the relationships between tasks as well.

OFMT: Matching performance

Eighteen SR participants (56%) performed one standard deviation above the Control group's mean score. Another 11 SRs (34%) had OFMT scores that were greater than the neurotypical mean score (but less than one standard deviation above the control mean), while an additional 3 SRs (9%) had a OFMT score that was either the same (n = 1) or below (n = 2) the neurotypical mean score (but less than one standard deviation below the control mean). At the group level, an independent t-test showed a significant difference in OFMT matching accuracy [$t(62) = 5.05$, $p < .001$, $d = 1.26$], with SRs more accurate (M = 82%, SD = 4.1%) than Control participants (M = 75%, SD = 6.2%).

Signal Detection Theory (Green & Swets, 1966) was used to characterise performance on the matching component of the OFMT, providing a measure of sensitivity (*d'*) and bias (criterion). A Mann-Whitney test showed a significant difference in *d'* ($U = 183.50$, $p < .001$, $r = 0.55$) with the SR group showing greater sensitivity (SR *d'* median: 1.99, range: 1.26 – 2.33; Control *d'* median: 1.56, range; 0.71 – 2.64) than the control group. A Mann-Whitney test revealed no differences in criterion scores ($U = 421.50$, $p = .11$, $r = 0.15$) between the two participant groups.

OFMT: Algorithmic Deviation

Seven SR participants (22%) performed two standard deviations below (indicating more accurate performance) the Control mean standard deviation score, while 13 SR participants (41%) performed one standard deviation below the Control mean. Another 8 SRs (25%) had a deviation score that was below the neurotypical deviation mean (but less than one standard deviation). One SR (3%) scored one standard deviation above the Control mean,

and three other SR (9%) scored two or more standard deviations above the control mean. At the group level, a Mann-Whitney test showed a significant difference in algorithmic deviation ($U = 214$, $p < .001$, $r = 0.50$) with SR participants having better deviation scores, or smaller deviation from algorithmic judgments (median = 19.08, range: 14.06 - 36.08), than the Control group (median = 22.42, range: 16.99 – 28.82).

GFMT

Twenty-five SR participants (78%) performed one standard deviation above the Control group's mean score. Another 6 SRs (19%) had GFMT scores that were greater than the neurotypical mean score (but less than one standard deviation above the Control mean), and one SR (3%) had a GFMT score of more than one standard deviation below the Control mean. At the group level, a Mann-Whitney test revealed a significant difference in GFMT accuracy ($U = 121$, $p < .001$, $r = 0.66$) with SR participants more accurate (median = 39, range: 25 - 40) than the Control group (median = 35, range: 20 - 40).

PI-20

Twenty-seven SR participants (84%) scored one standard deviation below the Control group's mean PI-20 score (where lower scores indicate superior self-reported face recognition ability). The remaining five SRs (16%) had a PI-20 score that was less than the neurotypical score mean (but less than one standard deviation below the control mean). At group level, a Mann-Whitney test revealed a significant difference in PI-20 scores ($U = 47$, $p < .001$, $r = 0.78$) with SR participants reporting lower scores (median = 26.50, range: 23 – 37) than the Control group (median = 41, range: 29 – 70) .

CFMT

All 32 SR participants (100%) performed one standard deviation above the Control group's CFMT mean score. At the group level, a Mann-Whitney test showed a significant difference in CFMT accuracy ($U = 4$, $p < .001$, $r = 0.86$) with SR participants more accurate (median = 71, range: 67 - 72) than the Control group (median = 57, range: 39 - 68).

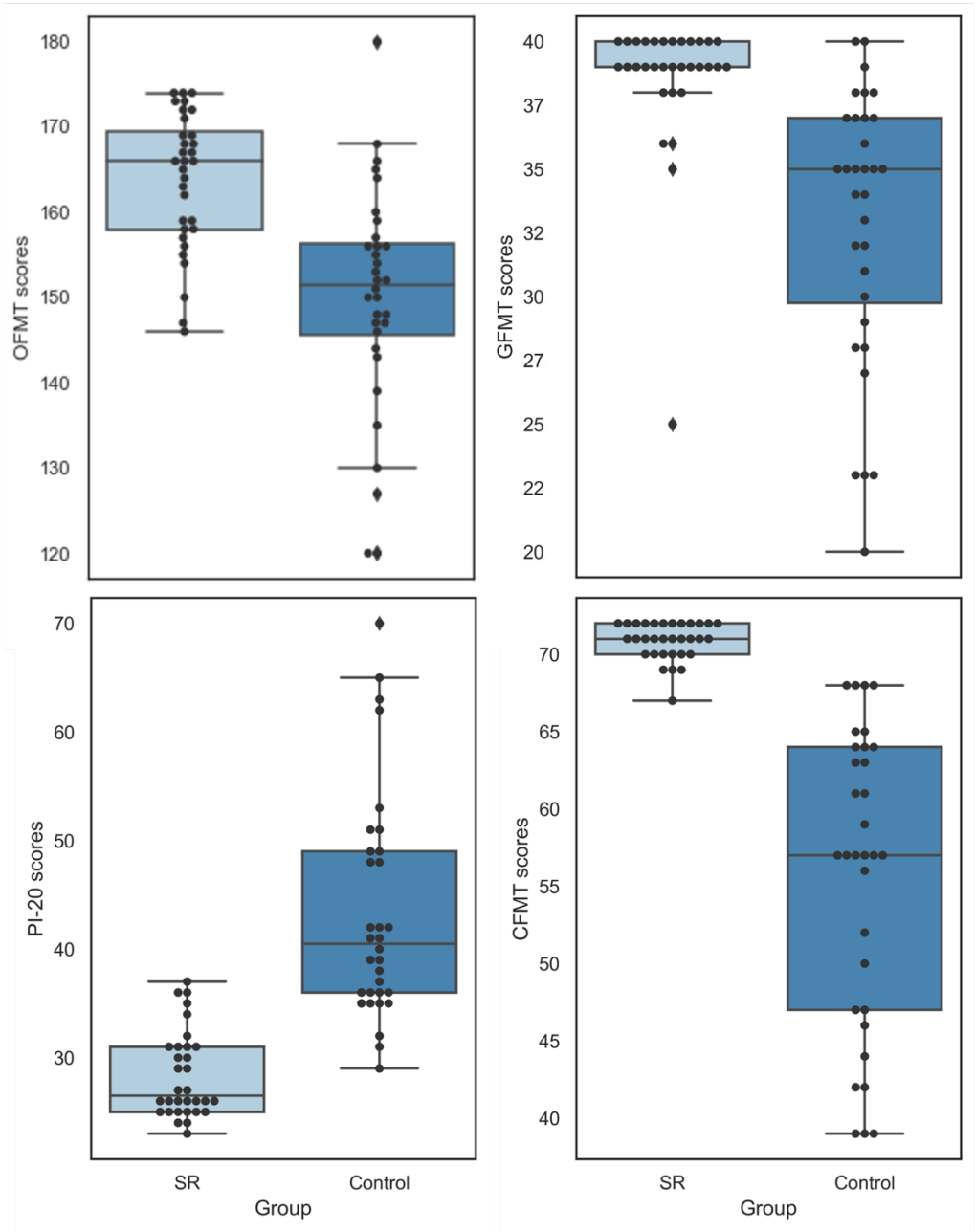


Figure 5.12. Group comparisons for all processing tasks.

Scores are shown for the OFMT (top, left), and algorithmic deviations (top, right), the GFMT (bottom, left), and the CFMT (bottom, right).

Table 5.6. Relationship between all face measures across all participants.

Two asterisks (**) denote a relationship significant at p < .01 level.

	n	*M (SD)*	Cambridge Face Memory Test (CFMT)	20-Item Prosopagnosia Index (PI-20)	Glasgow Face Matching Task (GFMT)	Oxford Face Matching Test (OFMT)
CFMT	64	64.5 (10.1)	-			
PI-20	64	36.1 (10.9)	-.63**	-		
GFMT	64	35.8 (5.0)	.54**	-.62**	-	
OFMT	64	156.9 (12.9)	.60**	-.57**	.60**	-

Table 5.7. Relationships between all tasks separated by group (Super recognizers or Control).

One asterisk (*) denotes significance at the p < .05 level and two asterisks (**) denote a relationship significant at p < .01 level.

		n	M (SD)	CFMT	PI-20	GFMT	OFMT
CFMT	**SR**	32	70.9 (1.2)	-			
	Control	32	56.0 (9.6)	-			
PI-20	**SR**	32	28.4 (4.0)	-.42*	-		
	Control	32	43.7 (10.4)	-.23	-		
GFMT	**SR**	32	38.5 (2.8)	.01	-.14	-	
	Control	32	33.0 (5.3)	.26	-.45*	-	
OFMT	**SR**	32	163.8 (8.0)	.39*	-.05	.08	-
	Control	32	149.9 (13.3)	.39*	-.38*	.54**	-

5.4.2. Effect of Face Perception on Face Matching

Group (SR vs Control), Deviation scores, and their interaction were entered into two regressions, the first predicting OFMT matching accuracy and the second GFMT matching

accuracy. For the OFMT analysis, Deviation scores were a significant predictor (β = -.42, t = -3.38, p = .001) of OFMT matching accuracy, suggesting that face perception ability is related to OFMT performance. Group was also a significant predictor of OFMT matching accuracy (β = .40, t = 3.69, p < .001), indicating that face matching was better in the SR group after accounting for face perception ability. The interaction between Group and Deviation scores was not a significant predictor (β = .22, t = 1.89, p = .06).

This same analysis was repeated using GFMT instead of OFMT scores as a robustness check. Group was again a significant predictor of GFMT matching accuracy (β = .52, t = 4.47, p < .001), suggesting that face matching was better in the SR group than in the Control group after accounting for face perception ability. Deviation scores were not a significant predictor (β = -.10, t = -0.73, p = .47) and the interaction between Group and Deviation was also not significant (β = .05, t = 0.39, p = .70).

5.4.3. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory

Group (SR vs control), Deviation scores, Face Matching scores (OFMT matching accuracy, and separately GFMT matching accuracy) and the interactions between Deviation scores and Group, and Face Matching and Group, were entered into regressions predicting CFMT scores. For the OFMT analysis, results showed no significant independent contributions of face perception (as measured by Deviation scores; β = -.10, t = -0.92, p = .36) or face matching (β = .20, t = 1.71, p = .09). Group was a significant predictor (β = .60, t = 5.94, p = < .001), indicating that face memory was superior in SRs even after controlling for face perception and face matching. The two interactions between Face Matching x Group and Deviation

scores x Group were not significant (Face Matching and Group: β = -.09, t = -0.92, p = .36; Deviation scores and Group: β = .12, t = 1.14, p = .26).

Similarly, for the analysis using GFMT scores, results showed no significant independent contributions of face perception (as measured by Deviation scores; (β = -.19, t = -1.87, p = .07) or face matching (β = .11, t = 0.91, p = .37). However, Group was a significant predictor (β = .62, t = 5.62, p < .001). The two interactions between Group x Face Matching scores and Group x Deviation scores were not significant predictors (Group x Face Matching: β = -.09, t = -.87, p = .39; Group x Deviation scores: β = .19, t = 1.95, p = .06).

5.5. Discussion

This study examined the performance of SRs on several face processing tasks, with the aim of deriving independent measures of face matching, face perception and face memory. Results showed that, when assessed independently, SRs had superior ability in all three aspects of face processing. Furthermore, this patten of data was replicated when matching scores for the GFMT were used instead of OFMT matching scores, indicating the robustness of these results. Group comparisons of performance on standard tests was aligned with this conclusion: the SR participants exhibited superior performance on all three face processing tests (OFMT, GFMT and CFMT), and reported better face recognition ability on the PI-20. SR participants also made similarity judgments of pairs of faces that were more similar to those made by leading computer vision algorithms than the matched control group.

When assessed at the group level then, the idea that SRs are at the opposite end of the face recognition spectrum to developmental prosopagnosics is supported by these data in comparison to the performance of individuals with developmental prosopagnosia when tested using the same approach (Stantić et al., 2022b). Both groups showed a consistent pattern of performance, such that they were either impaired or superior to average performers across face perception, matching and memory. Interestingly, this contrasts with individuals with autism who showed impaired face perception and memory, but intact face matching.

Examination of the scores of individual SRs show, by definition, little variation in CFMT scores. Of more interest is the fact that significant variation is seen on face matching tests (particularly the OFMT which avoids the problem frequently seen on the GFMT where ceiling effects are observed even among those with typical face processing ability, but especially with SRs; Dowsett & Burton, 2015; Stantić et al., 2022a; Dunn et al., 2020; Verhallen et al., 2017). Even more variation is seen on the face perception measure, indicating a wide range of differences between face similarity judged by algorithms and face similarity as judged by SRs.

The variation in face matching and face perception skill among individuals classed as SRs on the basis of their face memory performance is of interest for both applied and theoretical reasons. In the applied domain, SR individuals are sometimes recruited by police forces and intelligence agencies. If employed on tasks requiring excellent face memory this may be perfectly appropriate, but if employed for other face processing tasks where performance is largely governed by face matching or face perception, some SRs may be no better than the average member of the public. Theoretically, this finding is of

interest as it highlights the various routes by which one can achieve superior face memory performance. Individuals can range from demonstrating exceptional face perception, matching and memory, to exceptional face memory only. It is a limitation of this study, and indeed of the wider field, that appropriate control stimuli are not used to determine the degree to which superior performance is specific to faces, or common across stimulus categories. It is possible, for example, that individuals who exhibit superior face memory in the presence of typical face perception and matching, may have excellent memory for all stimuli, or have hyperphantasia (e.g., Zeman et al., 2020).

More broadly, the wide variation in face perception and face matching observed using the OFMT in the SR group may be a sign that the current measures used to identify SRs are not sufficient (or at least do not identify a homogenous group). In contrast to developmental prosopagnosics, SRs are frequently identified on the basis of superior performance on one test only (the CFMT+). Two strategies to improve this procedure are possible. The first is to administer the CFMT+ alongside other, more difficult tests of face memory and only consider individuals to be SRs if they perform at superior levels on multiple tests (in line with the DP diagnostic criteria). The second strategy is to redefine SRs as having superior performance across face perception, matching and memory, and use an approach similar to that adopted here.

Overall, this study derives independent measures of face perception, face matching, and face memory, in a group of SRs and in a matched control group. Data indicate that, as a group, SRs have superior face perception, matching and memory, but that individual differences are seen in the SR group in face perception and face matching,

Chapter 6: Face perception and memory in Autism Spectrum Disorder

An edited or partial version of this chapter has been published as the following journal articles:

- Stantić, M., Ichijo, E., Catmur, C., & Bird, G.P. (2021) Face Memory and Face Perception in Autism. *Autism: The International Journal of Research and Practice, 26*(1), 276-280

- Stantić, M., Brown, K. Catmur, C. & Bird, G. (under review). Independent Measurement of Face Perception, Face Matching, and Face Memory Reveals Impairments in Face Perception and Memory, but Not Matching, in Autism.

6.1. Abstract

It has been argued that autistic individuals have difficulties with face memory but typical face perception. However, only one previous study has examined both face memory and face perception in the same individuals, and this study was conducted with a small group of autistic children. Here face recognition was examined with a group of autistic adults using two face perception tasks (including one designed to avoid a neurotypical bias), and a standard test of face memory. Self-reported face recognition difficulties in everyday life were also recorded. Further, there are multiple psychological processes required in order for a face to be recognised. However, when testing face memory using tasks such as the Cambridge Face Memory Task (CFMT), it is rare for studies to attempt to account for individual differences in face perception and face matching in order to isolate variance in face memory specifically. First study found that the group of adults with autism scored lower than a matched neurotypical control group on all face tasks, and reported more problems with face recognition in everyday life. On the whole results suggest difficulties with both face perception and face memory in autistic adults, although it should be noted that a wide range of scores were recorded from the autistic individuals, with some scoring in the neurotypical range. Secondly, the OFMT was used to obtain an independent measure of face matching in addition to face perception in a group of 57 autistic adults and a matched neurotypical control group. Results revealed impaired face perception and memory in the individuals with autism, but intact face matching.

6.2. Introduction

Face recognition has been relatively well-studied in individuals diagnosed with Autism Spectrum Disorder (henceforth 'autism'), a condition defined by persistent challenges, difficulties or alterations in social communication and social interaction, and restricted, repetitive patterns of behaviour, interests, or activities (APA, 2013). Across approximately 150 studies results are mixed; with reports of both typical and impaired face recognition in autism. In their review, Wiegelt, Koldewyn**,** and Kanwisher (2012) argue that these mixed results can largely be explained by a selective deficit in face memory in the presence of intact face perception. With some exceptions (see Tang et al., 2015; Griffin et al., 2021), studies are in accordance with this view – the majority of studies requiring a face to be held in memory and compared to a subsequent exemplar reveal performance impairments in autistic volunteers, whereas studies involving tasks requiring two faces to be compared without memory demands (e.g., in a simultaneous identity matching task) tend to find typical performance by autistic volunteers.

Problematically however, as far as we are aware, there is only one previous study that compares face memory and face perception in the same autistic individuals (Gepner, DeGelder, & DeSchoner, 1996), and this study tested the performance of only seven children with autism. Furthermore, no previous study has assessed face matching independent of face perception in autism, nor face memory independent of face perception and matching. It is important to test face perception and face memory in the same individuals, because differential performance across groups of individuals in different studies may be explained by sample differences in, for example, age or gender ratio, or differences in experimental stimuli, rather than the type of face processing required for good performance. It is also important to assess face *matching* independent of face *perception* in autism, as it is possible

that an impairment in face matching produces what appear to be deficits in face perception and memory in autism (Stantić et al., 2021b; Robel et al., 2004). This chapter reports two studies which aim to use the novel nature of the OFMT to determine the degree to which variance in face perception and face matching explain variance in face memory in neurotypical and autistic individuals, and to determine which aspects of face processing are impaired in autistic individuals - particularly whether face memory is impaired in autism after accounting for any impairment in face perception and face matching.

6.3. Methods

6.3.1. Procedure

The procedure was as outlined in Chapter 2 (2.3. Common Methodology). Additionally, also in a randomized order, participants completed the matrix reasoning subtest of the WASI-II IQ test, and the AQ-50, a measure of autistic traits (Baron-Cohen et al., 2001).

6.3.2. Analysis Strategy

The analysis strategy was as outlined in Chapter 3 (3.3.3. Analysis Strategy).

Further, we determined whether autism impacts face perception, face matching (independent of face perception) and face memory (independent of face perception and face matching) using scores as calculated in 3.3.3. Analysis Strategy using a series of regression analyses in which group (autism vs control), predictors (e.g., face perception), and their interaction, was used to predict test scores (e.g., face matching). Including the

interaction term in regression models allows for the relationship between, e.g., face perception and face matching, to vary across groups.

