ANDROGYNOUS AND GENDERED FACE PERCEPTION

Androgynous and Gendered Face Perception

By Leigh Greenberg, B.Sc.

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University
Master of Science (2022)
Hamilton, Ontario (Psychology, Neuroscience & Behaviour)

TITLE: Androgynous and Gendered Face Perception
AUTHOR: Leigh Greenberg (McMaster University)
SUPERVISOR: Dr. Allison Sekuler & Dr. Patrick Bennett
NUMBER OF PAGES: ix, 43

Abstract

Studies in face perception often use androgynous faces as a tool. The common assumption about androgynous faces is that they lie at the halfway point of a continuum that features "male" and "female" at the extremes. However, this definition has not been verified by research. This thesis uses a variety of methods, with an emphasis on data-driven methods, to test common assumptions about androgynous faces. Chapter 2 compared morphed faces, which were created using the common definition of facial androgyny, to naturally androgynous faces. Although the two groups were rated as equally androgynous, the naturally androgynous faces were rated as significantly more feminine. Chapter 3 focused on understanding androgynous face perception while minimizing prior assumptions. In a series of experiments, participants handled androgyny-related tasks and stimuli in a way that was qualitatively and quantitatively different than their gendered counterparts. Overall, these results suggest androgyny as a category cannot be accurately summarized as halfway between male and female and that a more nuanced approach to studying face gender is needed.

Acknowledgements

To my supervisors, Dr. Patrick J. Bennett and Dr. Allison B. Sekuler: Thank you for your continued guidance and support. Your encouragement over these years has helped me grow as a student, researcher, and person. I would also like to dearly thank my committee member, Dr. Caroline Blais, for her wisdom and input on my projects.

I could not have made it to this point without the friendships I have made in the department of Psychology, Neuroscience & Behaviour. I will not think back to these years and think of isolation and quarantine; I will think of midday park trips, co-working on the couch, and giving haircuts in every possible location. To my fellow VisLab members, past and present: thank you for all your moral support, research feedback, company, and group puzzle breaks.

To my biological and chosen family: your support and patience has kept me sane, your drives to Hamilton got me to change out of my sweatpants, and your willingness to help me along this journey has made me feel so loved.

Contents

List of Figures

Declaration of Academic Achievement

I, Leigh Greenberg, declare that this thesis titled, "Androgynous and Gendered Face Perception" and the work presented in it are my own. I confirm that all chapters have been conceptualized, carried out, and written by me with the guidance and advice of my advisors and advisory committee.

Chapter 1

General Introduction

People are highly skilled in categorizing faces, both explicitly and implicitly (Wiese et al., 2008). Having this ability is important in deciding how to interact with an individual. Recognizing traits such as trustworthiness or mood, for example, is extremely useful for inciting favourable social interaction, and luckily we can do this with relative ease (Engell et al., 2007; Wild et al., 2001). Conversely, coming across a face that does not fit into a previously established category may be troubling, as we have less information available to inform how we will act with that individual.

1.1 Androgynous Faces in Research and the Validity of Morphing

Many researchers have investigated how people categorize faces, often using faces that may not belong to a specific category to control for category effects and maximize the effect of their variable of interest. Androgynous faces generally are considered to be an inter-category stimulus between the categories of male and female, and are commonly used in research. For example, Hess et al. (2009) assessed whether certain facial expression cues would bias viewers to categorize androgynous faces as being male or female. However, the use of androgynous faces is not limited to studies of gender perception; they have been used to investigate processes such as visual adaptation (Webster and MacLeod, 2011), the effect of empathy on face perception (Kosonogov et al., 2015), and face inversion and orientation effects (Watson and Clifford, 2006).

A way to visualize androgynous faces as an inter-category stimulus is to consider a continuum of all possible faces, sorted by masculinity and femininity; the most masculine faces would fall at one end, the most feminine faces would fall at the other end, and "androgynous" faces would fall at and around the midpoint (e.g. Webster et al., 2004;

Kloth et al., 2010). This would also imply that that the more masculine a face is, the less feminine it would be (and vice versa) and also that androgynous faces are equally masculine and feminine, or at least perceptually so. In practice, many researchers achieve this result by digitally manipulating femininity and masculinity, including morphing faces to achieve androgynous stimuli. Compared to sourcing natural (non-morphed) androgynous faces, this morphing technique provides a greater degree of stimulus control. Natural androgynous faces are not commonly used in research, meaning there is a lack of literature and knowledge about a subset of human faces. Further, no studies have investigated whether natural androgynous faces fit the same definition of "androgynous" used to describe digitally altered, equally masculine and feminine faces. The present thesis sought to investigate the validity of this common model.

Many studies have investigated which facial features cause a face to appear masculine or feminine. For example a man's nose is larger, and a woman's brows are more arched (Brown and Perrett, 1993). However, because androgyny has been conceptualized as an inter-category descriptor instead of its own category, the possibility of androgyny markers outside of "equally masculine and feminine" have not been explored. fMRI findings suggest that although people perceive gender categorically, initial activity in the fusiform face area is linearly correlated with the changes of how gendered a face objectively is, where "gendered" is defined as the opposite of androgyny (Freeman et al., 2010). In other words, the brain encodes the gender information in a face objectively, but then perceives the faces in a subjective, top-down approach. These findings lend support to the idea that perceptual categorization is determined not entirely by objective qualities of the photo, but at least to some degree by the category choices available to an individual at test.

Valentine's face-space theory proposes two possibilities for how people categorize masculine and feminine faces. Although he does not mention androgynous faces, the theory can be extrapolated to include them. His first proposal describes a prototype effect for perceiving gender: people have abstracted prototypes for both masculine and feminine faces from extensive experience, and they are comparing the faces they see to those prototypes to decide the face's gender (Valentine, 1991). Assuming "androgyny" follows the common definition, Valentine's theory might place androgynous faces approximately equally distant from masculine and feminine prototypes. Alternatively, it is possible that androgynous faces have their own prototype rather than simply being equally far from both masculine and feminine prototypes. However, due to the relative rarity of these faces, the prototype may not be as developed as gendered prototypes. To evaluate which

of these cases are true, researchers would need to make "androgynous" a category choice alongside "masculine" and "feminine" in face experiments. Although the present thesis does not aim to build on the face-space theory, it takes this point into consideration.

Valentine's alternate proposal is that face categorization is exemplar based. In this case, a given face is compared to every face that the perceiver has previously seen (exemplars), and the judgment given to that face depends on which exemplar it most closely resembles. In this framework, distinctive faces are dissimilar from any stored exemplars, whereas typical faces can have many similar exemplars (Valentine, 1991). In this scenario, androgynous faces could be considered distinct due to their relative rarity, making categorization difficult. Further, androgynous faces may be considered androgynous for varying reasons (e.g. having all features be gender netural, or having a mix of masculine and feminine features), meaning that even if an individual has many stored exemplars for androgynous faces, their distinctiveness from each other may make categorizing by proximity to existing exemplars near pointless. In the present thesis, we considered the possibility of faces being androgynous in different ways.

Some androgynous face morphs are created by using purely objective techniques, such as creating a face that is 50% male and 50% female (e.g. Riva et al., 2011). However, research has found that androgynous faces are not immune to adaptation effects, suggesting that these faces are not perceived purely by their objective gender composition. When adapted to male faces, for example, viewers subsequently judged previously "androgynous" faces as more feminine, meaning that their idea of what "androgynous" looked like had changed, becoming more masculine (Webster et al., 2004). To avoid assuming what an androgynous face looks like, some studies will use the point of subjective androgyny, where a face is judged as male half the time and female the other half of the time (e.g. Watson and Clifford, 2006). However, these faces are often still made by morphing along the male-female continuum, which limits what an androgynous face can look like.

Race is a social judgment that shares similarities with gender: although face race is not inherently discrete, there are socially defined categories. Given that these categories are imperfect, many faces would be judged as falling between categories or belonging to multiple categories. In this way, multiracial faces might be useful perceptual equivalents to androgynous faces. Ma et al. (2022) found that the categorization of multiracial faces was both dependent on culture (e.g. which races are culturally subordinate) and features (e.g. the presence of hair). Further, morphed multiracial faces were more likely to be

categorized as multiracial than real multiracial faces (Ma et al., 2022), implying that morphed faces are not always appropriate perceptual stand-ins for real faces.

1.2 Alternatives to morphing

One way to visualize androgynous faces while minimizing bias is to use a reverse correlation approach to create classification images. In this approach, a participant will be presented with two stimuli and asked to choose which one fits a given criteria better. Across trials, the differences between the two stimuli are randomly varied, typically by using random noise fields. By averaging the responses participants make, an image can be created that visualizes which parts of an image lead to certain classifications, aptly called a classification image (Murray et al., 2005)

Dotsch and Todorov (2012) used reverse correlation to create multiple classification images based on social traits such as trustworthiness and dominance. This study used a specific type of noise correlation called "reverse noise correlation", where the noise imposed on one stimulus option is the exact mathematical opposite of the other. This is done to maximize the difference between two stimulus options, instead of having the noise on each be independent, which is more common. A notable benefit to this approach is that it can significantly reduce the number of trials required to achieve a reliable classification image. The final classification image was layered onto the base face to visualize a prototype for the category of interest. We applied this technique in the present thesis with the goal of visualizing an androgynous face while minimizing pre-existing bias of what that may look like.

1.3 Current Thesis

The current thesis sets out to investigate whether androgyny can be considered its own category as opposed to existing between established categories. Another goal was to compare androgyny in morphed and natural faces, including establishing a data-driven visualization of androgyny. The work in Chapter 2 sought to verify the previous assumptions that masculinity and femininity strength are negatively correlated, and investigate where morphed and natural androgynous faces lie on such a plot. To investigate this, participants viewed either natural (Part 1) or natural and morphed (Part 2) faces and rated them on levels of masculinity, femininity, and androgyny. For these ratings, all scales were 1-5 Likert-type scales where 5 indicated a face being highly representative of that category and 1 represented a face being highly unrepresentative of that category.

To investigate the possibility of "different" androgynous faces in part A, we established a subset of faces that were labelled "naturally androgynous?" using participant ratings. To create morphed androgynous faces for Part 2, strongly masculine and feminine faces were morphed in equal weight using webmorph.org (DeBruine, 2018). These were labelled "morphed androgynous faces". Further, in a categorization task of natural faces, the category options provided were "man", "woman" "mix" (of a man and a woman), and "neither" (a man nor a woman").

Finally, to produce a minimally biased visualization of an androgynous face, we used a reverse correlation approach based on that of Dotsch and Todorov (2012). In Chapter 3, we investigated whether this technique could produce androgynous and gendered face images based only on participant input. We then verified these results by asking naive participants to make various judgments about the faces as a whole and on a part-by-part basis. To give context to some of our results, we also investigated how the base stimuli were perceived.

References

Brown, E. and Perrett, D. I. (1993). What gives a face its gender? *Perception*, 22(7):829–840.

DeBruine, L. (2018). Webmorph (beta release 2). *Zenodo*. doi: 10.5281/zenodo.1073696.

Dotsch, R. and Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, 3(5):562–571.

Engell, A. D., Haxby, J. V., and Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *Journal of cognitive neuroscience*, 19(9):1508–1519.

Freeman, J. B., Rule, N. O., Adams Jr, R. B., and Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, 20(6):1314–1322.

Hess, U., Adams, R. B., Grammer, K., and Kleck, R. E. (2009). Face gender and emotion expression: Are angry women more like men? *Journal of Vision*, 9(12):19–19.

Kloth, N., Schweinberger, S. R., and Kovács, G. (2010). Neural correlates of generic versus gender-specific face adaptation. *Journal of Cognitive Neuroscience*, 22(10):2345–2356.

Kosonogov, V., Titova, A., and Vorobyeva, E. (2015). Empathy, but not mimicry restriction, influences the recognition of change in emotional facial expressions. *Quarterly Journal of Experimental Psychology*, 68(10):2106–2115.

Ma, D. S., Kantner, J., Benitez, J., and Dunn, S. (2022). Are morphs a valid substitute for real multiracial faces in race categorization research? *Personality and Social Psychology Bulletin*, 48(1):95–104.

Murray, R. F., Bennett, P. J., and Sekuler, A. B. (2005). Classification images predict absolute efficiency. *Journal of Vision*, 5(2):5–5.

Riva, P., Sacchi, S., Montali, L., and Frigerio, A. (2011). Gender effects in pain detection: Speed and accuracy in decoding female and male pain expressions. *European Journal of Pain*, 15(9):985–e1.

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.

Watson, T. L. and Clifford, C. W. (2006). Orientation dependence of the orientation-contingent face aftereffect. *Vision Research*, 46(20):3422–3429.

Webster, M. A., Kaping, D., Mizokami, Y., and Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, 428(6982):557–561.

Webster, M. A. and MacLeod, D. I. (2011). Visual adaptation and face perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1571):1702–1725.

Wiese, H., Schweinberger, S. R., and Neumann, M. F. (2008). Perceiving age and gender in unfamiliar faces: Brain potential evidence for implicit and explicit person categorization. *Psychophysiology*, 45(6):957–969.

Wild, B., Erb, M., and Bartels, M. (2001). Are emotions contagious? evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences. *Psychiatry Research*, 102(2):109–124.