6.3.3. Participants

Fifty-eight autistic individuals were recruited to take part, but one participant was excluded for failing to pass attention checks. The final sample consisted of 57 autistic individuals (32 male; M_{age} = 42.02, SD = 13.39), and 57 individuals with no current or previous clinical diagnosis of autism (29 male; M_{age} = 41.81, SD = 9.27).

Autistic individuals were diagnosed by an independent clinician and met criteria for autism or autism spectrum on the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2000). Autistic individuals reported the following co-occurring conditions within the last 5 years: anxiety (4 participants), depression (4), ADHD (4), dyspraxia (3), and OCD (1). As expected, the autistic group had greater levels of autistic traits than the neurotypical group [$t(105) = 9.84$, $p < .001$], as assessed by the Autism-Spectrum Quotient (AQ-50; Baron-Cohen et al., 2001). However, the autistic and neurotypical groups did not differ significantly in terms of age [$t(112) = 0.10$, $p = .922$], gender [$X^2(1) = 0.32$, $p = .573$], or non-verbal IQ [$t(112) = 0.73$, $p = .468$] as measured by the matrix reasoning component of the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II; Wechsler, 2011).

Neurotypical participants were recruited via Prolific.co. Three originally recruited NT participants were replaced as they either 1) had a current or past psychiatric neurodevelopmental diagnosis, 2) used psychotropic meditation or 3) failed to attend to the task. All NT participants scored below cut-off (32) on the AQ-50.

6.4. Results

6.4.1. Group Differences

The autistic group performed significantly worse than the neurotypical control group on all tests.

OFMT: Matching performance

The performance of the autistic group *(M* = 138.42, *SD* = 12.41, Range = 102 - 165) was significantly worse than the neurotypical group (*M* = 145.02, *SD* = 12.30, Range = 112 - 168), $t(112) = 2.85$, $p = .005$. Forty of 57 autistic participants (70.0%) scored below the median neurotypical performance (indicating worse performance).

OFMT: Deviations

The deviation scores of the autistic group (*M* = 26.37, *SD* = 4.12, Range = 17.76 - 40.57) were significantly worse than the deviation scores of the neurotypical group (*M* = 24.42, *SD* = 3.43, Range = 16.14 - 30.96), $t(112) = 2.75$, $p = .007$. Thirty-nine out of 57 autistic participants (68.4%) scored above the median neurotypical score (indicating worse performance).

GFMT

A Shapiro-Wilk test of GFMT scores revealed a significant departure from normality ($W = 0.96$, $p = .002$). Therefore, a square transformation was applied to the total GFMT scores for statistical analysis following the implementation of the Box-Cox procedure (Box & Cox, 1964). The performance of the autistic group *(M* = 30.75, *SD* = 5.23, Range = 17-39, untransformed) was significantly worse than the neurotypical group (*M* = 33.16, *SD* =

3.84, Range = 24-40, untransformed), $t(112) = 2.69$, $p = .008$. Thirty-five of 57 autistic participants (61.4%) scored below the median neurotypical performance (indicating worse performance).

CFMT

The performance of the autistic group *(M* = 46.54, *SD* = 10.88, Range = 19-71) was significantly worse than the neurotypical group (*M* = 53.81, *SD* = 9.89, Range = 36-72), $t(112) = 3.73$, $p < .001$. Forty-four of 57 autistic participants (77.2%) scored below the median neurotypical performance (indicating worse performance).

The results for both groups on all tests are shown in Figure 6.13. Relationships between all measures are included in Table 6.8.

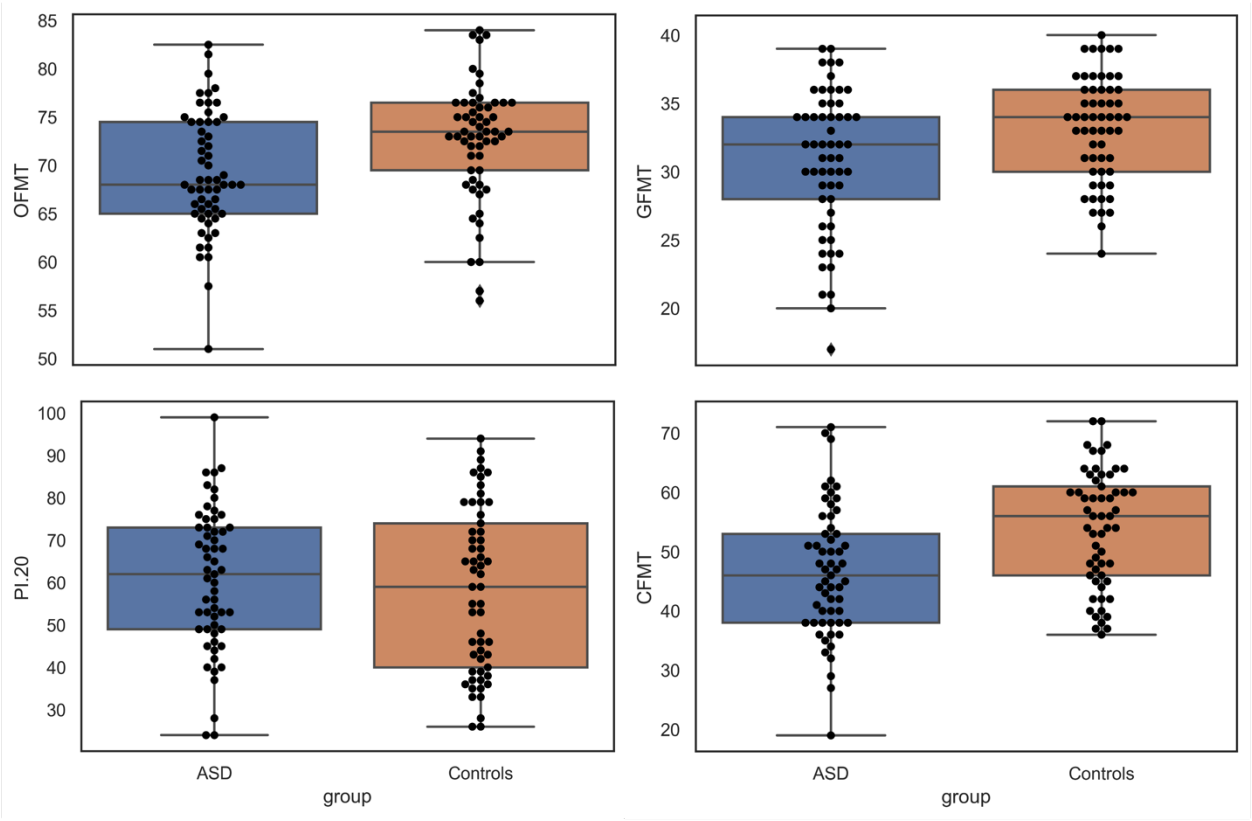


Figure 6.13. Group differences for all face processing tasks.

The boxes represent the interquartile scores, the horizontal lines in boxes represent group medians, and the whisker-lines span the full range of scores within each group (excluding any outliers, which are shown as separate dots). Matching tasks are shown in the top panel (OFMT, left; GFMT, right) whereas the bottom panel includes the memory task (CFMT, left) and the self-report measure of difficulties with face recognition (PI-20, right).

Table 6.8. Relationships between all tasks separated by group (Autism or Control).

Two asterisks (**) denote a relationship significant at p < .01 level

		CFMT	PI-20	GFMT	OFMT
CFMT	**Autism**	-			
	Control	-			
PI-20	**Autism**	-.14	-		
	Control	-.09	-		
GFMT	**Autism**	.50**	.03	-	
	Control	.54**	-.03	-	
OFMT	**Autism**	.51**	-.01	.61**	-
	Control	.48**	-.10	.54**	-

6.4.2. Effect of Face Perception on Face Matching

Group (Autism vs Control), Deviation scores, and their interaction were entered into two regressions, one predicting OFMT scores and the other GFMT scores. For the OFMT analysis, Deviation scores were a significant predictor (again showing that face perception abilities are related to face matching performance; β = -.611, t = 8.00, p < .001), Group was not a significant predictor (suggesting that face matching by autistic individuals was no different from neurotypical individuals after accounting for face perception; β = -.107, t = 1.42, p = .158), and the interaction between Group and Deviation scores was not a significant predictor (indicating that the relationship between face processing and face matching did not vary as a function of group; β = -.009, t = 0.12, p = .908). For the GFMT analysis, the pattern of significance was the same. Deviation scores were a significant predictor (β = -.456, t = 5.41, p < .001), Group was not a significant predictor (β = -.132, t = 1.59, p = .115), and the interaction between Group and Deviation scores was not a significant predictor (β = -.134, t = 1.64, p = .103).

6.4.3. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory

Group (Autism vs Neurotypical), Deviation scores, Face Matching Scores (OFMT and separately GFMT) and the interactions between Deviation scores and Group, and Face Matching and Group, were entered into a regression predicting CFMT scores. For the OFMT analysis, results demonstrated significant independent contributions of face perception (as measured by Deviation scores), and face matching (β = -.248, t = 2.44, p = .016; β = .319, t = 3.14, p = .002, respectively). Group was a significant predictor (β = -.187, t = 2.33, p = .022), indicating that face memory was worse in autism even after accounting for face perception and face matching scores. The two interactions were not significant predictors, indicating the relationship between face memory and face perception and face matching did not vary as a function of group (Group x Deviation scores: β = -.044, t = 0.445, p = .657; Group x Face Matching: β = -.016, t = 0.167, p = .868).

For the GFMT analysis, the pattern of significance was identical. Results demonstrated significant independent contributions of face perception (as measured by Deviation scores), and face matching (β = -.291, t = 3.28, p = .001; β = .365, t = 4.12, p < .001, respectively). Group was a significant predictor (β = -.169, t = 2.17, p = .032), while the two interactions were not significant predictors (Group x Deviation scores: β = -.034, t = 0.391, p = .697; Group x Face Matching: β = -.101, t = 1.17, p = .245).

6.5. Discussion

It has been claimed that autistic individuals are impaired on face memory but not face perception tasks (Wiegelt et al., 2012; but see Tang et al., 2015). In order to test this claim, it is essential to test face memory and face perception in the same individuals, as the high heterogeneity seen in autism may mean that sampling differences (random or otherwise) across studies manifest as artefactual differences in performance across test types. Further, it is crucial that measures of face perception, matching and memory can be assessed independently of each other. Accordingly, a group of autistic adults completed a test of face memory (CFMT) and two tests of face perception (GMFT and OFMT). Results showed that the group of autistic individuals performed worse on all three tests of face processing (OFMT, GFMT, and CFMT), and were worse at judging the similarity of two faces, than a matched control group., and also reported more problems with face recognition on the PI-20.

These data are not consistent with claims that face perception is spared in autism, as performance on both the GFMT and OFMT (with the latter designed to avoid a potential bias in favour of NT individuals) was lower in the autism group then the NT group. It is worth noting that impaired performance was not universal in the autism group, however, with 9 and 7 autistic individuals performing better than the median performance of the NT group on the GFMT and OFMT, respectively. Importantly, intact performance in some autistic individuals was not limited to face perception, but was also seen on the test of face memory where 5 autistic individuals performed better than the NT median. Interestingly, a single case of a participant with autism scoring below the NT median on the face perception tasks, and above the NT median on the memory task (CFMT) was observed. This result

highlights that group level comparisons do not allow the source of any atypical performance in the autism group to be identified.

However, all three behavioural tests included in this study require face matching, where face matching refers to the ability to judge whether two images of a face are from the same person, in addition to face perception. From the group level analysis alone, it is unclear whether atypically good or poor performance is caused by atypical perception, attention or memory. This is an important issue to be addressed by the following analysis that captures performance of these individual components independent of each other and is therefore able to shed additional light on the difference in performance by autistic participants. It is possible that autistic individuals are able to form an accurate perceptual representation of faces, but use sub-optimal decision criteria when deciding whether two facial images are from the same individual. If this is the case, it would be possible for face perception to be intact but poor performance to be observed on the OFMT and GFMT, consistent with claims of difficulty generalizing from exemplars in autism (Scherf, Luna, Kimchi, Minshew & Behrmann, 2008).

The second analysis is designed to address this possibility. When face perception was controlled for by accounting for the degree to which similarity judgements deviated from algorithmic judgements, the group of autistic individuals did not perform worse than the control group on either the OFMT or the GFMT. This pattern of results indicated that although face perception is, on average, impaired in the autistic group, face matching is not. The poor performance on the CFMT has been shown previously (Griffin et al., 2021) and is normally taken as evidence of impaired face memory in autism. These results go further, in demonstrating an autistic impairment in face memory even after accounting for

face perception and face matching. Overall then, results suggests that individuals with autism exhibit impaired face perception and face memory, but intact face matching.

These conclusions are consistent with the latest meta-analysis of face processing in autism (Griffin et al., 2021) in showing face perception and face memory difficulties in autism, but also show that the face memory impairment is not an artifact of face perception difficulties. In addition, and to our knowledge for the first time, these results show intact face matching ability in autism. In other words, given accurate perceptual representations, autistic individuals are no worse than neurotypical individuals in deciding whether two images are of the same face or different faces. Intact face matching in autism means that if face perception could be improved through clinical intervention, improved face recognition should follow (based on the results of first analysis).

Finally, it should be noted that, as is typical, a high degree of heterogeneity was observed in the autism group (Stantić et al., 2021b). A proportion of autistic individuals (between one- and two-fifths of the sample) performed better than the median neurotypical individual, and some performed excellently. Of course, wide-reaching conclusions about performance of any autistic group are limited given that specific data on socioeconomic status and educational attainment were not recorded, and we therefore cannot be sure that this pattern of results would hold across any group of autistic individuals. Further, the small sample of women in our population does not provide sufficient power to determine whether any of these effects interact with gender.

Chapter 7: Effects of age and gender on face perception and memory

An edited or partial version of this chapter has been published as the following journal article:

- Stantić, M., Hearne, B., Catmur, C., & Bird, G. (2021). Use of the Oxford face matching test reveals an effect of ageing on face perception but not face memory. *Cortex*, *145*, 226-235.

7.1. Abstract

Effects of ageing on both face perception and face memory have previously been reported. Previous studies, however, have not controlled for the effects of face perception when assessing face memory, meaning that apparent effects of ageing on face memory may actually be due to effects of ageing on face perception. Here, both face perception and face memory were assessed in a sample of adults ranging in age from 18 to 93, and the effect of age on face memory was assessed after controlling for face perception. Face perception was assessed using both a standard test and the Oxford Face Matching Test (OFMT), deliberately designed to avoid the bias towards younger, neurotypical samples that may be present in other tests. An effect of ageing on face perception was found using both tests, with the unbiased OFMT being more sensitive to the effect of age. Importantly, when controlling for face perception using the OFMT, no effect of age on face memory was found. Indicative scores on the OFMT from a sample of 989 participants are provided, broken down by age and gender.

7.2. Introduction

It has been widely reported that various aspects of cognition decline with age, including aspects of memory (Naveh-Benjamin et al., 2000; Light, 1991; Bowles et al., 1985) and visual processing (Roberts & Allen, 2016). Of particular interest, given its importance to normal social interaction, is an understanding of face perception and face memory processes in older adults, both of which have been studied extensively. Anecdotally, older adults complain of difficulties in recognizing people and 'matching' a person's face to their name (Schweich et al., 1992). Empirically, findings generally support an age-related decline in both face memory (Maylor, 1990; Maylor & Valentine, 1992; Crook & Larrabee, 1992; Duchaine, Germine, & Nakayama, 2007) and face perception (Duchaine, Germine, & Nakayama, 2007; Shah et al., 2015; Megreya & Bindemann, 2015, Habak, Wilkinson, & Wilson, 2008). Problematically, however, as far as we are aware, no previous study has attempted to examine the age-related decline in face memory after controlling for face perception. This is an issue because the cognitive processes involved in face perception tasks are also required for face memory tasks. Face perception tasks require a perceptual representation of each face to be formed and a judgement made as to whether the two faces are of the same individual. A face memory task, such as naming the face of a famous individual, requires forming a perceptual representation of the stimulus face and a judgement made as to whether that face matches faces of famous individuals stored in memory. Therefore, in order to discover whether age has an effect on face memory independent of its effect on face perception, one must control for face perception. If one does not, an apparent decline in face memory due to ageing may actually result from a decline in face perception, with face memory unaffected.

This discussion highlights the fact that face matching is required when performing both face perception and face memory tasks. Our representations of identities (e.g., "the person in front of me is Mary") are flexible and allow for storage of multiple exemplars, or facial images, of the same person (e.g., "Mary's hair is different than when we last met"). The process of face matching involves deciding whether a particular facial image is sufficiently similar to a facial identity for it to be an image of that identity. This process may be subject to systematic biases or random noise, and either biases or noise may scale with age. Indeed, there has been some suggestion that the decline in performance in older age could be related to an increased rate of false alarms in older adults (Shapiro and Penrod, 1986; Bastin & Van der Lindon, 2003; Bartlett et al., 1989). This would manifest as an increased likelihood of stating that a particular image is an exemplar of a given facial identity when in fact it is not (i.e., a bias towards the 'match' response in experimental tasks); or, in day-to-day experience, an increase in mistaking strangers for acquaintances.

Much like other atypical groups (e.g., those with developmental prosopagnosia or autism) who may use different strategies to process faces (Gauthier et al., 2009; Joseph & Tanaka, 2003; Avidan et al., 2011; DeGutis et al., 2012; Palermo et al., 2011), there have also been some reports of older neurotypical adults using different face processing strategies from younger adults. There is evidence for changes in the degree of holistic processing (increase (Konat, Bennett, & Sekuler, 2013) or decrease (Schwarzer, Kretzer, Wimmer, & Jovanovic, 2010; Wiese, Kachel, & Schweinberger, 2013)), or a differential reliance on internal vs external features (Meinhardt-Injac, Persike, & Meinhardt, 2014; Schwarzer, Kretzer, Wimmer, & Jovanovic, 2010) in older adults, but these effects are not always observed (Boutet & Meinhardt-Injac, 2019; Meinhardt-Injac, Boutet, Persike, Meinhardt, & Imhof, 2017; Menihardt-Injac, Persike, & Meinhardt, 2014). Given the possibility of a

change in how faces are processed with age, it is of particular importance that a sensitive and unbiased test is used to compare the face perception of older and younger adults. We suggest that using the recently developed Oxford Face Matching Test (OFMT; Stantić et al., 2022a) in an ageing population could assist answering many of the outstanding questions raised above. Accordingly, this study reports the face perception and face memory performance of a group of adults in which ages ranged from 18 to 93 in order to investigate the specific subprocesses affected in the ageing population using a more sensitive measure.