Chapter 2

Comparing morphed and natural androgynous faces

2.1 Introduction

Morphed faces are frequently used as perceptual stand-ins for real androgynous faces (e.g. Yang et al., 2011; Freeman et al., 2010). No previous work has explored the possibility that androgynous face morphs may be qualitatively or quantitatively different from real faces that are judged to be androgynous. A finding that these two types of androgynous faces are not perceived in identical ways could potentially undermine the validity of using morphed androgynous faces in certain types of face experiments.

Ma et al. (2022) compared biracial morphed faces (e.g. morphing one white face with one black face) to real faces of biracial or multiracial individuals. Morphs were judged as biracial more frequently than natural faces when both face types were masked to exclude hair. When the real faces were unmasked, they were judged as biracial as often as (masked) morphs were. Although this study focuses on race and not gender, both concepts are similar in that they are socially constructed categories. Further, in both cases, information about category belonging can be gleaned from looking at a person's face, although these observations are not always accurate to reality. Lastly, both biracial/multiracial faces and androgynous faces are more rare in everyday life relative to single-category faces. The findings of the Ma et al. (2022) study suggest that androgynous face morphs may be perceived differently from real androgynous faces.

This experiment explored the idea that androgyny and gender may have a more nuanced relationship than previously assumed. Further, it compares real androgynous faces to morphed ones, which have previously been using interchangeably. In Part 1,

participants answered questions about gender and androgyny in a sample of face stimuli. To investigate how naturally androgynous faces compare to morphed ones, morphed stimuli were created based on the information collected. Participants in Part 2 completed a similar task as in Part 1, but with the addition of morphed faces. The goal of this was to evaluate whether the various gender judgments change for morphed and natural androgynous faces.

2.2 Methods

2.2.1 Observers and Ethics

One hundred seven undergraduate students at McMaster University took part in this experiment (93 female; 17-57 years old, $M = 18$; $SD = 0.72$). Eighty-two participants took part in Part 1 and twenty-five participants took part in Part 2. All observers reported normal or corrected-to-normal vision. Participants were compensated with partial course credit for their participation. All participants provided written informed consent prior to the experiment. All protocols were approved by the McMaster Research Ethics Board.

2.2.2 Apparatus and Stimuli

Participants completed the experiment in groups of up to 30 in the McMaster LiveLab, an auditorium which allows for mass testing. Participants were provided with touch-screen tablets to make their responses while face stimuli were presented on a screen for 5 seconds each. 193 Caucasian face stimuli (91 female, 102 male) were sourced from the Radboud Faces Database (Langner et al., 2010), the Karolinska Directed Emotional Faces Database (Lundqvist et al., 1998), and the Psychological Image Collection at Stirling (Hancock, 2008). To minimize differences across face sets, the following manipulations were performed: the eyes, nose, and mouth were aligned, the average luminance was adjusted to be more uniform, and a uniform grey background was added. Face stimuli were presented in greyscale, without any masking to remove hair or ears. All stimuli were displayed with a uniform grey background and with consistent luminance across faces. All stimuli measured 1100x1200 pixels.

2.2.3 Experimental Design

After providing written consent, participants were given tablets and assigned seats in the LiveLab auditorium. Due to the nature of group testing, question conditions could

not be counterbalanced. The participants answered the questions "How androgynous is this face?", "How feminine is this face?", "How masculine is this face?", and "What gender is this face?", respectively. The first 3 questions used a 1-5 Likert-type scale, ranging from "Not at all" (1) to "Somewhat" (3) to "Extremely" (5), whereas the final question was forced-choice and participants had to choose from the options of "man", "woman", "both", or "neither". Participants responded to a single question for all 193 faces before moving on to the next question, meaning that they saw each face a total of 4 times. Face order was shuffled across question conditions. Participants were given verbal instructions for each question condition and informed that there were no correct answers. Because of the nature of the task, feedback was not provided. In Part 2, 36 morphed faces were added. Further, participants in Part 2 did not complete the final question "What gender is this face?" due to technical issues on the day of testing. Otherwise, all other experimental procedures remained the same in Part 2.

2.2.4 Morphed Stimuli

Morphed faces were created using webmorph.org (DeBruine, 2018), a free website which allows users to upload, delineate, and morph faces. Three types of face morphs were created: androgynous, masculine, and feminine. To create morphs, "highly masculine" and "highly feminine" faces were subsetted based on the ratings from Part 1, having a mean rating of at least 4 for their respective category. To create androgynous morphs, 12 highly masculine and 12 highly feminine faces were randomly selected from the subsets and paired so that each androgynous morph was comprised of one highly masculine and one highly feminine face. Masculine and feminine morphs were created to control for the possibility that the androgynous morphs may be perceived differently solely because they are computer morphs. To create the masculine morphs and feminine morphs, two faces were selected from the same category instead of one from each, creating 12 masculine morphs and 12 feminine morphs in addition to the 12 androgynous morphs. Contributing faces were only used once each across all morphs. All contributing faces were removed from the experiment for Part 2 and 36 new faces were added to preserve the same number of total stimuli. As all contributing faces were in greyscale, so were the morphs. All morph images were saved as JPGs.

2.3 Results

2.3.1 Initial Processing

For the three rating questions, ratings were combined across all observers and averaged together to create a mean rating for each face and each judgment. To establish a set of naturally androgynous face stimuli, any faces with a rating of at least 3 ("Somewhat androgynous") for the judgment "How androgynous is this face?" were deemed as being naturally androgynous faces. Of 193 faces, 19 were labelled naturally androgynous. All but three were women's faces. Naturally androgynous faces were judged to be as androgynous as the androgynous morphs created for Part 2 (p = 0.87) (Figure 2.1). For categorical judgments ("What gender is this face?"), data was combined across all participants and for each face the proportion of responses in each of the four options was calculated.

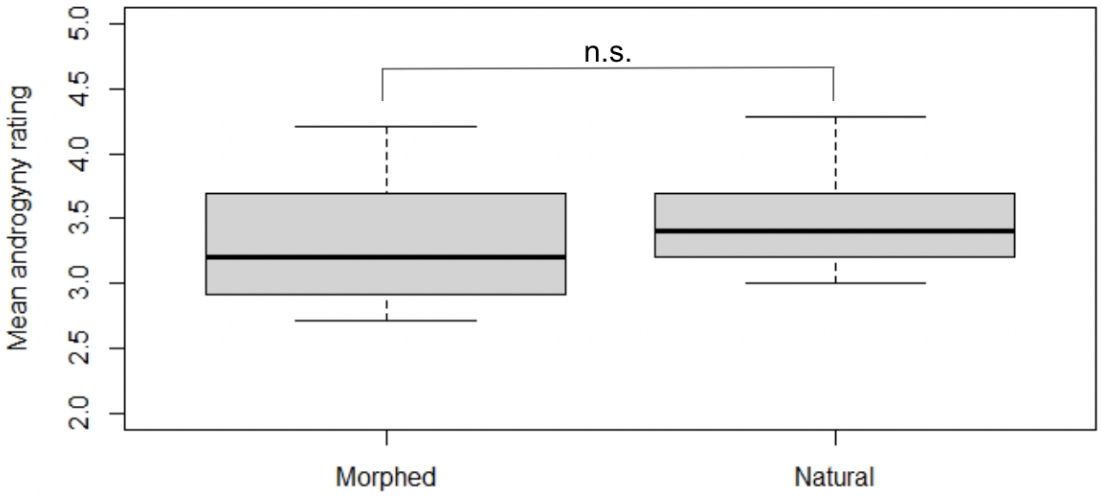


FIGURE 2.1: Averaged responses for the question "How androgynous is this face?" for the subsets of morphed androgynous faces and naturally androgynous faces.

2.3.2 Ratings Data

All rating information is visualized in Figure 2.2. The association between masculinity rating and femininity rating was strongly negatively correlated (adjusted R^2 = -0.83). The association for the androgynous faces only (natural and morphed) trends in the same direction but is much weaker (adjusted R^2 = -0.12). Morphed and naturally androgynous faces were not significantly different in their masculinity ratings, but naturally androgynous faces were significantly more feminine than the morphed androgynous faces

($p<0.05$). Both types of androgynous faces fell within a more constricted range of masculinity and femininity scores; androgynous faces, when averaged across participants, never received a masculinity rating under 2.17 or over 4.08. For femininity scores, androgynous faces were all within the bounds of 1.72 and 4.31. Conversely, many of the non-androgynous faces fell outside these bounds when considering both judgment types, as seen in Figure 2.2.

2.3.3 Categorical Data

Figure 2.3 shows, based on the gender of the face and divided into androgynous and non-androgynous stimuli, responses for non-androgynous face stimuli were correct (i.e. corresponding to the face's known gender) over 85% of the time. When the face was androgynous, "both" was a more common answer than either "man" or "woman". Although "woman" and "man" were chosen less frequently, observers were generally accurate to the face's known gender when they did use these options. However, it is worth noting that "woman" and "both" were chosen equally for women's faces, but "both" was chosen more often when judging a man's face. "Neither" was the least popular choice regardless of androgyny or the actual gender of the face, indicating that facial androgyny is perceived as being a mixture of genders as opposed to the lack of gender altogether.

2.4 Discussion

We found a strong negative correlation between masculinity and femininity in observer ratings. This is consistent with both previous research and the common understanding of the relationship between masculinity and femininity. However, we also found that different faces could be perceived as equally androgynous and simultaneously different in gender ratings, which goes against the idea that androgynous faces are equally masculine and feminine. Given these results, it is likely that there is something outside of balanced masculinity and femininity that drives androgyny judgments. This might be another social judgment (e.g., dominance, age), or the presence of specific physical attributes (e.g., ratios between face parts).

We also found that morphed androgynous faces are not perceived identically to natural androgynous faces. Despite being rated as equally androgynous on average, naturally androgynous faces were rated as significantly more feminine than the morphed faces. This is likely the case because most of these stimuli were in fact women's faces, meaning that observers were able to glean information about the correct gender while also acknowledging that the face looked very androgynous. Alternatively, it could be that

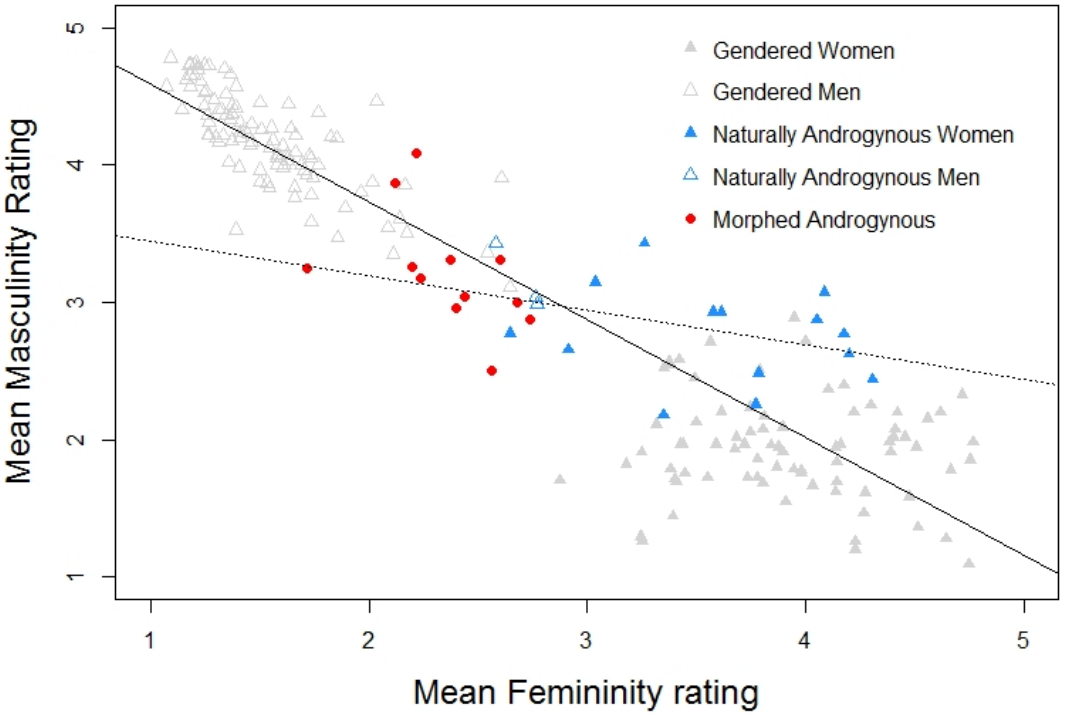


FIGURE 2.2: All face stimuli plotted by their average masculinity and femininity ratings. Faces are divided into androgynous or gendered (i.e. not androgynous). For non-morphed faces, the gender of the face is shown. Solid line represents the correlation for all faces. The dotted line represents the correlation for the androgynous faces only.