The study also had secondary aims of further validating the OFMT against the GFMT and CFMT in a large, age-diverse sample, and examining the effect of gender on OFMT performance. With respect to the effect of gender, previous research consistently indicates a small but significant difference in performance such that females tend to be better than males on both tasks of face memory and face perception (Lott et al., 2006; Duchaine & Nakayama, 2006a; Bowles et al., 2009; Duchaine et al., 2007). Finally, indicative OFMT scores organised by gender and age for almost one thousand participants are provided in order to aid interpretation of OFMT scores in future studies.

7.3. Methods

7.3.1. Participants

607 participants completed the task online, after which 16 participants were excluded for attention check failures and 8 additional participants were excluded for being potentially prosopagnosic (scoring both above 72 on the PI-20 and below 42 on the CFMT)**.** In the final sample there were 355 females (M_{age} = 47.4, SD_{age} = 17.6, range = 18-83), 224 males

(M_{age} = 53.8, SD_{age} = 15.6, range = 19-93), and 4 participants who would rather not say (M_{age} = 37.0, SD_{age} = 19.1, range = 26-65). Although age was used as a continuous variable across all analyses, stratified recruitment was used to ensure at least 90 participants in each decade-wide (X0-X9) age bracket up to 70 years old, with those over 70 treated as one group for the purposes of recruitment. After exclusions, all reported analyses include 90 participants in the 18-29 age group, 82 participants in the 30-39 age group, 88 participants in the 40-49 age group, 149 participants in the 50-59 age group, 87 participants in the 60-69 age group, and 87 participants in the 70+ age group. Gender distribution was not equalised across age brackets and the proportion of males varied between 15% and 56%. Participants were recruited through a combination of recruitment platforms, including social media, the recruitment platform Prolific, and participant databases.

7.3.2. Tasks

All tasks were included as described in Chapter 2. Additionally, an adapted measure of visual acuity (Heatherington, 1954) was taken to determine the participants' ability to see images clearly.

Visual acuity chart (Hetherington, 1954)

At the end of the study, participants were shown a standard Snellen visual acuity chart and asked to input the last line they could read without zooming in or getting closer to the screen. An exclusion criterion was defined to remove data from any participants whose scores were not within the 25-75th percentile norms for their age (none met this criterion). A significant correlation between age and visual acuity was observed (r = .40, p < .001),

however inclusion of visual acuity as a control variable did not alter the pattern of significance in any of the analyses reported below.

7.3.3. Procedure

The procedure was as outlined in Chapter 2 (2.3. Common Methodology).

7.3.4. Analysis Strategy

Analysis strategy was as outlined in Chapter 3 (3.3.3. Analysis Strategy).
Further, we determined whether age impacts face perception, face matching (independent of face perception), face memory (independent of face perception and face matching) using scores as calculated in 3.3.3. Analysis Strategy using a series of regression analyses in which age, predictors (e.g., face perception), and their interactions, were used to predict test scores (e.g., face matching). Including the interaction term in regression models allows for the relationship between, e.g., face perception and face matching, to vary across age.

7.4. Results

Mean, standard deviation of the mean, and the range, of scores obtained on each test separated by gender is reported in Table 7.9.

In order to determine the effect on age on face recognition a hierarchical multiple regression was conducted for each test which included linear, quadratic and cubic effects of age on task performance. The best fitting model is reported for each face processing measure.

7.4.1. Effect of Age on Face Perception

OFMT: The effect of age on OFMT performance was best explained by a regression model containing both linear and quadratic effects of age. A linear model showed age significantly predicted performance (β = -.419, t = -11.11, p < .001, R^2 = .175). The addition of the quadratic effect significantly increased the amount of variance explained (quadratic effect: β = -.116, t = -3.08, p = .002, R^2 = .189, R^2_{change} = .014), but the addition of the cubic effect did not (cubic effect: β = .050, t = .59, p = .553, R^2 = .189, R^2_{change} = .00). The model containing both linear and quadratic effects therefore showed that age was a significant predictor of OFMT performance (F(2, 579) = 67.33, p < .001) and this model explained 18.9% of the variance in performance, see Figure 7.14., upper left panel. Age was significantly correlated with the rate of false alarms (r(582) = .29, p < .001) and misses (r(582) = .17, p < .001) on the OFMT, with the correlation between age and false alarms significantly greater than the correlation between age and misses (z = 2.31, p = .02).

GFMT: The effect of age on GFMT performance was also best explained by a regression model containing both linear and quadratic effects of age. A linear model showed age significantly predicted performance (β = -.212, t = -5.22, p < .001, R^2 = .043). The addition of the quadratic effect significantly increased the amount of variance in performance explained (quadratic effect: β = -.177, t = -4.42, p < .001, R^2 = .073, R^2_{change} = .031), but the addition of the cubic effect did not (cubic effect: β = -.126, t = -1.39, p = .164, R^2 = .074, R^2_{change} = .003). The model containing both linear and quadratic effects therefore showed that age was a significant predictor of GFMT performance (F(2, 579) = 23.83, p < .001) and this model explained 7.6% of the variance in performance, see Figure 7.14., upper right panel. Given that GFMT allows for unlimited viewing time, it is possible for

the errors to be driven by faster response times (i.e., that older participants show a speed-accuracy trade-off such that they sacrifice accuracy for increased speed). However, the data indicate that the reduced accuracy with age could not be due to a speed-accuracy trade-off, as reaction time increased with age in a linear fashion ($\beta = .30$, $t = 7.65$, $p < .001$, $R^2 = .092$). In contrast to the OFMT, age was correlated with the rate of misses ($r(582) = .30$, $p < .001$) but not false alarms ($r(582) = -.02$, $p = .624$) on the GFMT, and the correlation between age and misses was significantly greater than that between age and false alarms ($z = -4.863$, $p < .001$).

Formal comparison of correlations (OFMT:age and GFMT:age) revealed that the OFMT was more sensitive to the linear effect of age than the GFMT ($z = 3.88$, $p < .001$). However, there was no significant difference in sensitivity between tests when comparing correlations with the quadratic effect of age ($z = 1.4$, $p = .16$).

7.4.2. Effect of Age on Face Memory

The effect of age on face memory was assessed in the same manner as the effect of age on face perception, using CFMT scores as the dependant variable. Results revealed that the effect of age on CFMT performance was also best explained by a regression model containing both linear and quadratic effects of age. A linear model showed age significantly predicted performance ($\beta = -.249$, $t = -6.21$, $p < .001$, $R^2 = .062$). The addition of the quadratic effect significantly increased the amount of variance explained by the model (quadratic effect $\beta = -.131$, $t = -3.28$, $p = .001$, $R^2 = .079$, $R^2_{change} = .017$), but the addition of the cubic effect did not (cubic effect: $\beta = -.104$, $t = -1.15$, $p = .249$, $R^2 = .077$, $R^2_{change} = .002$). The model containing both linear and quadratic effects therefore showed that age was a significant predictor of CFMT performance ($F(2, 579) = 24.96$, $p < .001$),

and this model explained 7.9% of the variance in performance, see Figure 7.14., lower left panel.

7.4.3. Effect of Age on Self-Reported Difficulties with Face Recognition

A hierarchical multiple regression revealed that PI-20 scores were best explained by a linear effect of age ($\beta = .153$, $t = 3.73$, $p < .001$, $R^2 = .023$), see Figure 7.14., lower right panel.

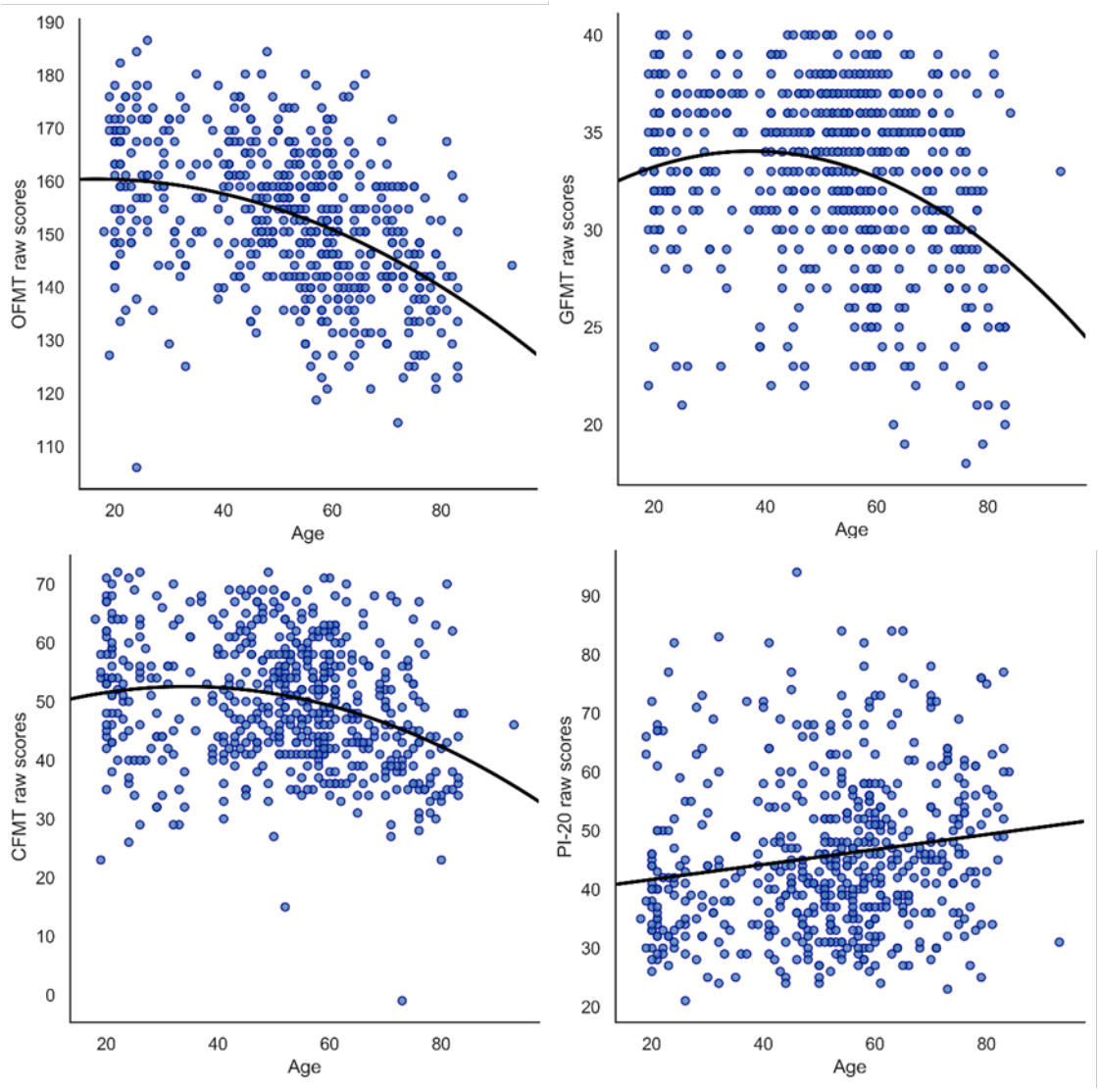


Figure 7.14. Performance on all face perception measures shown as a function of age with the best fit line for the model that explains most variance.

Upper left: Oxford Face Matching Test (OFMT) performance shown as a function of age with the best fit line for the quadratic regression model; Upper right: Glasgow Face Matching Test (GFMT) performance shown as a function of age with the best fit line for the quadratic regression model; Lower left: Cambridge Face Memory Test (CFMT) performance shown as a function of age with the best fit line for the quadratic regression model. Note that the relationship between age and CFMT performance does not hold after accounting for the effect of age on face perception using the OFMT; Lower right: 20-Item Prosopagnosia Index (PI-20) scores shown as a function of age with the best fit line for the linear regression model, note that higher scores on the PI-20 indicate a higher incidence of self-reported face difficulties in everyday life.

7.4.5. Effect of Face Perception on Face Matching

Age, Deviation scores, and their interaction were entered into two regressions, the first predicting OFMT matching accuracy and the second GFMT matching accuracy. For the OFMT analysis, Deviation scores were a significant predictor (β = -1.85, t = -4.77, p < .001) of OFMT matching accuracy, suggesting that face perception ability is related to OFMT performance. Age was not a significant predictor of OFMT matching accuracy (β = -.06, t = -.32, p = .75), indicating that face matching was not related to age after accounting for face perception ability. The interaction between Group and Deviation scores was also not a significant predictor (β = -.01, t = -.45, p = .65).

This same analysis was repeated using GFMT instead of OFMT scores as a robustness check. Age was again not a significant predictor of GFMT matching accuracy (β = .13, t = 1.68, p = 0.10), suggesting that face matching was not related to age after accounting for face perception ability. Deviation scores were also not a significant predictor (β = 0.08 , t = 0.49, p = .62), but the interaction between Age and Deviation was significant (β = .01, t = -2.21, p = .03), suggesting that the relationship between face perception and face matching on the GFMT varied as a function of age.

7.4.6. Effect of Face Perception and Face Matching (Independent of Perception) on Face Memory

Age, Deviation scores, Face Matching scores (OFMT matching accuracy, and separately GFMT matching accuracy) and the interactions between Deviation scores and Age, and Face Matching and Age, were entered into regressions predicting CFMT scores.

For the OFMT analysis, results showed no significant independent contribution of face perception (as measured by Deviation scores; β = -.07, t = -.15, p = .88) or age (β = .01, t = .02, p = .99). The two interactions between Face Matching x Age and Deviation scores x Age were not significant (Face Matching and Age: β = .01, t = .43, p = .66; Deviation scores and Age: β = -.01, t = -.76, p = .45). The independent contribution of face matching was significant (β = .29, t = 2.59, p = .01), indicating that face matching was related to face memory after accounting for face perception ability.

Similarly, for the analysis using GFMT scores, results showed no significant independent contributions of face perception (as measured by Deviation scores; (β = -.43, t = -1.17, p = .24) or age (β = .52, t = 3.32, p = .06). The two interactions between Age x Face Matching scores and Age x Deviation scores were not significant predictors (Age x Face Matching: β = -.01, t = -1.86, p = .06; Age x Deviation scores: β = -.01, t = -1.44, p = .15). However, the independent contribution of face matching was again a significant predictor (β = .95, t = 1.93, p < .001).

7.4.7. Effect of Gender – All Measures

See Table 7.9. for gender-specific means on all measures. A small, but significant effect of gender was observed for OFMT performance (t(577) = 3.76, p < .001, d = 0.32), whereby females scored higher than males. There was no interaction between age and gender (F(61, 449) = 1.03, p = .41). The female advantage was also significant for the GFMT t(577) = 3.24, p = .001, d = 0.27), CFMT (t(577) = 2.45, p = .014, d = 0.21), and PI-20 (t(577) = -2.11, p = .035, d = 0.18). There were no interactions between age and gender for any of the other tests: GFMT (F(61, 449) = 1.23, p = .12); CFMT (F(61, 449) = .91, p = .68); or PI-20 (F(61, 449) = 1.04, p = .41).

Table 7.9. Descriptive statistics for all participants across all face perception measures.

The data here showing average performance, standard deviation, and range of scores for all face perception tasks (OFMT, GFMT, CFMT and PI-20) split by gender. The somewhat lower CFMT scores observed here compared to previous studies are likely driven by the increased representation of participants over 60 years of age, who score below previously reported mean scores (M=46.78).

	M*±*SD	**Range [min-max]**
OFMT	153.80±13.06	110-190
Male	151.29±12.97	110-181
Female	155.42±12.77	121-190
Other	170.00±18.08	151-187
GFMT	32.92±4.48	18-40
Male	32.15±4.75	18-40
Female	33.38±4.25	20-40
Other	36.67±3.51	33-40
CFMT	50.43±10.40	15-72
Male	49.05±10.28	27-72
Female	51.22±10.36	15-72
Other	60.33±13.20	46-72
PI-20	45.05±12.65	21-94
Male	46.44±12.76	24-84
Female	44.16±12.57	21-94
Other	47.00±5.29	43-53

7.4.5. Relationships Between OFMT and Other Face Processing Measures

OFMT scores showed moderate-to-strong positive relationships with GFMT (r = .45, p < .001) and CFMT (r =.50, p < .001) scores, as previously found in the initial validation of the OFMT (Stantić et al., 2022a). Correlations between all other tests (including OFMT scores on match and mismatch trials separately) are shown in Table 7.10.

Table 7.10. Relationships between all face perception measures for all participants.
Two asterisks denote significance at the level of p < .001. CFMT = Cambridge Face Memory Test, PI-20 = 20-Item Prosopagnosia Index, GFMT = Glasgow Face Matching Test, OFMT = Oxford Face Matching Test, Match = accuracy on same/match trials of the OFMT; Mismatch = accuracy on different/mismatch trials of the OFMT.

	M	***SD***	**CFMT**	**PI-20**	**GFMT**	**OFMT**	**OFMT match trials**	**OFMT mismatch trials**
CFMT	50.43	10.40	-					
PI-20	45.05	12.65	-.26**	-				
GFMT	32.92	4.48	.41**	-.28**	-			
OFMT	153.90	13.05	.50**	-.28**	.45**	-		
Match	71.95	13.72	.30**	-.19**	.20**	.60**	-	
Mismatch	81.50	11.98	.19**	-.09*	.27**	.39**	-.49**	-

OFMT Indicative Scores by Age and Gender

The sample described above was combined with previously collected data from 406 neurotypical adults (276 female, 130 male) making a combined sample of 989 individuals (631 women, 354 men, 4 other/would rather not say). Mean (and SD) OFMT scores from

this sample are reported below organised by age, and further subdivided by gender (Table 7.11.).

Table 7.11. Distribution of OFMT scores by age and gender.

Age brackets are defined to be inclusive (e.g., participants in the age bracket 30-39 include both 30-year-olds and 39-year-olds.) Participants who defined their gender as 'other' or 'would rather not say' are included in the average age scores but not the gender-specific scores.

		n	**OFMT** ***M*±*SD***	**OFMT Range**
18-29		409	157.12±13.23	107-187
	Male	101	150.75±13.85	107-179
	Female	305	159.10±12.26	107-187
30-39		113	154.18±14.71	104-180
	Male	42	148.83±15.83	104-179
	Female	70	157.00±12.90	119-180
40-49		119	156.45±12.25	126-190
	Male	64	153.97±11.94	126-181
	Female	55	159.33±12.08	128-190
50-59		160	153.01±12.37	119-181
	Male	53	152.94±10.70	122-174
	Female	106	153.07±13.22	119-181
60-69		91	148.63±12.53	120-180
	Male	43	148.35±12.42	120-176
	Female	47	149.28±12.58	128-180
70+		93	142.92±11.24	115-171
	Male	45	141.76±12.18	115-171
	Female	48	144.02±10.29	123-168

7.5. Discussion

This study aimed to test the effect of ageing on face perception using both a standard test (GFMT) and a novel test (OFMT), which relies on unbiased algorithmic judgments to determine difficulty of items. The effect of ageing on face memory was also tested using the CFMT. Crucially, the effect of age on face memory was assessed both before, and after, accounting for any effects on face perception. The study also assessed the relationship between the OFMT and the other face measures, and the effect of gender on all measures.