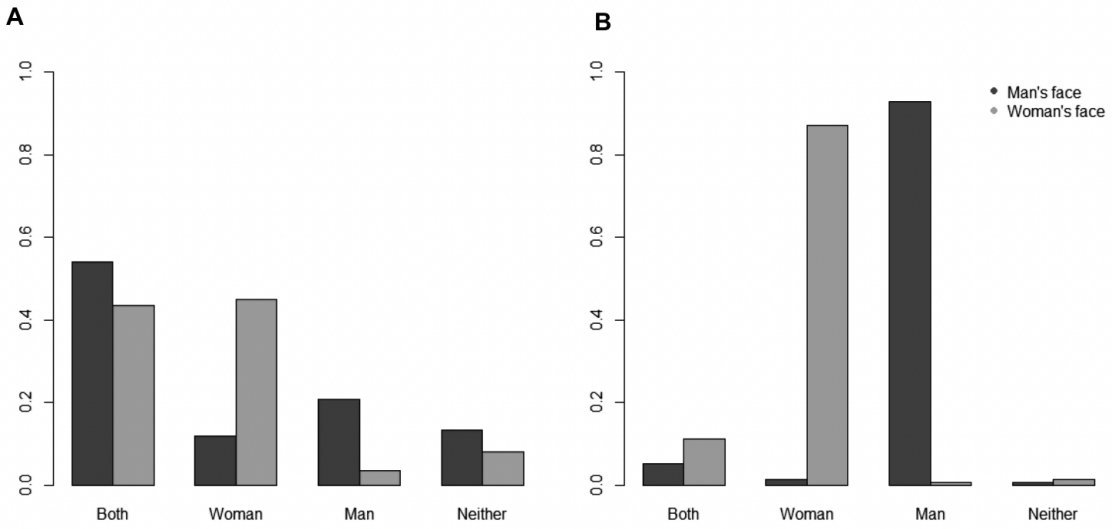


Figure 2.3: How often each of the four response options were chosen in response to "What gender is this face?", based on the gender of the face and whether the face was naturally androgynous (A) or not (B).

the morphed faces were biased towards looking more male. This finding suggests that the abstract concept of androgyny needs to be more carefully operationalized in future studies, as androgynous faces are not a heterogeneous category. Further, as suggested by the Ma et al. (2022) study, morphed faces may be suitable for some experimental purposes but not others. If a researcher is interested in presenting gender-ambiguous stimuli, morphed faces might be suitable, as they are as ambiguous as real androgynous faces. Conversely, if the goal is to specifically visualize faces that are half feminine and half masculine, extra pre-testing steps should be considered, such as verifying that the stimuli are perceived as equally masculine and feminine or judged as male and female equally.

In the categorical data, we found that for androgynous faces, women's faces were equally likely to be labelled "woman" as they were "both" man and woman. Conversely, androgynous men's faces were more than twice as likely to be chosen as "both" compared to the accurate gender judgment. This might suggest that participants had an easier time assigning binary gender to androgynous women's faces, acknowledging that they are women's faces that look androgynous, but having a harder time making a similar judgment for men's faces. For example, if a man's face appeared androgynous, it was harder to continue to judge it as being a man's face. This difference between judging androgyny in men's and women's faces could be due to the specific stimuli that we used, the low number of men's faces in our naturally androgynous face set, or a reflection

of a social acceptance for androgynous women's faces but not androgynous men's faces. Lastly, we found that when given the option to judge androgynous and non-androgynous faces as "both" man and woman or "neither" man nor woman, *both* was a far more popular choice, suggesting that when a face falls outside of a binary gender, observers judge it as belonging to more than one gender (or having qualities of more than one gender), as opposed to having an absence of gender altogether. Given the typical definition of facial androgyny as being equally masculine and feminine, "equal" could imply that a face has gender information that is equal across the contributing genders, or that the face does not belong to (or have features from) any gender, equally. Although both alternatives are possible in theory, this finding shows that observers bias toward the former.

References

DeBruine, L. (2018). Webmorph (beta release 2). *Zenodo*. doi: 10.5281/zenodo.1073696.

Freeman, J. B., Rule, N. O., Adams Jr, R. B., and Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, 20(6):1314–1322.

Hancock, P. (2008). Psychological image collection at sterling (pics). http://pics.psych.stir.ac.uk.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388.

Lundqvist, D., Flykt, A., and Ohman, A. (1998). Karolinska directed emotional faces [database of standardized facial images]. *Psychology Section, Department of Clinical Neuroscience, Karolinska Hospital, S-171*, 76.

Ma, D. S., Kantner, J., Benitez, J., and Dunn, S. (2022). Are morphs a valid substitute for real multiracial faces in race categorization research? *Personality and Social Psychology Bulletin*, 48(1):95–104.

Yang, H., Shen, J., Chen, J., and Fang, F. (2011). Face adaptation improves gender discrimination. *Vision research*, 51(1):105–110.

Chapter 3

Visualizing and analyzing androgynous faces using classification images

3.1 Experiment 1

3.1.1 Introduction

Classification images (CIs) are an excellent tool for visualizing faces in a data-driven way. Although there are many techniques for creating classification images, we chose to use reverse correlation. We specifically chose to model our experiments after those of Dotsch and Todorov (2012) as they were able to visualize complex social categories such as dominance and trustworthiness in just 300 trials. This is in contrast to most classification image studies which require several thousand trials (Gold et al., 2004; Sekuler et al., 2004; Creighton et al., 2019). A typical method of generating CIs would have two random noise fields added to two identical base images. Participants are asked to choose between two stimuli. In a reverse correlation approach, the two noise fields are not independent; they are completely negatively correlated, meaning that a dark pixel in one noise field will be light in the other. This detail maximizes the difference between the stimuli and therefore reduces the number of trials needed to establish a CI.

In the following study, we aimed to visualize androgynous and gendered faces with the hope of establishing how people conceptualize these categories.

3.1.2 Methods

Participants

Three hundred fifty nine participants were recruited through Prolific, a platform that connects participants to online studies. All participants indicated that they were fluent in English and were at least 18 years of age. The median completion time was 18.5 minutes. Nine participants who completed the study in less than seven minutes were not included in our analyses, leaving a total of 350 participants (139 female, 18-74 years old, $M = 32.3$, $SD = 10.9$)in the final sample. All protocols were approved by the McMaster University Research Ethics Board.

Stimuli

Stimuli were images of faces embedded in noise. Ten faces were used from the Gold et al. (1999) face set. Five of the identities were female, five were male. All faces were Caucasian, facing forward displaying a neutral expression, and cropped with an oval window to omit hair and ears. All face images were 256x256 pixels.