With respect to face perception, both the OFMT and GFMT revealed a linear and a quadratic relationship with age whereby the effect of age on face perception increased as age increased (i.e., the rate of age-related decline in face perception ability increased as age increased). The OFMT proved to be the more sensitive measure. The increased sensitivity of the OFMT likely stems from at least three factors: its larger number of trials compared to the GFMT, the absence of a celling effect for superior performers, as well as its use of algorithmic similarities to determine the difficulty of individual items.

The analysis investigating independent contributions of face perception to face matching and face perception and face matching on face memory revealed a unique pattern for the ageing neurotypical population. Face perception impacted face matching using both OFMT and GFMT scores, but it was only in the GFMT analysis that the interaction between age and perception was also significant, indicating that the relationship between perception and matching varied as a function of group. Previous analysis (see Figure 7.14.)

indicated that this is likely driven by ceiling effects on GFMT performance in the younger groups (which was not present using the OFMT).

Importantly, face memory was not affected by age after controlling for face perception and face matching. The CFMT scores showed an independent contribution of matching ability to face memory, but age was not a significant factor after accounting for face perception and face matching. To our knowledge, this is the first study to examine the impact of ageing on face memory independent of face perception, and so this result is in need of replication. However, if reliable, this result suggests interventions to improve face recognition in older adults should primarily be targeting face perception.

As outlined in the Introduction, the matching, or decision-making aspect, of these tasks is also potentially susceptible to effects of ageing. There have been previous suggestions that the rate of false alarms increases in age (Shapiro & Penrod, 1986; Bastin & Van der Lindon, 2003; Bartlett et al., 1989), and this effect was partially supported in these data. Age was significantly associated with an increased rate of both false alarms and misses on the OFMT, but the correlation between age and false alarms was stronger than between age and misses. On the GFMT, however, the correlation with age was significant only for misses and not false alarms, with the correlation between age and misses significantly stronger than that for age and false alarms.

With respect to OFMT validation, as expected, correlations between OFMT scores and the GFMT and CFMT were moderate-to-strong. Data from a combined sample of nearly one thousand participants were used to provide indicative scores for various age groups, and for males and females within those age groups, on the OFMT. Furthermore, a small but

significant effect of gender was observed for OFMT scores (and all other tests) indicating a female advantage. McKelvie (1981) proposed that gender differences in face processing are the result of different interests, suggesting that men are less interested in faces than women and therefore worse at processing them. However, research on a range of cognitive tasks tends to find a slight female advantage in performance (Van Exel et al., 2001), a result interpreted as suggesting that females try harder on experimental tests than males. Thus, females are either slightly better at face processing than males, and/or put more effort into taking part in experimental studies in general.

It is worth noting a potential limitation with the study relating to the age of the stimulus faces. Stimulus faces in the GFMT are on average approximately 23 years old, and the CFMT includes faces of approximately the same average age. Stimulus faces in the OFMT are of a wider range of ages, resulting in an older average age of approximately 45 years. It has previously been shown that, similar to the other-ethnicity effect (Meissner & Bringham, 2001), an equivalent age-effect exists such that older adults recognize faces of other older adults better than those of younger adults (Lamont, Stewart-Williams, & Podd, 2005; Kuefner et al., 2008). Therefore, it is possible that the relatively young age of stimulus faces in the GFMT and CFMT explain some of the age effects on those tasks, such that as the age of the participant gets further from the age of the stimulus faces, performance decreases. This cannot account for the age effect on the OFMT however, as the age of the stimulus faces corresponds to the mean age of participants (meaning that performance would be impaired at both the younger and older ends of the age spectrum).

A more substantial limitation of this study is that the specificity of any effects of ageing with respect to stimulus class cannot be established, i.e., it cannot be ascertained whether

the impaired face perception seen in older adults is specific to faces, or whether similar effects would be observed for comparable non-face stimuli. There are certainly effects of ageing on visual perception in general, including on colour perception, temporal resolution, visual acuity, motion perception, and a loss of high spatial frequencies (see Roberts & Allen, 2016, for a review), making it plausible that age-related decline in face perception is the result of more general effects of ageing on visual perception. In addition, due to the length of the testing session, no other measures of cognitive decline were included in this study, which may have identified any abnormal cognitive decline in the older participants not captured by a simple visual acuity test. It is, of course, likely that general cognitive decline associated with normal ageing affects both face matching and face memory, as well as a number of other cognitive capacities.

Further limitations of the study are shared with all studies that adopt a similar design. Most obvious is the fact that cross-sectional studies cannot separate true effects of age from cohort effects. Also, due to the online nature of our study, the test of visual acuity had to be conducted remotely. This meant it was not possible to control testing conditions as much as in the lab. Furthermore, it is possible that older people have less access to technology than younger individuals. Older individuals do not tend to use computers as much as younger individuals, and tend to perceive their use as less relevant to their everyday lives (Selwyn, 2004) and be more anxious about, or less interested in, using computers (Laguna and Babcock, 1997; Van Deursen & Helsper, 2015). Consequently, the online nature of this study may have served as more of a selection criterion in our older participants such that only older adults who are more confident with technology took part. To avoid this as much as possible, extensive support was offered in setting up the experiment, and no feedback was received to indicate that the use of technology had been

an obstacle to taking part, or that participants struggled to understand the instructions for any of the tasks.

In conclusion, using an unbiased and sensitive test of face perception, a significant reduction in face perception with age was observed, but this was not accompanied by an effect of age on face matching nor an effect of age on face memory. Instead, results suggested that effects of age on face memory and face matching were explained by effects on face perception. Distributions of scores on the OFMT broken down by age and gender are provided in order to facilitate comparisons in future studies of social cognition across the lifespan.

Chapter 8: General Discussion

In this dissertation I presented a series of experiments that explore the relationship between face perception, face matching and face memory in typical and atypical populations. This work sought to further scientific understanding of how we perceive and recognize faces in everyday life, particularly focusing on wide individual differences observed in the general population on all face recognition tasks. Thus far, the field of face perception and recognition has seen vastly different batteries of tests deployed to capture various aspects of the recognition process. The resulting findings are heterogeneous and oftentimes contradictory, with gaps in what the existing tests can measure, as well as major outstanding debates on seemingly basic questions, such as whether face perception and memory are both affected in atypical groups. or whether these abilities dissociate in typical groups.

This thesis sought to address these two outstanding questions in the field. Specifically, by creating and validating a new, more sensitive and objective test of face perception and matching (**Chapter 2**), I was able to investigate independent contributions of face recognition subprocesses to face recognition. This furthers our understanding of how each of the stages of the process functions in typical and atypical groups. I focused on understanding advantages and limitations of the novel test (**Chapter 2**), its performance over multiple administrations and ability to capture individual differences in typical participants (**Chapter 3**), and atypical groups, including people with developmental prosopagnosia (**Chapter 4**), super recognition (**Chapter 5**), autism (**Chapter 6**), and the elderly (**Chapter 7**). In these groups, I also utilized a novel analysis strategy to understand

how face perception and face matching interplay, as well as how differences in these abilities influence face memory performance.

This chapter will provide a summary of all findings across reported experiments and draw conclusions across chapters. This will allow an integrated account of how existing theoretical models of typical and atypical face recognition ought to be updated. It is crucial that we understand the nature of these differences on both extremes of the spectrum. Only by doing so will we be able to design appropriate applications, from those in national security and policing to those in social and educational integration.

In the next section, a thematic summary is provided to organize key findings of the presented empirical chapters, followed by their implications for theory and research. After that, I will outline limitations and future directions. Finally, I outline opportunities to extend work in applied domains for both poor and exceptional performers.

8.1. Development of better tests

A thorough review of existing face perception tests revealed clear drawbacks to using any of them to study the whole population. Limitations of existing tests grouped around three issues. The first is that while some tests were designed to capture performance at one end of the spectrum – poor or superior – they lacked sensitivity when measuring performance in typical groups. Tests designed to capture performance in typical groups exhibited ceiling or floor effects, and thus were unable to distinguish between individuals at the ends of the typical distribution and outlying individuals. Second, tests were biased in that they were developed to be sensitive to variance in neurotypical populations. Third, none of

these tests had the ability to independently assess perception and matching. This made it impossible to determine the relationship between these two abilities, particularly in atypical groups, and between these abilities and face recognition.

I presented the development of the Oxford Face Matching Test, a novel measure of face perception performance that relies on objective algorithmically derived similarity of faces to determine test item difficulty, to overcome these problems. Reliance on algorithms for test development meant the assessment of item difficulty was not biased to neurotypical populations. The test was validated for use in typical and atypical groups, as well as tested for reliability over time.

8.1.1. Findings and implications

As discussed in Chapter 2, results showed that the test is sensitive enough to both capture individual differences in a typical group and measure performance when administered in person and remotely, as well as in atypical groups (Chapters 4, 5, and 6). It was notable that the OFMT does not have issues of ceiling or floor performance, with no participants scoring at chance level or at 100% accuracy across more than a thousand participants who completed the test to date. This was particularly pertinent given that the test was deployed in large groups of independently diagnosed prosopagnosics and super recognisers, both groups that are difficult for researchers to study because existing tests often do not have enough sensitivity to differentiate performance within the group because lots of participants score at floor or ceiling. The OFMT was also found to show consistent results over time, with good reliability over two testing sessions (Chapter 2).

Further, given the novel approach to making the OFMT, a relationship between the algorithmically derived and human-provided similarity of face pairs was investigated. A moderate correlation was observed. This suggests that algorithmic similarity was not simply derived by using a human-like perception mechanism, and that instead the convolutional neural networks utilized for these ratings use different mechanisms. That the relationship between human and algorithmic similarity was not found to be stronger further emphasizes the importance of using unbiased, objective metrics to determine difficulty, and subsequently pick test items in the future. This is because a moderate relationship indicates that neurotypical humans might deploy different strategies to determine similarity, so relying on their performance could therefore bias the choice of items to neurotypical judgments. Using algorithms instead of typical human raters, regardless of the size and diversity of the sample, reduces bias to neurotypicality and makes the test more appropriate for application in atypical groups.

8.1.2. Limitations and future directions

In practical terms, using algorithms instead of large samples of typical participants allows for development of multiple parallel forms of the test. This could be important in many research settings. For instance, application in cultures where faces are not predominantly white would benefit from a test version being developed using faces that better represent the local demographic (Lindsay, Jack, & Christian, 1991; Valentine & Bruce, 1986; Ellis & Deregowski, 1981). Because we know that there is a strong effect of age on perception showing that humans are very good at recognizing faces up to about their own age, and poor at recognizing faces older than their own age (Anastasi & Rhodes, 2006), testing children on OFMT-like measures should be adjusted accordingly as well.

There are also, of course, disadvantages to developing tests using this method. Most importantly there is, of course, the concern that choice of algorithms itself could be biased. Specifically, if one were to use a single algorithm, any biases present in the training set of that algorithm, and accordingly, the model that determines similarity between two faces, would be potentiated in the choice of items included in each difficulty bin, and therefore available for test inclusion. This could be an issue even if multiple algorithms were used if those algorithms utilized similar strategies to determine similarity, therefore greatly reducing the benefit of using multiple assessment models. Finally, using a group of algorithms with a sole algorithm utilizing a substantially different strategy for determining similarity would possibly result in mathematical irregularities when similarities were averaged across models.

These are addressable challenges. Here I have decided to include three different, independently developed algorithms to combat possible issues with any individual model. To alleviate some of the concerns outlined above, I included algorithms that have not used a shared database of images in their training sets nor relied on publicly available code (each algorithm was developed for commercial deployment). Further, I chose these three algorithms because they were amongst those that scored best in the independent testing conducted by the U.S. National Institute for Standards and Technology. Inclusion of additional algorithms into the similarity assessment stage would have obviously alleviated further concerns, though marginal returns of including further algorithms likely decline quickly. It is crucial, however, that researchers always include more than one algorithm, that they seek to include ones that perform well in independent testing, and that they average out performance of as many algorithms as is feasible.

Furthermore, additional checks should be deployed to ensure that oddities in strategy by any one model on a particular test item do not influence the researchers' choice of items included in the test itself. For instance, one could choose to include only those pairs whose similarities do not significantly deviate between algorithms. In the case of the OFMT, I deployed a reasonably conservative check such that pairs were only kept in consideration if they were rated as quite similar by all three algorithms (allowable difference was defined as 10%). A researcher looking into a different research question might be keen to allow more variation amongst algorithms but integrating some level of checking for outlying ratings ought to be implemented.

In recent years, application of these algorithms has been questioned after reports of large differences in accuracy levels across ethnicities and genders (Mazura et al., 2012; Klare et al., 2012; Irhebhude et al., 2019; Buolamwini & Gebru, 2018). Critics have suggested that the accuracy of face recognition algorithms as a whole ought to be questioned, as well as that using them in real-world applications is inappropriate until training sets more accurately represent said real world and resulting models show no bias in accuracy. These are valid concerns that ought to be considered carefully when algorithms are deployed in research. However, given their wide application in everyday life, these models are now subject to thorough scrutiny. As a result, they are regularly tested for effectiveness and bias during objective challenges organized by independent government agencies (e.g., Recognition Grand Challenge, Multiple Biometric Grand Challenge, Face Recognition Vendor Test). It is promising that recent tests showed no notable differences in accuracy of recognition across different ethnicities for leading algorithms (McLaughlin & Castro, 2020). One expects that algorithmic accuracy will continue to improve across demographic

groups, and that the most accurate models will be the most commercially marketable ones, deployable across a wider range of situations. However, the issue of differences in accuracy and poorly designed training sets is an important one that warrants continuous scrutiny and challenging of assumptions.

8.2. Towards understanding individual differences in typical groups

With the recent shift in research towards studying individual differences, one of the crucial aims of the OFMT was to have the ability to measure individual differences in perception and memory. I sought to develop a test that would include enough items to meaningfully measure the performance of all typical perceivers while avoiding issues of floor or ceiling effects.

8.2.1. Findings and implications

Chapters 2 (The development and validation of the OFMT) and 3 (The individual differences in neurotypical population) show that this effort was successful. The Long-Form OFMT can be administered both online and in-person with no notable differences in patterns of results. Importantly, a very large sample of over a thousand participants of various demographics who completed the OFMT gives confidence that the test is well-balanced in difficulty of both match and mismatch items.

Relationships between the OFMT and other tests used throughout this dissertation (GFMT, CFMT, PI-20) were consistent across samples and relationships between other tests in line

with published literature (Bobak et al., 2016; Bobak, Dowsett, & Bate, 2016; Burton et al., 2010; Russell et al., 2009; Shah et al., 2015; Balsdon et al., 2021; Dunn et al., 2020). A known finding was replicated showing very small relationships between objective face processing performance and scores on PI-20, the self-report measure of own awareness about face recognition ability, in the typical population (Palermo et al., 2017; Bobak, Mileva & Hancock, 2019; Arizpe et al., 2019). The strength of this relationship increased with inclusion of atypical samples. This is probably explained by a previously documented improved ability of atypical groups to assess their own performance (possibly driven by interactions with researchers during diagnostic batteries in which the results of these tests were disclosed). Importantly, the stability of other relationships (e.g., the OFMT with the CFMT and GFMT) in typical samples reported throughout the dissertation indicates that there is a stable amount of shared variance in these tasks.

Finally, the use of OFMT in a large neurotypical sample revealed important patterns about the relationship between face perception and matching, two abilities that have not traditionally been easily empirically dissociable. I found a wide range of differences in how much participants deviated from algorithmic similarities when providing formal similarity judgments of face pairs. Importantly, this metric was a strong predictor of matching performance on both the OFMT and the GFMT. The relationship revealed that those who deviated more from the algorithms were less accurate on the matching tasks. This likely indicates that the ability to accurately and precisely judge similarity of two faces is causally related to one's ability to determine whether two faces are the same or different (i.e., face matching). Indeed, it is possible that the process of face matching involves a distinct step of determining face similarity and subsequently using that similarity as an index to make the same/different decision (see Figure 8.15. for suggestion

of mechanism). Of course, there will still be some variance in face matching performance even for individuals who show equal face perception accuracy, which might be explained by different thresholds for what constitutes 'sameness', or a more accurate translation from similarity to the likelihood of two face images being of the same identity.

Data also showed that face memory in typical groups was a function of both accurate face perception, and face matching measured independently of face perception. This finding has two important implications. Firstly, it strengthens the theoretical account that suggests that remembering a face involves a distinct step of comparing an exemplar of a face in front of you to a stored representation in memory (face matching), which is further reliant on accurately perceiving the face so that it can be appropriately compared to the stored representation (see sketch of a model in Figure 8.15.). Secondly, this finding suggests that difficulties in face memory could stem from a number of prerequisite processes. Assessing individual differences in face memory would therefore benefit from including tests of perception and matching to more accurately define the profile of an individual's deficit.

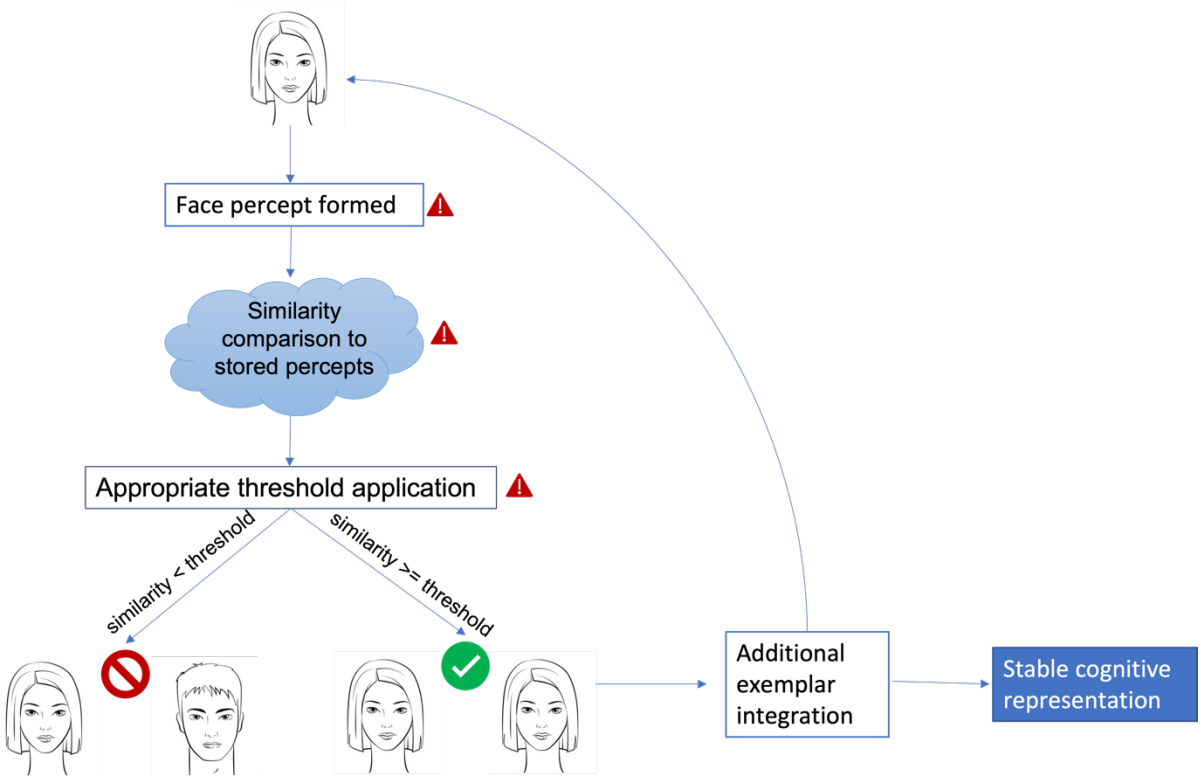


Figure 8.15. A suggested update to the existing models of face recognition subprocesses.

I suggest that each encounter of a face prompts a formation of a face percept, which is compared in similarity to stored face representations. A set similarity index is required for the encountered face exemplar to be determined to be of the same person – if a similarity is judged to be lower, the face will be rejected as "different" than the stored representation, and if the similarity is at or above the threshold, it will be accepted as "same" as the stored representation. This additional exemplar is then integrated into the identity's representation, making it more stable over time. Issues can occur at various stages of this process (marked in Figure with red flags). For instance, people with autism might form an inaccurate percept due to their different face scanning patterns. Alternatively, wide individual differences have been observed in precision of similarity judgments between two face images in typical (and particularly atypical) populations. Finally, an inappropriate threshold for similarity might be applied, resulting in an erroneous acceptance of a face as the same or an erroneous rejection as different. Any of these would impact one's ability to form a more stable cognitive representation of a person's face over time.

8.2.2. Limitations and future directions

Of course, one limitation of this approach is that an overarching effort factor can explain these relationships. While the OFMT does deploy attention checks that have been shown to work very well at capturing no-effort responses, they are nevertheless not necessarily able to capture low-effort participants. It is possible that participants who are less engaged in the task provide less precise judgments of face similarity, therefore making their face perception performance seem poorer; meanwhile; the lack of engagement could be driving these participants to click at random more frequently, therefore making their face matching performance seem poorer. The challenge of implementing more sophisticated attention checks (e.g., by showing two very similar, but not identical images of a person), however, is that they might genuinely be too difficult for prosopagnosics and therefore erroneously exclude them from the study. A better solution might be an implementation of an independent measure of effort.

Further, reported literature, as well as present results, lack a good set of control measures utilizing other categories of visual stimuli (i.e., non-faces). As part of the investigations into face specificity of cognitive mechanisms, a number of other visual categories have been deployed in memory paradigms, such as cars, houses, horses, flowers, etc. (Dennett et al., 2012; Duchaine & Nakayama, 2005). However, due to lack of sensitivity (Gray, Biotti, & Cook, 2015; Geskin & Behrmann, 2020) and concerns about gender differences in performance, possibly driven by differential expertise (McGugin et al., 2012), these have been deployed sparsely outside of research on domain-specificity. As a result, there is a general lack of control tasks in the field. This leaves open the question of whether any observed differences in performance reflect a face-specific deficit or general impairments in perception, matching, or memory.

8.3. Towards understanding individual differences in atypical groups

This dissertation also set out to test sizeable groups of atypical performers (developmental prosopagnosics, super recognizers, people with autism, elderly adults) for two reasons. Firstly, understanding the performance of atypical groups was crucial for validating the novel test and understanding whether the included items were meaningfully able to measure performance at both extremes of the spectrum. Secondly, the debate about how memory and perception performance dissociate in these groups has a long history and the use of a particular set of tests in the same sample allowed us to address those questions. Importantly, the specific nature of the OFMT allowed dissociation between face perception and matching performances, providing increased granularity into our understanding of atypical performance.

8.3.1. Findings and implications

Participants at the lower end of the performance spectrum – those with autism and developmental prosopagnosia – showed a subtly different pattern of results when face perception, matching, and memory were independently assessed. At the group level, both populations showed consistently lower scores on all tests, including the CFMT, GFMT and OFMT, as well as on the PI-20. Further, both groups showed differences in perception such that their judgments of face pair similarity varied more from algorithmic similarity than those of matched neurotypical controls.

However, an interesting difference emerged in analyses that dissociate performance between perception and matching, as well as contributions of those two processes on face memory. Developmental prosopagnosics were impaired at every stage of the process, even when accounting for previous stages – they showed pervasive impairments in face matching after accounting for face perception, as well as impairments in face memory after accounting for face perception and face matching. People with autism, however, did not. They were also impaired on face perception, as noted above, but when that impairment was accounted for, showed no differences in face matching tasks. Importantly, their face memory remained impaired compared to controls when accounting for face perception and matching. This dissociation in findings challenges the previous assumption that nature of poor performance is the same in autism and prosopagnosia.

The dissociation also suggests that examination of potential subgroups of poor face recognition performers might be worthwhile. Much has been made of the heterogeneity of results in autism and prosopagnosia, but perhaps it actually reflects true subset differences that were difficult or impossible to capture with existing measures. If replicated and examined at an individual level, these results would indicate that the subgroups of developmental prosopagnosics might exist whose behavioural patterns align more closely with autistic participants, but that another subset of developmental prosopagnosics might not show the same profile.

Exploration of potential subtypes of autism and prosopagnosia is also interesting in the context of a wider understanding of these differences. For instance, there have been attempts to teach or train face recognition performance in prosopagnosia (DeGutis, Cohan, & Nakayama, 2014), but these have so far mainly focused on featural training processes

that were supposed to help individuals make more salient identity representations. However, our results indicate that early perception is already impaired in prosopagnosia, showing perhaps that any training programs need to be focused on an even earlier stage of the face recognition process. Further, our data showed that matching was impaired in prosopagnosia as well, even when perception performance was accounted for, indicating that prosopagnosic participants apply a different decision-making threshold when deciding if two faces are of the same person or different people. Given that we know, for instance, that pairs of people make better face recognition decisions than individual participants (Dowsett & Burton, 2015; Ritchie et al., 2022), future research might explore whether pairing developmental prosopagnosics with typical participants in training paradigms aimed to improve face perception results in improved face matching performance (or, subsequently, face memory performance).

With regards to people with autism, there have been debates about whether their deficits in face recognition are explained by other issues observed in autism, such as difficulties generalizing from exemplars or processing global shapes (Scherf et al., 2008) or difficulties integrating local features into a unified whole (Behrmann et al., 2006; Happe 1999; Happe & Frith, 2006). Current results suggested that there is a difference in similarity judgments between people with autism and neurotypical participants. This difference might result from people with autism focusing on different facial features or differential feature weighting (e.g., encoding more salient features of the face rather than integrating the local features into a unified whole). However, our data indicated that once these differences are controlled for, people with autism did not show different matching abilities. This implies that their threshold for deciding what constitutes sufficient evidence for deciding that two faces are of the same person or different people is not different from

that of neurotypical matched controls. This likely indicates two things. Firstly, issues with perception in autism might have to do with different patterns of scanning a face (Pelphrey et al., 2002; Joseph & Tanaka, 2003), resulting in less salient representations of a face. Importantly, issues with face memory (even after accounting for perception and matching performance) did not seem to be driven by a difficulty in matching an exemplar to a stored representation of an identity, given that we show matching itself is not impaired in autism. Instead, it is likely that differences in memory are driven by the memory demand of the task itself. This conclusion would be strengthened if a control memory task was available that uses a different visual category, as it would dissociate between difficulties with faces and general visual memory difficulties.

Our final group of interest comprised of a sizeable sample of elderly adults. Here, a clear pattern of poorer performance was observed, with elderly participants scoring more poorly on all tests (CFMT, OFMT, GFMT) as well as accurately reporting more difficulties with face recognition (using the PI-20 questionnaire). The relationship between age and performance on these tasks revealed that the rate of decline increases as age increased, indicating that decline was slower in younger elderly adults than in older groups.

However, when it came to dissociating face recognition subprocesses in elderly adults, their pattern of performance did not mirror either the prosopagnosic group nor the group with autism. Specifically, perception was shown to be impaired, but when impaired perception was accounted for, no differences were observed on face perception or face memory tasks.

This is a novel finding that needs to be replicated. However, if replicated, it implies that changes we observe in face recognition abilities in elderly adults are not likely driven by demands in memory. Instead, it is more likely that face perception is affected, possibly due to changes in low-level or mid-level vision (Roberts & Allen, 2016), which have been reported to affect elderly adults' ability to recognize other stimulus categories. Differences in low- and mid-level vision are likely to influence the ability to match and mismatch faces presented next to each other. Given that in everyday life matching tasks always involve a memory component (e.g., one matches an exemplar of a face in front of them to a stored representation), it is inevitable for one to experience a difficulty. It is likely that these issues in low- and mid-level vision lead to formation of less stable representations of faces (e.g., more fuzziness on the edges of a face space portion that represents a particular identity, see Figure 8.16.), which in turn makes it more difficult to compare an exemplar to the cognitive representation. However, what is crucial is that, if there is no difference in memory after accounting for perception performance, any interventions or training programs that seek to improve face recognition in the elderly ought to focus on face perception abilities.

8.3.2. Future direction and theoretical updates

A lot of discussion about possible training programs has also centered around understanding mechanisms that underlie superior recognition abilities in SRs. Indeed, our findings replicated consistent reports of SRs both reporting fewer issues in everyday face recognition as well as actually scoring very well on objective tests (including the CFMT+, OFMT, and GFMT). Further, SRs were significantly more likely to make similarity judgments about pairs of faces aligned to face recognition algorithms. Most telling,

however, is the analysis examining differences in perception, matching and memory. This work showed that SRs showed superior performance on matching tasks even when perception was accounted for, as well as on memory tasks even when perception and matching were accounted for. The profile of this mirrors that observed in DPs, who were found to have deficits in all subprocesses. This is a novel contribution examining the nature of superiority in SRs that has important implications in a few domains.

Firstly, training paradigms that seek to increase the number of SRs for deployment in national security settings will need to focus on improving performance across all domains – perception, matching and memory – rather than focusing simply on improving matching ability. Further, these findings underscore the possible importance of relying on algorithms in these training paradigms as algorithms can provide a useful 'ground truth' to measure participants' performance against. Finally, these results highlighted the importance of using multiple measures to diagnose SRs rather than a single test as is commonly practiced now. That SRs showed superior ability in all domains of face recognition underlines that the selection of people in this group ought to include assessment on tests measuring more than one domain.

More broadly, these data provided increased granularity into the face recognition process. This detail allowed us to update existing models about face recognition mechanisms in typical and atypical populations alike. Neurotypical results showed the relationship between perception and matching when measured independently, implying that precision of encoding new exemplars into existing representations of faces is an important prerequisite for ability to face match. Importantly, these two abilities were previously empirically inseparable. However, these data showed that there are considerable individual

differences in the relationship between these abilities across participants. Most interestingly, this leaves open a possibility that participants deploy decision-making thresholds of varied appropriateness to determine whether two faces are the same or different. Specifically, there is likely variance in how flexible and accurate decision-making rules are that different participants apply to determining sameness, and it is likely that differences in these can explain some of the remaining differences in face matching performance after accounting for face perception ability.

However, more interesting updates to our understanding of face recognition mechanisms were driven by findings in atypical populations. While different patterns of performance were observed in people with autism, developmental prosopagnosia and healthy elderly adults, differences in organization of face space can explain these different profiles. In these groups, I observed a difference in perception performance compared to typical controls, implying that the ability to encode exemplars in relation to other exemplars of the same person (or of different people) might be impaired. These atypicalities could result in two types of changes to the multidimensional face space in those participants. We might observe misshapen representations (if people have a poor sense of similarity to other faces and therefore erroneously assign exemplars to mistaken identities) or fuzzy, poorly defined edges of person representations (if people's similarity judgments are simply imprecise or uncertain). See Figure 8.16. As a result, these representations could overlap with representations of people close to them in face space. Comparing people who are reasonably similar to each other, therefore, becomes a challenging task producing more errors of 'mistaken' identities, with participants stating that two different images represent the same person.

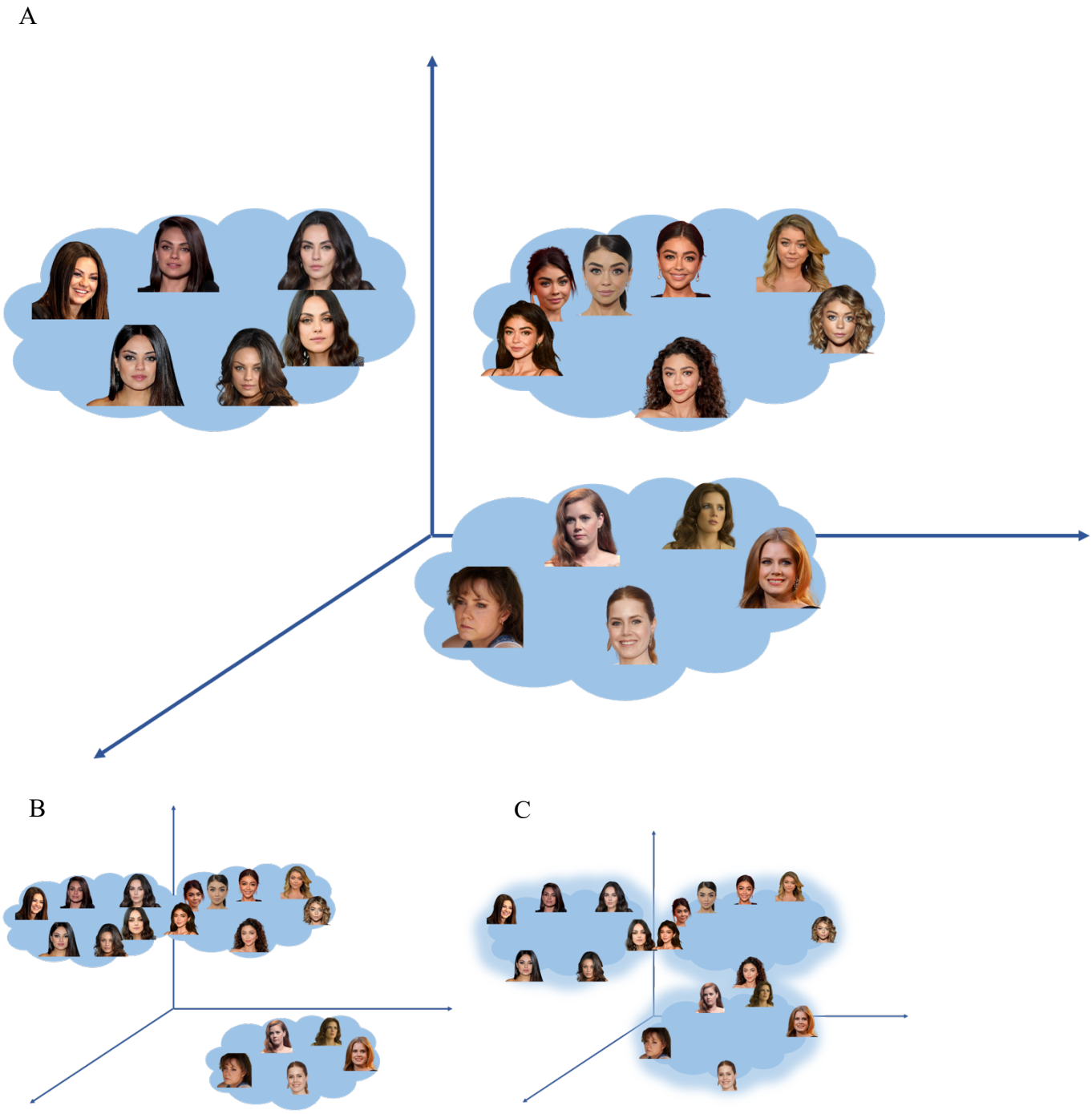


Figure 8.16. Theoretical comparison of 3-D face space representations in typical and atypical participants. In panel A, a simple representation of a three-dimensional face space, as it might be organized in a typical participant. Clear limits of each person's identity are driven by certainty and accuracy of similarity judgments of face pairs. These clear perceptual delineations underscore one's ability to correctly tell faces together or apart. Meanwhile, panels B and C show what an atypical participant's representation might look like, from misshapen representations that lead to overlap (B) to perceptual uncertainty resulting in fuzziness of the cognitive representation (C), both of which could result in more errors of face recognition.

8.4. General limitations

8.4.1. Terminology and diagnostic criteria

A crucial challenge in current visual perception research, particularly in the field of face recognition, is a lack of established, widely accepted theoretical frameworks for some key constructs. As was outlined in more detail in the introduction, diagnostic criteria for super recognizers and developmental prosopagnosics still vary widely between research groups. As a result, traits of participants included in different samples might vary quite substantially, most notably in the severity and consistency of their symptoms. Importantly, groups that deploy less conservative measures of these differences are probably more likely to include more participants and therefore work with larger samples. As a result, they would be more likely to identify subtypes of prosopagnosia or super recognition where subtypes do not in reality exist. It is possible that those groups would cluster participants based on their performance on novel measures, potentially identifying a subgroup with less pronounced performance atypicality or a subgroup with deficits pronounced on one type of measure but not others (e.g., people who might not actually meet the definition for DP or SR). These differences between groups could be explained by selection practices and would disappear with use of more conservative selection criteria. The field would benefit from agreed upon standards for diagnoses, as well as from shared databases of patient test scores.

Further, even where only semantics are involved, key terms such as ‘face perception’ and ‘face matching’, for instance, have been used interchangeably in the literature despite actually referring to two separate stages of the face recognition process. As a result, findings have been contradictory and mixed, with some studies finding differences between groups on perception tests and others discovering no differences (Wiegelt,

Koldewyn, & Kanwisher, 2012 for review; Biotti, Grey, & Cook, 2019). This is mostly a limitation of methodological approaches as the field has mostly relied on tests that only measure face matching performance and cannot capture perception ability independent of matching. However, because of this limitation, inconsistent language has often been used to report study results.