To create the noise for one trial, a 64x64 field of random scalar values were drawn from a standard normal distribution using the MATLAB `randn` function. These values were multiplied by a root-mean-square (RMS) value of 0.2. This noise field was then scaled up to match the size of the face stimuli (256x256 pixels), meaning that each pixel in the original noise field took up a 2x2 pixel region after rescaling. The negative noise field was created by multiplying all values in the first noise field by -1. The base face was multiplied by an RMS value of 0.1, then each noise field then had the base face added to it. As these two stimuli were represented by contrast values (ranging from -1 to 1) they needed to be converted to be displayed as images to the participant. To do this, we used customized software to convert the contrast values to luminance values, which were then converted to bit-stealing numbers that corresponded to an RGB lookup table. Then, a function was applied which selected the closest RGB value (0 to 255) that matched the luminance value. This process was repeated 300 times to create pairs of stimuli for every trial. Note that this method did not correct for non-linearities in the display. Also, the same 300 noise pairs were used for all participants and all face identities.

This task was performed on laptops and desktop computers, and therefore the actual size of the stimuli varied across participants. We measured stimulus size on a variety of laptops and desktops, and the height of the 256 x 256 stimulus ranged from 5 to 6.2 cm (5.39 to 6.67 degrees). The face itself measured about 200 pixels tall, meaning

that on-screen, it ranged from 3.9 to 4.84 cm (4.21 to 5.22 degrees). Participants were instructed to view the stimuli from a consistent and comfortable distance. The average viewing distance used by 4 naive observers was 53 cm.

Procedure

On each trial, a participant would see a blank screen for 500ms followed by a pair of stimuli presented 275 pixels to the left and right of centre. Depending on the condition, participants were asked to select the face that appeared more androgynous, female, or male. The stimuli remained on the screen until the participant made a selection, at which point the next trial would automatically start. The relevant prompt remained on the screen during all trials. Responses were made by pressing keys on a computer keyboard. Pressing the F key indicated choosing the left face and pressing the J key indicated pressing the right. Participants viewed 300 such trials with self-timed breaks at the 100 and 200 trial marks.

Each participant saw one randomly assigned face identity which remained constant across trials. In other words, the base face remained the same and only the noise changed across trials. Each participant made one judgement (i.e., androgynous, female, or male) about the stimuli shown on each trial. Participants were assigned randomly to the three judgment conditions with the constraint that there were 100 participants per group for the gendered conditions and 150 per group in the androgynous condition. Participants were told that there were no right or wrong selections and to simply select the face that they thought was the best answer. Further, participants were told that they had unlimited time to respond on any given trial but that they should usually aim to respond within a few seconds. To minimize researcher intervention, participants were not given a definition of male or female, with the assumption that anyone fluent in English would have their own notions of these categories. However, participants performing the androgyny task were instructed that "An androgynous face can be thought of as a face which cannot be categorized as simply male or female". This instruction was included after some English-fluent participants completing the experiments in Chapter 2 were unclear on the meaning of androgyny. This definition of androgyny was intended to be as objective as possible so as not to impose researcher expectations while still clarifying what concept the participants should be evaluating. This definition was used in Experiments 1-4 in any condition where androgyny was mentioned.

3.1.3 Results

The noise fields from all of the selected images (i.e. the more androgynous/male/female on each of 300 trials) were averaged together to create the classification image (CI) for that category. The non-selected images also were averaged together to create the anti-classification image (anti-CI). Like the individual trials, the CI and anti-CI are negatively related to each other. Finally, data from participants assigned to the same identity and judgment were averaged together, resulting in a total of 30 pairs (10 identities x 3 judgments) instead of one pair per participant.

To investigate which areas of the CIs and anti-CIs were responsible for transforming the base images, we used the stat4CI toolbox (Chauvin et al., 2005) to identify any significantly light or dark pixel clusters in the CIs using their cluster test function. Significant clusters can be interpreted as areas of the CI (above a certain size threshold, which is calculated based on the size of the smallest expected feature in the face) that are associated with the given judgment. The CIs were smoothed with a Gaussian filter (σ = 20 pixels) and Z-transformed. Then a two-tailed cluster test was performed (a_{fw} = 0.05; t_{crit} = 3.66). The analysis was performed on thirty CIs (3 tasks x 10 faces). No meaningful significant or significant clusters were found among any of the CIs. This failure to find significant clusters occurred even with smaller values of σ. This finding might suggest that the CIs are not visualizing the intended categories. Alternatively, it could be that this technique is not sensitive enough to find the diagnostic areas in our CIs.

3.1.4 Discussion

The lack of meaningful or consistent significant clusters in the pixel-wise analysis could suggest that there is no difference between the CI and a field of random Gaussian noise. This may be due to our use of the same 300 noise fields across all faces and participants, which is not typical in a classification image experiment. It is difficult to say with certainty whether this decision contributed to the lack of effect in our cluster analysis. However, given that visual evaluations concluded that most CI/anti-CI pairs had a distinct effect on the base face, it is more likely that the technique we used for cluster analysis was simply not sensitive to the structure of our CIs. One alternate method to pixel-wise analysis is a behavioural study in which naive participants are asked to make judgments about the CI/anti-CI pairs and their agreement (or lack thereof) is investigated as a measure of effect of CI. We conduct this study in Experiment 2.

3.2 Experiment 2

3.2.1 Introduction

The aim of Experiment 2 is to validate whether the results of Experiment 1 were due to a true lack of difference between CI and anti-CI or a poorly suited analysis tool. To do this, we created a behavioural experiment where naive viewers evaluated the CI and anti-CI pairs in a task similar to that of Experiment 1. If the analysis in Experiment 1 was valid in concluding that there were no differences, we would expect to see no difference in the behavioural study. However, if there are differences, it may be that the tool we chose to analyse the images was simply not compatible with our CIs.

3.2.2 Methods

Observers and Ethics

Two hundred fifty participants participants (101 female, 3 unknown; 18-54 years old, $M = 25.5$, $SD = 7$) were recruited through Prolific. All participants indicated that they were fluent in English prior to participating. Median completion time was 3.6 minutes. Participants were compensated at a rate of £7.50 per hour. Pavlovia users who participated in Experiment 1 were prevented from participating in the current experiment so as to ensure all participants were naive to the study. All protocols were approved by the McMaster Research Ethics Board.

Stimuli and Experimental Design

Stimuli were the CI and anti-CI images from Experiment 1 with the base faces added in: thirty pairs of images were created from the ten base face images and the three judgment conditions (androgyny, masculinity, femininity) by layering the base face at an RMS of 0.1 and the noise fields at an RMS of 0.2, ending up with two images: one base face with CI and one base face with anti-CI. Examples of the resulting images are shown in Figure 3.3.

The trials were set up in the same way as in Experiment 1: participants saw the face pairs after a 500ms delay and the images remained on the screen until a keyboard selection was made, at which point the next trial would immediately start. Participants were given the same instructions and questions as in Experiment 1, and were naive to the fact that they were viewing CI/anti-CI pairs. The order of stimuli within a block were randomized. Similarly to Experiment 1, the same definition of androgyny was provided for the androgyny trials. Unlike Experiment 1, the judgment tasks were within-groups.