Here I sought to address these challenges to the fullest extent possible. Specifically, I focused on using stringent criteria for selection of developmental prosopagnosics by relying on multiple diagnostic tests that ensured that the included participants did in fact perform poorly on multiple measures. In the case of super recognizers, a standard criterion of superior performance on a single diagnostic test of face memory was applied, though it is possible that not all of those participants would show superior performance on other tests (Noyes, Phillips, & O'Toole, 2017; de Haas, 2022). In fact, some of the included SRs scored in the typical range on the OFMT. It would be interesting for future consortium research to explore whether these participants show typical profiles on other tests as well. This is further evidence that using a single test to determine SR status likely does not constitute a sufficient standard of evidence and that a test battery of specifically designed and highly difficult tests should be used in the future.

A good potential solution to this issue is an emerging practice of sharing participant databases across research groups. Specifically, inclusion of all test scores that a particular participant has achieved in a single database rather than having the scores kept across multiple laboratories will be a substantial improvement on current practices. Not only will such databases provide a rich wealth of information about the patterns of performance and any behavioural subtypes in atypical groups, but they will also likely surface differences

between true poor / superior performers and individuals who may appear as such but only score atypically on one measure, rather than consistently. Further, their use will remove the concern of practice effects on tests applied multiple times by different research groups.

Further to the issue of choosing participants who are included into atypical group comparisons appropriately, the studies in this thesis have relied on the original version of the GFMT (Burton, White, & McNeill, 2010) rather than the updated version published this year (White et al., 2022). Because the updated version of the task was not available when the research outlined here was conducted, I utilized the GFMT as the existing gold standard in face matching research. Therefore, all conclusions made in this dissertation can only apply to the original version. While the updated version makes important improvements on the short versions of the GFMT most commonly deployed by researchers, it remains inadequate at measuring performance across the full spectrum of abilities and still suffers from neurotypicality bias in the method of item choice. Given that the paper outlining the development of an updated test version provides no data about the performance relationships between the original and updated versions of the task, it is difficult to speculate about how the novel version of the task might differ from the old one in our sample. Further, the updated version has not yet been compared to gold-standard tasks, so it is difficult to know whether relationships between tasks have changed with deployment of the GFMT2.

8.4.2. Sample

The clinical sample of participants with autism used in this dissertation included highly functional adults. It is therefore impossible to know how findings of these studies might generalize to those with more severe impairments, who represent the majority of autistic

people. While tasks deployed in these experiments generally do not require verbal ability, completing them requires a good understanding of instructions. Developing versions that do not require any reading would be difficult but developing versions that demonstrate an appropriate way to complete the task might aid participants with more severe cognitive impairments. This would be a useful extension of current work validating the existing findings in a more inclusive sample of autistic participants.

8.4.3. Stimuli

Finally, the OFMT is somewhat limited in the demographics of faces it features as images that it includes are all white. We know of a consistent difference in recognition performance between own-ethnicity and other-ethnicity faces such that one is much better at recognizing faces of their own ethnicity (Tanaka, Kiefer & Bukach, 2004; Rhodes et al., 2009; Anzures et al., 2013). Given that the OFMT features only Caucasian faces, it is reasonable to speculate that I might be observing better performance in Caucasian participants compared to participants of other ethnicities, though insufficiently complete demographic data prevent empirical testing of this hypothesis. However, this issue could be addressed easily in the future. One of the benefits of the OFMT is that alternate and parallel versions of the test can be developed disproportionately faster than before.

For instance, tests studying developmental changes in face recognition have been particularly difficult to obtain. This is partly because children's images are subject to additional security protections. However, it is mostly because the pace of developmental research makes it impractical to choose items and validate a test in large groups of children, as they can be difficult to engage in long paradigms. The OFMT does not suffer from this issue. Due to lack of children's images in the training datasets, there are still

some differences in how well algorithms perform on adult faces compared to children's (Michalski, Yiu, Malec, 2018; Michalski, 2018; Srinivas et al., 2019), but these gaps in performance are being bridged continuously. I expect, therefore, that a children's version of the OFMT could be developed in the future to allow developmental research into changes and maturation of our face recognition ability.

8.5. Future directions

As outlined above, future research will need to focus on defining constructs, agreeing on consistent diagnosis frameworks and more clearly articulating which processes tasks are measuring. Deployment of more sensitive tests will be particularly helpful in studying individual differences in the typical population and better understanding subtle differences in atypical groups. This is likely to result in fewer reports of putative heterogeneity and more focus on meaningful true subtypes of various disorders.

Alongside those advances, research will be furthered in two main directions – firstly, in development of other tests of face recognition subprocesses using algorithmic similarities or relying on them to dynamically create idiosyncratic task versions, and secondly, to better understand the nature of both human and algorithmic processing by comparing them.

Further forays into the first research direction are already taking place. Recent studies have introduced perception tests that use triads assessed for similarity by a face recognition algorithm instead of face pairs (Hahn et al., 2021; Jeckeln et al., 2020). These tests, instead of asking participants to determine whether a presented pair of faces is of the same person

or different people, use the odd-one-out method, asking participants instead to pick an odd identity out of each triad (which is constructed to include two images of the same person and one 'foil'). Like the OFMT, these tests show a wide range of performances by typical participants. Additionally, they can be split into shorter sub tests that can be reliably administered to the same groups of participants when researchers wish to avoid multiple administrations of an identical measure. These tests still need to be validated in atypical populations and have their performance examined for reliability and training effects over time. Due to concurrent development, these have not yet been compared to traditional two-item matching tests (such as the OFMT and the GFMT), so future research should examine the relationship their performances.

An obvious next methodological advance is developing a face memory test using algorithmic similarities. The literature review included in the introduction showed that the most commonly used face memory test is the CFMT (Duchaine & Nakayama, 2006a). It is deployed to dissociate between prosopagnosics and typical performers, and its extended alternative, the CFMT+ (Russell et all., 2009), to dissociate between super recognizers and typical performers. However, the difficulty on the test is achieved through use of non-realistic changes to stimuli, rather than mirroring real life difficulty by increasing the memory load or perceptual similarity of faces. Specifically, the CFMT has three stages of difficulty (the CFMT+ features four) such that increased difficulty in the critical last stage, which provides most of the variation in scores, is achieved through introduction of Gaussian noise overlaid to the images (or rotation of the faces' viewpoints). Obviously, the semi-transparent occlusion and changes in angle do make the task more difficult, though neither of these mirrors face memory difficulties in everyday life. These come more frequently from faces changing as a result of passing time or outward changes in

appearance, or sometimes due to perceptual similarity of two known identities (of different people) and the subsequent likelihood of one 'confusing' the two identities.

A face memory test sensitive to real-life changes would provide a more useful measure of face memory than the CFMT. One can imagine that, in a memory paradigm, the similarity difference between faces presented at learning and at test might influence one's ability to correctly recall a face. Specifically, you would expect that images of faces that are more similar to each other would be more likely to be accurately recognized as the same compared to faces that are more distinct from each other, in line with the logic of the OFMT. It would therefore be possible to manipulate the difficulty of each trial such that images at test are more or less similar to the images presented during encoding. Of course, the similarity of distractors – other images presented as foils in the testing phase – would also need to be considered and varied across trials. Importantly, however, there may even exist a possibility of dynamically changing task difficulty to present more difficult trials to participants who are performing well and easier trials to those who are performing poorly. Using a staircase procedure like this might have important practical considerations as it could allow significantly shorter testing sessions.

Regardless of its length to complete, a well-developed memory test that relies on algorithmic similarity ought to have the capacity to measure performance of super recognizers, developmental prosopagnosics and typical perceivers at the same time, effectively replacing the use CFMT and CFMT+ as diagnostic tools in the field. This would also provide a measure of face memory ability on the same scale for all participants, making a range of novel investigations possible that are now difficult as scores cannot be compared across tests (CFMT+ scores can be recalculated into CFMT scores, but not vice

versa). It might be of interest to determine whether a particular type of difficulty manipulation (e.g., making the learned/test image combination more difficult or making the distractors more similar to the target image) in the test makes participants from any of these three populations more susceptible to making errors.

Use of algorithms also offers an opportunity to move towards idiosyncratic test versioning. Traditionally, producing a test alternative for each participant based on their performance, calibration of difficulty or individual familiarity was logistically impossible or created psychometric issues with test administration. This was because development of parallel versions required that specific items for a particular participant be assessed by neurotypical groups ahead of test administration, making such development prohibitively expensive and challenging. Alternatively, participants would be assessed only using images they reported familiarity with (Duchaine, 2000), which resulted in different numbers of test items being administered to each participant.

However, these challenges can be remedied by using algorithms to support development of individual versions of these tests. Algorithms can yield similarity indices for a virtually unlimited number of face combinations, making it possible for tests to be both individualized to participants and matched in length. As an additional benefit, test versions could be matched in difficulty to ensure scores can be compared across participants. Concerns about psychometric concerns would be alleviated, as all participants complete the same number of trials in their testing session.

Similarly, algorithmic similarities of faces can be applied to distinguish between leading accounts of face learning. Efforts to study face learning by examining individual

differences in this ability (Longmore, Liu, & Young, 2008) have yielded two main hypotheses about how humans learn to associate images of faces with their corresponding identities. One posits that this process is underpinned by exemplar-based encoding (Lewis, 2004; Longmore, Liu, & Young, 2008) and other argues for encoding of an average (Giese & Leopold, 2005; Rhodes & Jeffrey, 2006; Benson & Perret, 1993; Burton et al., 2005; Jenkins & Burton, 2011). These hypotheses disagree on whether humans encode every exemplar of a particular identity they ever encounter or if, instead, they encode an average of exemplars of one identity they have encountered over time. Differentiation between these hypotheses is empirically difficult as they often make the same predictions and subtle differences regarding which exemplar(s) is/are encoded have been challenging to dissociate using traditional methods. However, relying on algorithmic calculations of similarity between face images could resolve this difficulty by providing more precise similarity judgments that allow researchers to distinguish between these alternative explanations.

Finally, researchers have relied on algorithms to compare human and machine performance outright. As machines surpassed humans in accuracy (Taigman et al., 2014), qualitative comparisons became more prominent (Hancock, Somai, & Mileva, 2020; Abudarham, Shkiller, & Yovel, 2019; Parde et al., under review; Cavazos et al., 2021; Colón et al, 2020). This is largely driven by the idea that good deep (convolutional) neural networks can explain not only cognitive patterns in human performance, but that their models correlate very well with brain activity as well (Kubilius et al., 2019; Mehrer et al., 2021; Yamins et al., 2014; Zhuanga et al., 2021). These suggestions have been challenged recently, with some papers finding that the relationships between model layers and neural activity reflect poor brain activity measures rather than a true correlation between the two

(Bowers et al., 2022, under review; Sexton & Love, 2022; Elmoznino & Bonner, under review). These early findings acknowledge that further work will be necessary to better elucidate the relationship between neural activity and neural network performance.

Overall, work presented in this dissertation shows that reliance on human participants for rating and assessment of items to include in a face perception test is a flawed strategy prone to neurotypicality bias and rife with practical challenges. Relying instead on face recognition algorithms provides an opportunity to further research in novel directions by minimizing bias to neurotypicality, as well as removing practical challenges of assessing a very large number of test item options. This approach allows for development of more sensitive, more reliable and more widely applicable tests, and also for independent measurement of subprocesses of face recognition. This provides an exciting avenue to updating existing theoretical frameworks to better highlight differences in atypical populations and explain the wide range of typical performance.

The most important consequence of these changes to theoretical understandings are applied efforts on both ends of the spectrum. Development of training programs that can permanently and meaningfully change face recognition ability has long been a focus of researchers interested in prosopagnosia and autism, particularly given that these efforts have so far have shown limited efficacy. The underlying problem with this work was that the apparent heterogeneity of included samples made it difficult to understand the profile of the deficit in these groups and made it accordingly challenging to devise a training protocol. However, our data indicate that examining performance by dissociating steps of the process can offer useful insight into the stages at which the difference arises. As a

result, future work should be better equipped to address challenges at (an) appropriate point(s) in the process.

On the other hand, applied research looking into improvements of national security, surveillance or footage reviews has increasingly relied on use of technology instead of humans (e.g., face scanning at borders rather than human review), likely as a result of algorithmic accuracy surpassing performance of even the most accurate humans even in most difficult situations (Phillips et al., 2018; Parde et al., 2022). This approach is likely to facilitate speedier border crossings with fewer errors. However, issues might arise with populations that are not proportionally represented in training sets (e.g., people of colour, children). Notably, even the latest generation of DCNNs show some differences in performance on exemplars of ethnicities that are less represented in training sets, mostly towards Asian faces (Gavazos et al., 2020), indicating that deployment in applied situations might give rise to a disproportionate number of issues with some racial groups over others. This issue might be exacerbated by the review that is likely to follow algorithmic malfunctions and is commonly done by a border officer of a majority ethnicity. Own-ethnicity bias makes it more likely an incorrect decision is reached. Some solutions to this issue might again be informed by research, which shows that collaboration of multiple humans of any ethnicity improves face identification accuracy (Bruce et al., 2001), but also that multi-racial pairs and groups can counteract own-ethnicity bias as well (Gavazos, Jeckeln, & O'Toole, 2021). Therefore, devising a system that relies on more than one border patrol officer's judgment, particularly if multi-racial pairs are deployed, would be a wise strategy that reduces the likelihood of subsequent review mistakes. Overall, research seems consistent in that both high performers and algorithms show very

low rates of errors, so finding ways of fusing abilities of both groups promises to improve outcomes of critical face recognition tasks in applied settings (Gavazos et al., 2021).

References

Abudarham, N., Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. *Cognition*, *182*, 73-83.

Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of cognitive neuroscience*, *13*(2), 232-240.

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

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.).

Anastasi, J. S., & Rhodes, M. G. (2006). Evidence for an own-age bias in face recognition. *North American Journal of Psychology*, *8*(2).

Anzures, G., Quinn, P. C., Pascalis, O., Slater, A. M., Tanaka, J. W., & Lee, K. (2013). Developmental origins of the other-race effect. *Current directions in psychological science*, *22*(3), 173-178.

Arizpe, J. M., Saad, E., Douglas, A. O., Germine, L., Wilmer, J. B., & DeGutis, J. M. (2019). Self-reported face recognition is highly valid, but alone is not highly discriminative of prosopagnosia-level performance on objective assessments. *Behavior research methods*, *51*(3), 1102-1116.

Ashwin, C., Chapman, E., Colle, L., & Baron-Cohen, S. (2006). Impaired recognition of negative basic emotions in autism: A test of the amygdala theory. *Social neuroscience*, *1*(3-4), 349-363.

Avidan, G., Tanzer, M., & Behrmann, M. (2011). Impaired holistic processing in congenital prosopagnosia. *Neuropsychologia*, *49*(9), 2541-2552.

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

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, *3*(1), 1-13.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of autism and developmental disorders*, *31*(1), 5-17.

Bartlett, J. C., & Leslie, J. E. (1986). Aging and memory for faces versus single views of faces. *Memory & cognition*, *14*(5), 371-381.

Barton, J. (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., & Corrow, S. L. (2016). Recognizing and identifying people: A neuropsychological review. *Cortex*, *75*, 132-150.

Barton, J. J., & Corrow, S. L. (2016). The problem of being bad at faces. *Neuropsychologia*, *89*, 119-124.

Bastin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology*, *17*(1), 14.

Bate, S., & Tree, J. J. (2017). The definition and diagnosis of developmental prosopagnosia. *Quarterly Journal of Experimental Psychology*, *70*(2), 193-200.

Bate, S., Bennetts, R., Hasshim, N., Portch, E., Murray, E., Burns, E., & Dudfield, G. (2019). The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(3), 363.

Bate, S., Frowd, C., Bennetts, R, Hasshim, N., Murray, E., Bobak, A.K., Wills, H. & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, *3*, 22.

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Portch, E., Murray, E., & Dudfield, G. (2019a). The consistency of superior face recognition skills in police officers. *Applied Cognitive Psychology*, *33*(5), 828-842.

Bate, S., Parris, B., Haslam, C., & Kay, J. (2010). Socio-emotional functioning and face recognition ability in the normal population. *Personality and Individual Differences*, *48*(2), 239-242.

Bate, S., Portch, E., & Mestry, N. (2021). When two fields collide: Identifying "super-recognisers" for neuropsychological and forensic face recognition research. *Quarterly Journal of Experimental Psychology*, *74*(12), 2154-2164.

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: face-blind from birth. *Trends in cognitive sciences*, *9*(4), 180-187. doi:10.1016/j.tics.2005.02.011

Behrmann, M., Avidan, G., Leonard, G. L., Kimchi, R., Luna, B., Humphreys, K., & Minshew, N. (2006). Configural processing in autism and its relationship to face processing. *Neuropsychologia*, *44*(1), 110-129.

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*(7), 1130-1149.

Belanova, E., Davis, J. P., & Thompson, T. (2018). Cognitive and neural markers of super-recognisers' face processing superiority and enhanced cross-age effect. *Cortex*, *108*, 92-111.

Benton, A. L., Sivan, A. B., Hamsher, K. De S., Varney, N. R., & Spreen, O. (1983). *Contribution to Neuropsychological Assessment*. New York: Oxford University Press.

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

Biotti, F., Gray, K. L. H., & Cook, R. (2019). Is developmental prosopagnosia best characterised as an apperceptive or mnemonic condition? *Neuropsychologia*, *124*, 285-298. doi:10.1016/j.neuropsychologia.2018.11.014

Biotti, F., Wu, E., Yang, H., Jiahui, G., Duchaine, B., & Cook, R. (2017). Normal composite face effects in developmental prosopagnosia. *Cortex*, *95*, 63-76.

Bird, G., & Cook, R. (2013). Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism. *Translational psychiatry*, *3*(7), e285-e285.

Bird, G., Press, C., & Richardson, D. C. (2011). The role of alexithymia in reduced eye-fixation in autism spectrum conditions. *Journal of autism and developmental disorders*, *41*(11), 1556-1564.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological science*, *15*(10), 674-679.

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex*, 82, 48-62.

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PloS one*, *11*(2), e0148148.

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

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2019). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly journal of experimental psychology*, *72*(4), 872-881.

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-Movement Strategies in Developmental Prosopagnosia and "Super" Face

Recognition. *Quarterly Journal of Experimental Psychology*, *70*(2), 201-217. doi:10.1080/17470218.2016.1161059

Bodamer, J. (1947). Die Prosop-Agnosie. *Arch. Psychiat.*

Boraston, Z. L., Corden, B., Miles, L. K., Skuse, D. H., & Blakemore, S. J. (2008). Brief report: Perception of genuine and posed smiles by individuals with autism. *Journal of Autism and Developmental Disorders*, *38*(3), 574-580.

Boraston, Z., Blakemore, S. J., Chilvers, R., & Skuse, D. (2007). Impaired sadness recognition is linked to social interaction deficit in autism. *Neuropsychologia*, *45*(7), 1501-1510.

Boucher, J., Lewis, V., 1992. Unfamiliar face recognition in relatively able autistic children. *J. Child Psychol. Psychiatry* 33, 843–859.

Boutet, I., & Meinhardt-Injac, B. (2019). Age differences in face processing: the role of perceptual degradation and holistic processing. *The Journals of Gerontology: Series B*, 74(6), 933-942.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... & Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *psyArxiv*: https://doi.org/10.31234/osf.io/5zf4s

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., Rivolta, D., Willson., C.E. & 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*(5), 423-455.

Bowles, N. L., & Poon, L. W. (1985). Aging and retrieval of words in semantic memory. *Journal of gerontology*, *40*(1), 71-77.

Brewer, R., Bird, G., Gray, K. L., & Cook, R. (2019). Face perception in autism spectrum disorder: Modulation of holistic processing by facial emotion. *Cognition*, *193*, 104016.

Brewer, R., Bird, G., Gray, K. L., & Cook, R. (2019). Face perception in autism spectrum disorder: Modulation of holistic processing by facial emotion. *Cognition*, 193, 104016.

Brinley, J. F. (1965). Cognitive sets, speed and accuracy of performance in the elderly. *Behavior, aging and the nervous system*, *114*, 49.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305-327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207.

Bruck, M., Cavanagh, P., & Ceci, S. J. (1991). Fortysomething: Recognizing faces at one's 25th reunion. *Memory & Cognition*, *19*(3), 221-228.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior research methods*, *42*(1), 286-291.

Cavazos, J. G., Jeckeln, G., & O'Toole, A. (2021). Collaboration to improve cross-race face recognition: Wisdom of the multi-racial crowd?.

Cavazos, J. G., Jeckeln, G., Hu, Y., & O'Toole, A. J. (2021). Strategies of Face Recognition by Humans and Machines. *Deep Learning-Based Face Analytics* (pp. 361-379). Springer, Cham.

Cohan, S., Nakayama, K., & Duchaine, B. (2016). Broadly superior: many, but not all, visual and non-visual abilities are strong in face super-recognizers. *Journal of Vision*, *16*(12), 74-74.

Colón, Y. I., Parde, C., Castillo, C., Cavazos, J., & O'Toole, A. (2020). Facial Expression Information in Humans and DCNNs. *Journal of Vision*, *20*(11), 600-600.

Conway, J. R., Catmur, C., & Bird, G. (2019). Understanding individual differences in theory of mind via representation of minds, not mental states. *Psychonomic bulletin & review*, *26*(3), 798-812.

Cook, R., & Biotti, F. (2016). Developmental prosopagnosia. *Current Biology*, *26*(8), R312-R313.

Cook, R., Brewer, R., Shah, P., & Bird, G. (2013). Alexithymia, not autism, predicts poor recognition of emotional facial expressions. *Psychological science*, *24*(5), 723-732.

Cook, R., Shah, P., Gaule, A., Brewer, R., & Bird, G. (2015). Autism and Developmental Prosopagnosia: A Cross-Disorder Study. *Journal of Vision, 15*(12), 1211-1211. doi:10.1167/15.12.1211

Corrow, S. L., Dalrymple, K. A., & Barton, J. J. (2016). Prosopagnosia: current perspectives. *Eye and brain*, *8*, 165.

Corrow, S. L., Dalrymple, K. A., & Barton, J. J. (2016). Prosopagnosia: current perspectives. *Eye Brain, 8,* 165-175. doi:10.2147/eb.S92838

Crook III, T. H., & Larrabee, G. J. (1992). Changes in facial recognition memory across the adult life span. *Journal of Gerontology*, *47*(3), P138-P141.

Cuve, H. C., Castiello, S., Shiferaw, B., Ichijo, E., Catmur, C., & Bird, G. (2021). Alexithymia explains atypical spatiotemporal dynamics of eye gaze in autism. *Cognition*, *212*, 104710.

Cuve, H. C., Gao, Y., & Fuse, A. (2018). Is it avoidance or hypoarousal? A systematic review of emotion recognition, eye-tracking, and psychophysiological studies in young adults with autism spectrum conditions. *Research in Autism Spectrum Disorders*, *55*, 1-13.

Dalmaijer, E. S., Nord, C. L., Astle, D. E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics, 23* (1) https://doi.org/10.1186/s12859-022-04675-1

Dalrymple, K. A., & Palermo, R. (2016). Guidelines for studying developmental prosopagnosia in adults and children. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(1), 73-87.

Dalrymple, K. A., Corrow, S., Yonas, A., & Duchaine, B. (2012). Developmental prosopagnosia in childhood. *Cogn Neuropsychol, 29*(5-6), 393-418. doi:10.1080/02643294.2012.722547

Dalrymple, K. A., Fletcher, K., Corrow, S., das Nair, R., Barton, J. J., Yonas, A., & Duchaine, B. (2014). “A room full of strangers every day”: The psychosocial

impact of developmental prosopagnosia on children and their families. *Journal of psychosomatic research*, *77*(2), 144-150.

Dalrymple, K. A., Garrido, L., & Duchaine, B. (2014). Dissociation between face perception and face memory in adults, but not children, with developmental prosopagnosia. *Developmenta Cognitive Neuroscience, 10*, 10-20. doi:10.1016/j.dcn.2014.07.003

Davies, S., Bishop, D., Manstead, A.S., Tantam, D., 1994. Face perception in children with autism and Asperger's syndrome. *Journal of Children Psychology and Psychiatry* 35, 1033–1057.

Dawson, G., Webb, S.J., McPartland, J., 2005. Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies. *Developmental Neuropsychology. 27, 403–424.*

De Gelder, B., & Rouw, R. (2000). Configural face processes in acquired and developmental prosopagnosia: Evidence for two separate face systems?. *Neuroreport*, 3145-3150.

de Haas, B. (2022). What's a super-recogniser?. *Neuropsychologia*, *166*, 107805.

DeGutis, J., Cohan, S., & Nakayama, K. (2014). Holistic face training enhances face processing in developmental prosopagnosia. *Brain*, *137*(6), 1781-1798.

DeGutis, J., Cohan, S., Mercado, R. J., Wilmer, J., & Nakayama, K. (2012). Holistic processing of the mouth but not the eyes in developmental prosopagnosia. *Cognitive Neuropsychology*, *29*(5-6), 419-446.

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, dissociations from face memory, and expertise effects. *Behavior research methods*, *44*(2), 587-605.

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British journal of psychology*, *106*(3), 433-445.

Duchaine, B. C. (2000). Developmental prosopagnosia with normal configural processing. *Neuroreport*, *11*(1), 79-83.

Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition test. *Neurology*, *62*(7), 1219-1220.

Duchaine, B. C., & Nakayama, K. (2006b). Developmental prosopagnosia: a window to content-specific face processing. *Current opinion in neurobiology*, *16*(2), 166-173.

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*(6), 713-720.

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of cognitive neuroscience*, *17*(2), 249-261.

Duchaine, B., & Nakayama, K. (2006a). The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*(4), 576-585. doi:10.1016/j.neuropsychologia.2005.07.001

Duchaine, B., & Yovel, G. (2015). A revised neural framework for face processing. *Annual review of vision science*, *1*, 393-416.

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology, 24*(4), 419-430. doi:10.1080/02643290701380491

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*(2), 104-113.

Dunn, J. D., Summersby, S., Towler, A., Davis, J. P., & White, D. (2020). UNSW Face Test: A screening tool for super-recognizers. *PloS one*, *15*(11), e0241747.

Ellis, H. D., & Deregowski, J. B. (1981). Within-race and between-race recognition of transformed and untransformed faces. *The American Journal of Psychology*, 27-35.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception*, *8*(4), 431-439.

Elmoznino, E., & Bonner, M. F. (2022). High-performing neural network models of visual cortex benefit from high latent dimensionality. *bioRxiv*.

Faja, S., Webb, S. J., Merkle, K., Aylward, E., & Dawson, G. (2009). Brief report: Face configuration accuracy and processing speed among adults with high-functioning autism spectrum disorders. *Journal of autism and developmental disorders*, *39*(3), 532-538.

Faubert, J. (2002). Visual perception and aging. *Canadian Journal of Experimental Psychology*, 56(3), 164.

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in MTurk and other online samples. *Industrial and Organizational Psychology*, *8*(2), 196-202.

Fry, R., Wilmer, J., Xie, I., Verfaellie, M., & DeGutis, J. (2020). Evidence for normal novel object recognition abilities in developmental prosopagnosia. *R Soc Open Sci, 7*(9), 200988. doi:10.1098/rsos.200988

Funk, F., & Todorov, A. (2013). Criminal stereotypes in the courtroom: Facial tattoos affect guilt and punishment differently. *Psychology, Public Policy, and Law*, *19*(4), 466.

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

Fysh, M. C., & Ramon, M. (2022). Accurate but inefficient: Standard face identity matching tests fail to identify prosopagnosia. *Neuropsychologia, 165,* 108119. doi:10.1016/j.neuropsychologia.2021.108119

Fysh, M. C., Stacchi, L., & Ramon, M. (2020). Differences between and within individuals, and subprocesses of face cognition: Implications for theory, research and personnel selection. *Royal Society Open Science*, *7*(9), 200233.

Gaigg, S. B., Cornell, A. S., & Bird, G. (2018). The psychophysiological mechanisms of alexithymia in autism spectrum disorder. *Autism*, *22*(2), 227-231.

Garrido, L., Duchaine, B., & Nakayama, K. (2008). Face detection in normal and prosopagnosic individuals. *Journal of Neuropsychology*, *2*(1), 119-140.

Gauthier, I., Klaiman, C., & Schultz, R. T. (2009). Face composite effects reveal abnormal face processing in Autism spectrum disorders. *Vision Research*, 49(4), 470-478.

Gepner, B., Gelder, B. D., & Schonen, S. D. (1996). Face processing in autistics: Evidence for a generalised deficit?. *Child Neuropsychology*, 2(2), 123-139.

Geskin, J., & Behrmann, M. (2018). Congenital prosopagnosia without object agnosia? A literature review. *Cognitive Neuropsychology*, *35*(1-2), 4-54.

Giese, M. A., & Leopold, D. A. (2005). Physiologically inspired neural model for the encoding of face spaces. *Neurocomputing*, *65*, 93-101.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*(1), 32-41.

Golarai, G., Grill-Spector, K., Reiss, A.L., 2006. Autism and the development of face processing. *Clin. Neurosci. Res.* 6, 145–160.

Grady, C. L., Maisog, J. M., Horwitz, B., Ungerleider, L. G., Mentis, M. J., Salerno, J. A., ... & Haxby, J. V. (1994). Age-related changes in cortical blood flow activation during visual processing of faces and location. *Journal of Neuroscience*, *14*(3), 1450-1462.

Gray, K. L. H., Bird, G., & Cook, R. (2017). Robust associations between the 20-item prosopagnosia index and the Cambridge Face Memory Test in the general population. *Royal Society Open Science, 4*(3), 160923. doi:10.1098/rsos.160923

Gray, K. L., Biotti, F., & Cook, R. (2019). Evaluating object recognition ability in developmental prosopagnosia using the Cambridge Car Memory Test. *Cognitive Neuropsychology*, *36*(1-2), 89-96.

Gray, K. L., Bird, G., & Cook, R. (2017). Robust associations between the 20-item prosopagnosia index and the Cambridge Face Memory Test in the general population. *Royal Society Open Science*, *4*(3), 160923.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: John Wiley.

Griffin, J. W., Bauer, R., & Scherf, K. S. (2021). A quantitative meta-analysis of face recognition deficits in autism: 40 years of research. *Psychological Bulletin, 147*(3), 268-292. http://dx.doi.org/10.1037/bul0000310

Grueter, M., Grueter, T., Bell, V., Horst, J., Laskowski, W., Sperling, K., ... & Kennerknecht, I. (2007). Hereditary prosopagnosia: the first case series. *Cortex*, *43*(6), 734-749.

Habak, C., Wilkinson, F., & Wilson, H. R. (2008). Aging disrupts the neural transformations that link facial identity across views. *Vision research*, *48*(1), 9-15.

Hahn, C., Yates, A., Phillips, P. J., O'Toole, A., Jeckeln, G., Cavazos, J., & Hu, Y. (2021). Face Identification Proficiency Test Designed Using Item Response Theory.

Hancock, P. J. (2021). Familiar faces as islands of expertise. Cognition, 214, 104765.

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

Hancock, P. J., Somai, R. S., & Mileva, V. R. (2020). Convolutional neural net face recognition works in non-human-like ways. *Royal Society Open Science*, *7*(10), 200595.

Happé, F. (1999). Autism: cognitive deficit or cognitive style?. *Trends in cognitive sciences*, *3*(6), 216-222.

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of autism and developmental disorders*, *36*(1), 5-25.

Happé, F., Cook, J. L., & Bird, G. (2017). The structure of social cognition: In (ter) dependence of sociocognitive processes. *Annual review of psychology*, *68*, 243-267.

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

Hecaen, H., & Angelergues, R. (1962). Agnosia for Faces (Prosopagnosia). *Archives of Neurology, 7*, 24-32.

Hedley, D., Brewer, N. i Young, R. (2011). Face recognition performance of individuals with Asperger syndrome on the Cambridge Face Memory Test. *Autism Research*, *4*(6), 449- 455.

Hobson, R. P. (1986). The autistic child's appraisal of expressions of emotion. *Journal of Child psychology and Psychiatry*, *27*(3), 321-342.

Humphreys, K., Minshew, N., Leonard, G. L., & Behrmann, M. (2007). A fine-grained analysis of facial expression processing in high-functioning adults with autism. *Neuropsychologia*, *45*(4), 685-695.

Irhebhude, M., Celestine, A. U., & Aget, B. J. (2019). Exploring the Efficacy of Face Recognition Algorithm on Different Skin Colors. *IUP Journal of Information Technology*, *15*(1).

Jackson, M. C., Counter, P., & Tree, J. J. (2017). Face working memory deficits in developmental prosopagnosia: Tests of encoding limits and updating processes. *Neuropsychologia*, *106*, 60-70.

Jackson, M. C., Counter, P., & Tree, J. J. (2017). Face working memory deficits in developmental prosopagnosia: Tests of encoding limits and updating processes. *Neuropsychologia, 106*, 60-70. doi:10.1016/j.neuropsychologia.2017.09.003

Jeckeln, G., Hu, Y., Cavazos, J. G., Yates, A. N., Hahn, C. A., Tang, L., ... & Phillips, P. J. (2020). Building calibrated face-identification ability test with Item Response Theory. *Journal of Vision*, *20*(11), 579-579.

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

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

Joseph, R. M., & Tanaka, J. (2003). Holistic and part-based face recognition in children with autism. *Journal of child psychology and psychiatry*, 44(4), 529-542

Kennerknecht, I., Grueter, T., Welling, B., Wentzek, S., Horst, J., Edwards, S., & Grueter, M. (2006). First report of prevalence of non-syndromic hereditary prosopagnosia (HPA). *American Journal of Medical Genetics Part A*, *140*(15), 1617-1622.

Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2016). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism, 20*(4), 442-462. doi:10.1177/1362361315588200

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789-1801.

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2013). Effects of aging on face identification and holistic face processing. *Vision research*, 88, 38-46.

Kress, T., & Daum, I. (2003). Developmental prosopagnosia: A review. *Behavioural neurology*, *14*(3-4), 109-121.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... & DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in neural information processing systems*, *32*.

Kuefner, D., Macchi Cassia, V., Picozzi, M., & Bricolo, E. (2008). Do all kids look alike? Evidence for an other-age effect in adults. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(4), 811.

Laguna, K., & Babcock, R. L. (1997). Computer anxiety in young and older adults: Implications for human-computer interactions in older populations. *Computers in human behavior*, *13*(3), 317-326.

Lamont, A. C., Stewart-Williams, S., & Podd, J. (2005). Face recognition and aging: Effects of target age and memory load. Memory & cognition, 33(6), 1017-1024.

Lewis, M. (2004). Face-space-R: Towards a unified account of face recognition. *Visual Cognition*, *11*(1), 29-69.

Light, L. L. (1991). Memory and aging: Four hypotheses in search of data. *Annual review of psychology*, *42*(1), 333-376.

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

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

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. Journal of Experimental Psychology: Human Perception and Performance, 34(1), 77.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... & Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of autism and developmental disorders, 30(3), 205-223.

Lott, L. A., Haegerstrom-Portnoy, G., Schneck, M. E., & Brabyn, J. A. (2005). Face recognition in the elderly. *Optometry and Vision Science*, *82*(10), 874-881.

Masi, I., Wu, Y., Hassner, T., & Natarajan, P. (2018, October). Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI*) (pp. 471-478). IEEE.

Maylor, E. A. (1990). Recognizing and naming faces: Aging, memory retrieval, and the tip of the tongue state. *Journal of Gerontology*, 45(6), P215-P226.

Maylor, E. A., & Valentine, T. (1992). Linear and nonlinear effects of aging on categorizing and naming faces. *Psychology and Aging*, 7(2), 317.

Mazura, J. C., Juluru, K., Chen, J. J., Morgan, T. A., John, M., & Siegel, E. L. (2012). Facial recognition software success rates for the identification of 3D surface reconstructed facial images: implications for patient privacy and security. *Journal of digital imaging*, *25*(3), 347-351.

McConachie, H. R. (1976). Developmental prosopagnosia. A single case report. *Cortex*, *12*(1), 76-82.

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision research*, *69*, 10-22.

McKelvie, S. J., Standing, L., Jean, D. S., & Law, J. (1993). Gender differences in recognition memory for faces and cars: Evidence for the interest hypothesis. *Bulletin of the Psychonomic Society*, *31*(5), 447-448.

McKone, E., Hall, A., Pidcock, M., Palermo, R., Wilkinson, R. B., Rivolta, D., Yovel, G., Davis, J.M. & O'Connor, K. B. (2011). Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test–Australian. *Cognitive neuropsychology*, *28*(2), 109-146.

McLaughlin, M., & Castro, D. (2020). The Critics Were Wrong: NIST Data Shows the Best Facial Recognition Algorithms Are Neither Racist Nor Sexist. *Information Technology and Innovation Foundation.*

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

Megreya, A. M., & Bindemann, M. (2015). Developmental improvement and age-related decline in unfamiliar face matching. *Perception*, *44*(1), 5-22.

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

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

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118.

Meinhardt-Injac, B., Boutet, I., Persike, M., Meinhardt, G., & Imhof, M. (2017). From development to aging: holistic face perception in children, younger and older adults. *Cognition*, 158, 134-146.