This was done as there were only 30 stimuli to judge. The judgments were performed in blocks and the order of blocks was randomly shuffled for each participant. The question/task for each block matched the source of the stimuli. For example, for the androgynous judgment participants were shown the 10 pairs (one for each face identity) of CI and anti-CI images generated from the androgynous condition in Experiment 1 and were asked about which face appeared more androgynous. Similarly, for the male and female judgments participants were shown the 10 pairs of CI and anti-CI images from the male and female conditions in Experiment 1 and were asked the relevant question. We also chose to have the participants do every block twice, totaling 60 trials, so that we could evaluate whether people judged the faces consistently. As each face x question block was seen twice throughout the experiment, we ensured that the sides were counterbalanced such that the position of the faces was reversed the second time the block was seen. For example, if the CI was on the left the first time a given block was seen (and the anti-CI was on the right), then the CI would be on the right in the second showing (with the anti-CI on the left).

On each trial, the identity, task, and gender of the face was recorded. In total, there were 30 different stimuli pairs: each of the ten identities was shown three times, once for each task. For each of these 30 pairs, an agreement score was calculated, which represented the percent of participants that selected the CI as being more androgynous, male, or female.

3.2.3 Results

To evaluate how participants judged the CIs and anti-CIs, we plotted their percent agreement score as a function of judgment and face identity in Figure 3.1. Percent agreement was high for all face identities in the male and female conditions, but varied significantly with face identity in the androgynous condition. For the male and female CIs, agreement was over 75% for all identities. For the androgynous CIs, only half fell above this threshold. Of the remaining five CIs, only one fell above the range of chance, which was established by calculating a 95% confidence interval for a model that assumed every observer responded randomly (Figure 3.1). As a whole, all three tasks were significantly above chance ($p<0.05$). However, the tasks were not equal: a Bartlett test showed a significant difference ($p<0.05$) indicating that variance was not homogeneous across levels of Judgment. A Welch ANOVA (using `oneway.test` with var.equal set to FALSE) evaluating the three tasks indicated that there was a significant difference between groups ($F(2,14.527) = 11.628$, $p<0.05$). Specifically, agreement in the androgyny-judging task was significantly lower than agreement in the male or female judging tasks (Tukey's HSD

test, $p<0.05$ between Male-Androgyny and Female-Androgyny; $p>.05$ for Male-Female), as seen in Figure 3.2A). Removing the four at-chance faces identified in Figure 3.1 dramatically decreased the variances across the three tasks but did not change the results of significance testing (Figure 3.2B).

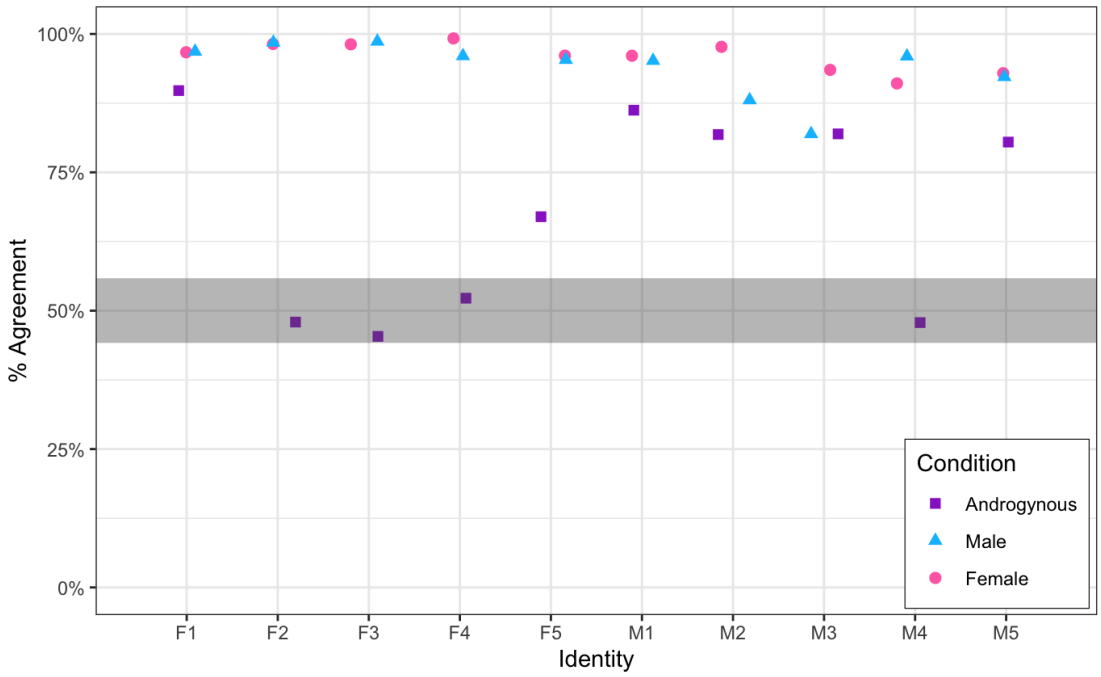


FIGURE 3.1: Averaged agreement for each judgment and face identity. The grey bar shows the bounds of random chance. % Agreement represents the percent of participants who agreed that the CI was more androgynous, male, or female than the anti-CI.

Visual Analysis of Stimuli

All stimuli are visualized in Figure 3.3. As expected from the percent agreement shown in Figure 3.1, the male and female judgment results are overall easier to distinguish when comparing the effect of the CI and the anti-CI on the base image. In the androgyny judgment results, some pairs are easily distinguished such as F1 and M1. Others, such as F2 and M4, which were determined to be at-chance for classifications, are difficult or impossible to visually distinguish in a meaningful way. Overall, in the gendered (male and female) judgment conditions, there seem to be consistent differences in lip thickness, expression, and darkness around the eyes in a way that might that suggest make-up. In the androgynous judgment condition, differences are overall more subtle even in faces that were above-chance in agreement. In more discernible cases, the CI tends to make