Meinhardt-Injac, B., Persike, M., & Meinhardt, G. (2014). Holistic processing and reliance on global viewing strategies in older adults' face perception. *Acta psychologica*, 151, 155-163.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3.

Michalski, D., Yiu, S. Y., & Malec, C. (2018, February). The impact of age and threshold variation on facial recognition algorithm performance using images of children. In *2018 international conference on biometrics* (pp. 217-224). IEEE.

Michalski, D. J. (2018). *The impact of age-related variables on facial comparisons with images of children: algorithm and practitioner performance* (Doctoral dissertation, Adelaide University). https://hdl.handle.net/2440/111184

Mondloch, C. J., Le Grand, R., & Maurer, D. (2002). Configural face processing develops more slowly than featural face processing. *Perception*, *31*(5), 553-566.

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

Murray, E. & Bate, S. (2020) Diagnosing Developmental Prosopagnosia: Repeated Assessment using the Cambridge Face Memory Test. *Royal Society Open Science*, 7, 200884.

Murray, E., Hills, P. J., Bennetts, R. J., & Bate, S. (2018). Identifying hallmark symptoms of developmental prosopagnosia for non-experts. *Scientific Reports*, *8*(1), 1-12.

Nador, J. D., Zoia, M., Pachai, M. V., & Ramon, M. (2021). Psychophysical profiles in super-recognizers. *Scientific reports*, *11*(1), 1-11.

Naveh-Benjamin, M., Craik, F. I., Perretta, J. G., & Tonev, S. T. (2000). The effects of divided attention on encoding and retrieval processes: The resiliency of retrieval processes. *The Quarterly Journal of Experimental Psychology Section A*, *53*(3), 609-625.

Nishimura, M., Doyle, J., Humphreys, K., & Behrmann, M. (2010). Probing the face-space of individuals with prosopagnosia. *Neuropsychologia*, *48*(6), 1828-1841.

Nishimura, M., Rutherford, M. D., & Maurer, D. (2008). Converging evidence of configural processing of faces in high-functioning adults with autism spectrum disorders. *Visual Cognition*, 16(7), 859-891.

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

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recogniser?. In *Face processing: Systems, disorders and cultural differences* (pp. 173-201). Nova Science Publishers Inc.

O'Toole, A. J., & Castillo, C. D. (2021). Face recognition by humans and machines: Three fundamental advances from deep learning. *Annual Review of Vision Science*, *7*, 543-570.

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of nonverbal behavior*, *34*(2), 83-110.

Owsley, C. (2016). Vision and aging. *Annual review of vision science*, 2, 255-271.

Palermo, R., Rivolta, D., Wilson, C. E., & Jeffery, L. (2011). Adaptive face space coding in congenital prosopagnosia: Typical figural aftereffects but abnormal identity aftereffects. *Neuropsychologia*, *49*(14), 3801-3812.

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., ... & McKone, E. (2017). Do people have insight into their face recognition abilities?. *Quarterly journal of experimental psychology*, *70*(2), 218-233.

Palermo, R., Willis, M. L., Rivolta, D., McKone, E., Wilson, C. E., & Calder, A. J. (2011). Impaired holistic coding of facial expression and facial identity in congenital prosopagnosia. *Neuropsychologia*, *49*(5), 1226-1235.

Parde, C. J., Strehle, V. E., Banerjee, V., Hu, Y., Cavazos, J. G., Castillo, C. D., & O'Toole, A. J. (2022). Twin identification over viewpoint change: A deep convolutional neural network surpasses humans. *arXiv preprint arXiv:2207.05316*.

Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of autism and developmental disorders*, *32*(4), 249-261.

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32*(1), 74-85. doi:https://doi.org/10.1016/j.imavis.2013.12.002

Phillips, P. J., Moon, H., Rauss, P., & Rizvi, S. A. (1997). The FERET evaluation methodology for face-recognition algorithms. IEEE, 137-143.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5), 295-306.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, JC., Castillo, C. D., Chellappa, R., White, D., O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences, 115*(24), 6171-6176. doi:doi:10.1073/pnas.1721355115

processes. *Neuropsychologia*, *106*, 60-70.

Pruette, J. R. (2013). Autism diagnostic observation schedule-2 (ADOS-2). *Google Scholar*, 1-3.

Pyo, J., & Maxfield, M. G. (2021). Cognitive effects of inattentive responding in an MTurk sample. *Social Science Quarterly*, *102*(4), 2020-2039.

Ramon, M. (2021). Super-Recognizers–a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, *158*, 107809.

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British journal of psychology*, 110(3), 461-479.

Ramon, M., Bobak, A. K., & White, D. (2019). Towards a 'manifesto' for super-recognizer research. *British Journal of Psychology*, *110*(3), 495-498.

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.

Rezlescu, C., Susilo, T., Wilmer, J. B., & Caramazza, A. (2017). The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1961.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision research*, *46*(18), 2977-2987.

Rhodes, G., Locke, V., Ewing, L., & Evangelista, E. (2009). Race coding and the other-race effect in face recognition. *Perception*, *38*(2), 232-241.

Ritchie, K. L., Flack, T. R., Fuller, E. A., Cartledge, C., & Kramer, R. S. (2022). The pairs training effect in unfamiliar face matching. *Perception*, 03010066221096987.

Robel, L., Ennouri, K., Piana, H., Vaivre-Douret, L., Perier, A., Flament, M. F., & Mouren-Siméoni, M. C. (2004). Discrimination of face identities and expressions in children with autism: Same or different? *European Child & Adolescent Psychiatry*, *13*(4), 227–233. https://doi.org/10.1007/S00787-004-0409-8

Roberts, K. L., & Allen, H. A. (2016). Perception and cognition in the ageing brain: a brief review of the short-and long-term links between perceptual and cognitive decline. *Frontiers in aging neuroscience*, 8, 39.

Robertson, D. J., Black, J., Chamberlain, B., Megreya, A. M., & Davis, J. P. (2020). Super-recognisers show an advantage for other race face identification. *Applied Cognitive Psychology*, *34*(1), 205-216.

Rossion, B., Caldara, R., Seghier, M., Schuller, A., 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*, 2381- 2385.

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological science*, *19*(2), 109-111.

Rule, N. O., & Ambady, N. (2009). She's got the look: Inferences from female chief executive officers' faces predict their success. *Sex Roles*, *61*(9), 644-652.

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

Rutherford, M. D., Clements, K. A., & Sekuler, A. B. (2007). Differences in discrimination of eye and mouth displacement in autism spectrum disorders. *Vision research*, *47*(15), 2099-2110.

Rutter, M., Le Couteur, A., & Lord, C. (2003). Autism diagnostic interview-revised. *Los Angeles, CA: Western Psychological Services*, *29*(2003), 30.

Scherf, S. K., Luna, B., Kimchi, R., Minshew, N., & Behrmann, M. (2008). Missing the big picture: Impaired development of global shape processing in autism. *Autism research*, *1*(2), 114-129.

Schiltz, C., Sorger, B., Caldara, R., Ahmed, F., Mayer, E., Goebel, R., & Rossion, B. (2006). Impaired face discrimination in acquired prosopagnosia is associated with abnormal response to individual faces in the right middle fusiform gyrus. *Cerebral Cortex, 16*, 574-586.

Schultz, R.T., 2005. Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. *Int. J. Dev. Neurosci*. 23, 125–141.

Schwarzer, G., Kretzer, M., Wimmer, D., & Jovanovic, B. (2010). Holistic face processing among school children, younger and older adults. *European Journal of Developmental Psychology*, 7(4), 511-528.

Schweich, M., Van der Linden, M., Brédart, S., Bruyer, R., Nelles, B., & Schils, J. P. (1992). Daily-life difficulties in person recognition reported by young and elderly subjects. *Applied Cognitive Psychology*, *6*(2), 161-172.

Selwyn, N. (2004). The information aged: A qualitative study of older adults' use of information and communications technology. *Journal of Aging studies*, *18*(4), 369-384.

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science advances*, *8*(28), eabm2219.

Shafto, M. A., Burke, D. M., Stamatakis, E. A., Tam, P. P., & Tyler, L. K. (2007). On the tip-of-the-tongue: neural correlates of increased word-finding failures in normal aging. *Journal of cognitive neuroscience*, *19*(12), 2060-2070.

Shah, P., Bird, G., & Cook, R. (2016). Face processing in autism: Reduced integration of cross-feature dynamics. *Cortex*, *75*, 113-119.

Shah, P., Catmur, C., & Bird, G. (2016). Emotional decision-making in autism spectrum disorder: the roles of interoception and alexithymia. *Molecular autism*, *7*(1), 1-10.

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): a self-report instrument for identifying developmental prosopagnosia. *Royal Society open science*, *2*(6), 140343.

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin*, *100*(2), 139.

Sinha, P., Balas, B., Ostrovsky, Y., & Russell, R. (2006). Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11), 1948-1962.

Smith, H. M., Andrews, S., Baguley, T. S., Colloff, M. F., Davis, J. P., White, D., ... & Flowe, H. D. (2021). Performance of typical and superior face recognizers on a novel interactive face matching procedure. *British Journal of Psychology*, *112*(4), 964-991.

Sorger, B., Goebel, R., Schiltz, C., & Rossion, B. (2007). Understanding the functional neuroanatomy of acquired prosopagnosia. *Neuroimage, 35*, 836- 852.

Srinivas, N., Ricanek, K., Michalski, D., Bolme, D. S., & King, M. (2019). Face recognition algorithm bias: Performance differences on images of children and adults. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*.

Stacchi, L., Huguenin-Elie, E., Caldara, R., & Ramon, M. (2020). Normative data for two challenging tests of face matching under ecological conditions. *Cognitive research: principles and implications*, *5*(1), 1-17.

Stacchi, L., Huguenin-Elie, E., Caldara, R., & Ramon, M. (2020). Normative data for two challenging tests of face matching under ecological conditions. *Cognitive research: principles and implications*, *5*(1), 1-17.

Stantić, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., Catmur, C., Bird, G. (2022a). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavioral Research Methods, 54*(1), 158-173.

Stantić, M., Browne, K.., Ichijo, E., Catmur, C., & Bird, G.P. (under review) Independent Measurement of Face Perception, Face Matching, and Face Memory Reveals Impairments in Face Perception and Memory, but Not Matching, in Autism.

Stantić, M., Catmur, C., Duchaine, B., & Bird, G. (2020). Human face matching performance is predicted by deviation from algorithmic similarity. *Journal of Vision*, *20*(11), 508-508.

Stantić, M., Hearne, B., Catmur, C. & Bird, G. (2021a). Use of the Oxford Face Matching Test Reveals an Effect of Ageing On Face Perception But Not Face Memory. *Cortex, 145*, 226-235.

Stantić, M., Ichijo, E., Catmur, C., & Bird, G. (2021b) Short Report: Face Memory and Face Perception in Autism. *Autism*.

Stantić, M., Pounder, Z., Bate, S., Catmur, C., & Bird, G.P. (under review) Individuals who are 'super-recognisers' show superior performance on independent measures of face perception, face memory and face matching.

Stantić, M., Pounder, Z., Bate, S., Susilo, T., Catmur, C., & Bird, G. (2022b). Individuals with developmental prosopagnosia show independent impairments in face perception, face memory and face matching. *Cortex*, *157*, 266-273.

Stollhoff, R., Jost, J., Elze, T., & Kennerknecht, I. (2011). Deficits in long-term recognition memory reveal dissociated subtypes in congenital prosopagnosia. *PLoS One*, *6*(1), e15702.

Stollhoff, R., Jost, J., Elze, T., & Kennerknecht, I. (2011). Deficits in long-term recognition memory reveal dissociated subtypes in congenital prosopagnosia. *PLoS ONE, 6*(1), e15702. doi:10.1371/journal.pone.0015702

Susilo, T., & Duchaine, B. (2013). Advances in developmental prosopagnosia research. *Current Opininions Neurobiology, 23*(3), 423-429. doi:10.1016/j.conb.2012.12.011

Susilo, T., McKone, E., Dennett, H., Darke, H., Palermo, R., Hall, A., Pidcock, M., Dawel, A., Jeffrey, L., Wilson, C. E., Rhodes, G. (2010). Face recognition impairments despite normal holistic processing and face space coding: Evidence from a case of developmental prosopagnosia. *Cognitive Neuropsychology, 27*(8), 636-664. doi:10.1080/02643294.2011.613372

Sussman, A. B., Petkova, K., & Todorov, A. (2013). Competence ratings in US predict presidential election outcomes in Bulgaria. *Journal of Experimental Social Psychology*, *49*(4), 771-775.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).

Tanaka, J. W., Kiefer, M., & Bukach, C. M. (2004). A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study. *Cognition*, *93*(1), B1-B9.

Tang, J., Falkmer, M., Horlin, C., Tan, T., Vaz, S., & Falkmer, T. (2015). Face recognition and visual search strategies in autism spectrum disorders: Amending and extending a recent review by Weigelt et al. *PloS One*, *10*(8), e0134439.

Teunisse, J. P., & De Gelder, B. (2003). Face processing in adolescents with autistic disorder: The inversion and composite effects. Brain and cognition, 52(3), 285-294.

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

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work?. *PloS one*, *14*(2), e0211037.

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

Towler, J., Fisher, K., & Eimer, M. (2017). The cognitive and neural basis of developmental prosopagnosia. *Quarterly Journal of Experimental Psychology*, *70*(2), 316-344.

Towler, J., Fisher, K., & Eimer, M. (2017). The cognitive and neural basis of developmental prosopagnosia. *Quarterly Journal of Experimental Psychology*, 70(2), 316-344.

Towler, J., Fisher, K., & Eimer, M. (2018). Holistic face perception is impaired in developmental prosopagnosia. *Cortex, 108*, 112-126. doi:10.1016/j.cortex.2018.07.019

Turk, M. A., & Pentland, A. P. (1991, January). Face recognition using eigenfaces. In *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition* (pp. 586-587). IEEE Computer Society.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1), 71-86.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *43*(2), 161-204.

Valentine, T., & Bruce, V. (1986). The effect of race, inversion and encoding activity upon face recognition. *Acta psychologica*, 61(3), 259-273.

Valentine, T., & Bruce, V. (1986). The effect of race, inversion and encoding activity upon face recognition. *Acta psychologica*, *61*(3), 259-273.

Van Deursen, A. J., & Helsper, E. J. (2015). A nuanced understanding of Internet use and non-use among the elderly. *European journal of communication*, 30(2), 171-187.

Van Exel, E., Gussekloo, J., De Craen, A. J. M., Bootsma-Van Der Wiel, A., Houx, P., Knook, D. L., & Westendorp, R. G. J. (2001). Cognitive function in the oldest old: women perform better than men. *Journal of Neurology, Neurosurgery & Psychiatry*, *71*(1), 29-32.

Ventura, P., Carmo, J. C., Souza, C., Martins, F., Leite, I., Pinho, S., Barahona-Correa, B., & Filipe, C. N. (2018). Holistic processing of faces is intact in adults with autism spectrum disorder. *Visual Cognition*, 26(1), 13-24.

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision research*, *141*, 217-227.

Verosky, S. C., Turk-Browne, N. B., & Todorov, A. (2013). Social group knowledge biases face perception. *Journal of Vision*, *13*(9), 865-865.

Wallace, S., Coleman, M. & Bailey, A. (2008). Face and object processing in autism spectrum disorders. *Autism Research*, *1*(1), 43-51.

Warrington, E. K. (1984). *Recognition Memory Test*. Windsor, England: NFER-Nelson.

Webster, M. A., & MacLeod, D. I. (2011). Visual adaptation and face perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1702-1725.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-*

Weigelt, S., Koldewyn, K., & Kanwisher, N. (2012). Face identity recognition in autism spectrum disorders: A review of behavioral studies. *Neuroscience & Biobehavioral Reviews*, *36*(3), 1060-1084.

White, D., Guilbert, D., Varela, V. P. L., Jenkins, R., & Burton, A. M. (2022). GFMT2: A psychometric measure of face matching ability. *Behavioral Research Methods, 54*(1), 252-260. doi:10.3758/s13428-021-01638-x

White, D., Guilbert, D., Varela, V. P., Jenkins, R., & Burton, A. M. (2022). GFMT2: A psychometric measure of face matching ability. *Behavior research methods*, *54*(1), 252-260.

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic bulletin & review*, *21*(1), 100-106.

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017). Face matching impairment in developmental prosopagnosia. *Quarterly Journal of Experimental Psychology, 70*(2), 287-297. doi:10.1080/17470218.2016.1173076

Wiese, H., Kachel, U., & Schweinberger, S. R. (2013). Holistic face processing of own- and other-age faces in young and older adults: ERP evidence from the composite face task. *Neuroimage*, 74, 306-317.

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225-230.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K. & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. Proceedings of the National Academy of sciences, 107(11), 5238-5241.

Wilson, C. E., Palermo, R., Schmalzl, L., & Brock, J. (2010). Specificity of impaired facial identity recognition in children with suspected developmental prosopagnosia. *Cognitive Neuropsychology, 27*(1), 30-45. doi:10.1080/02643294.2010.490207

Woodbury-Smith, M. R., Robinson, J., Wheelwright, S., & Baron-Cohen, S. (2005). Screening adults for Asperger syndrome using the AQ: A preliminary study of its diagnostic validity in clinical practice. *Journal of autism and developmental disorders*, *35*(3), 331-335.

Woodhead, M. M., Baddeley, A. D., & Simmonds, D. C. V. (1979). On training people to recognize faces. *Ergonomics*, *22*(3), 333-343.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619-8624.

Yardley, L., McDermott, L., Pisarski, S., Duchaine, B., & Nakayama, K. (2008). Psychosocial consequences of developmental prosopagnosia: A problem of recognition. *Journal of psychosomatic research*, *65*(5), 445-451.

Young, A. W., & Burton, A. M. (2018). Are we face experts? Trends in cognitive sciences, 22(2), 100-110.

Young, A. W., & Noyes, E. (2019). We need to talk about super-recognizers Invited commentary on: Ramon, M., Bobak, AK, & White, D. Super-recognizers: From the lab to the world and back again. British Journal of Psychology. *British Journal of Psychology*, *110*(3), 492-494.

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., Hattersley, A., Heuerman-Williamson, B., Jones, K., MacKisack, M., Winlove, C. (2020). Phantasia–The psychological significance of lifelong visual imagery vividness extremes. *Cortex, 130,* 426-440. doi:https://doi.org/10.1016/j.cortex.2020.04.003

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3).

Zorowitz, S., Niv, Y., & Bennett, D. (2021). Inattentive responding can induce spurious associations between task behavior and symptom measures. *psyArxiv:* https://doi.org/10.31234/osf.io/rynhk