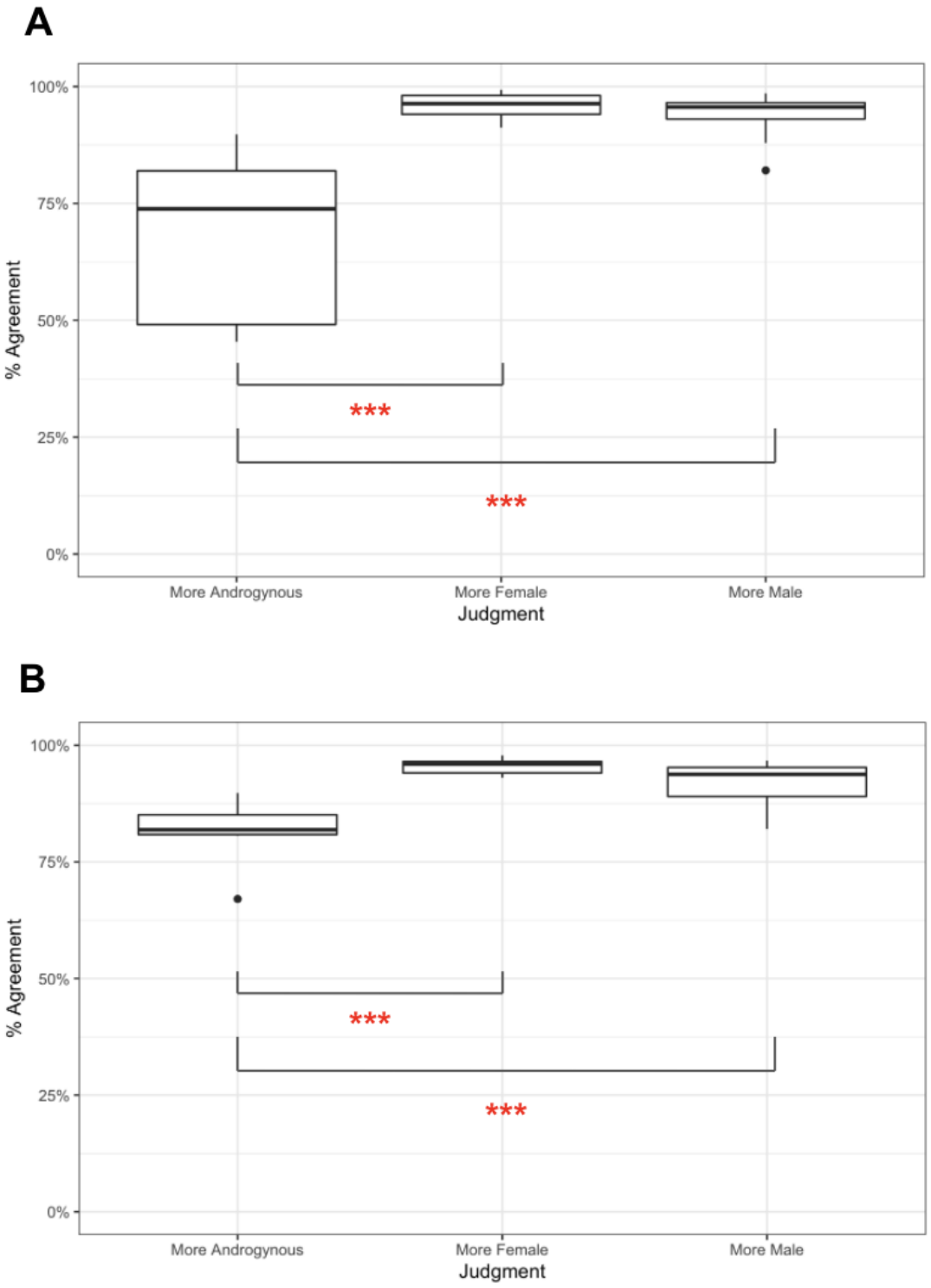


Figure 3.2: Boxplots of percent agreement scores for each judgment. Panel (A) shows data for all faces. Panel (B) shows the results after removing four faces (F2, F3, F4 and M4) where agreement was

the base face appear more like the opposing binary gender (e.g. the CI in F1 makes the face appear more male, whereas the CI in M1 makes the face appear more female).

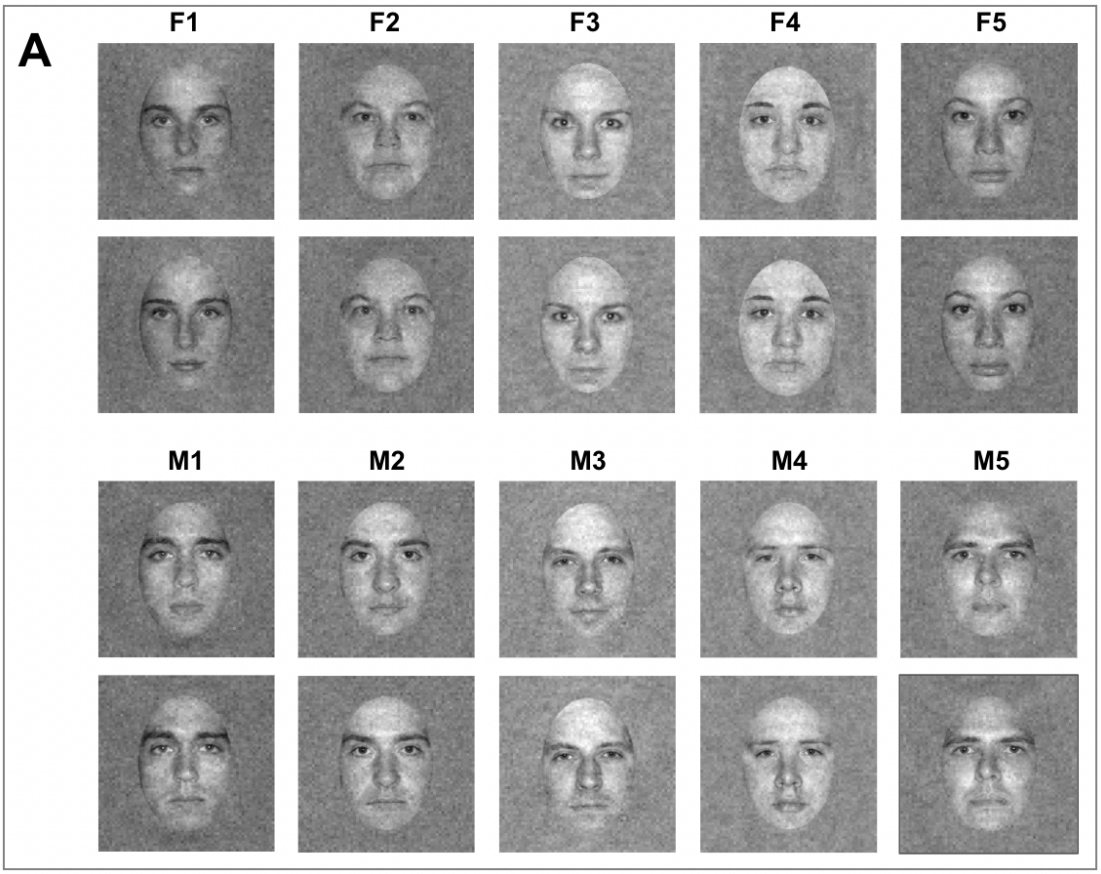
3.2.4 Discussion

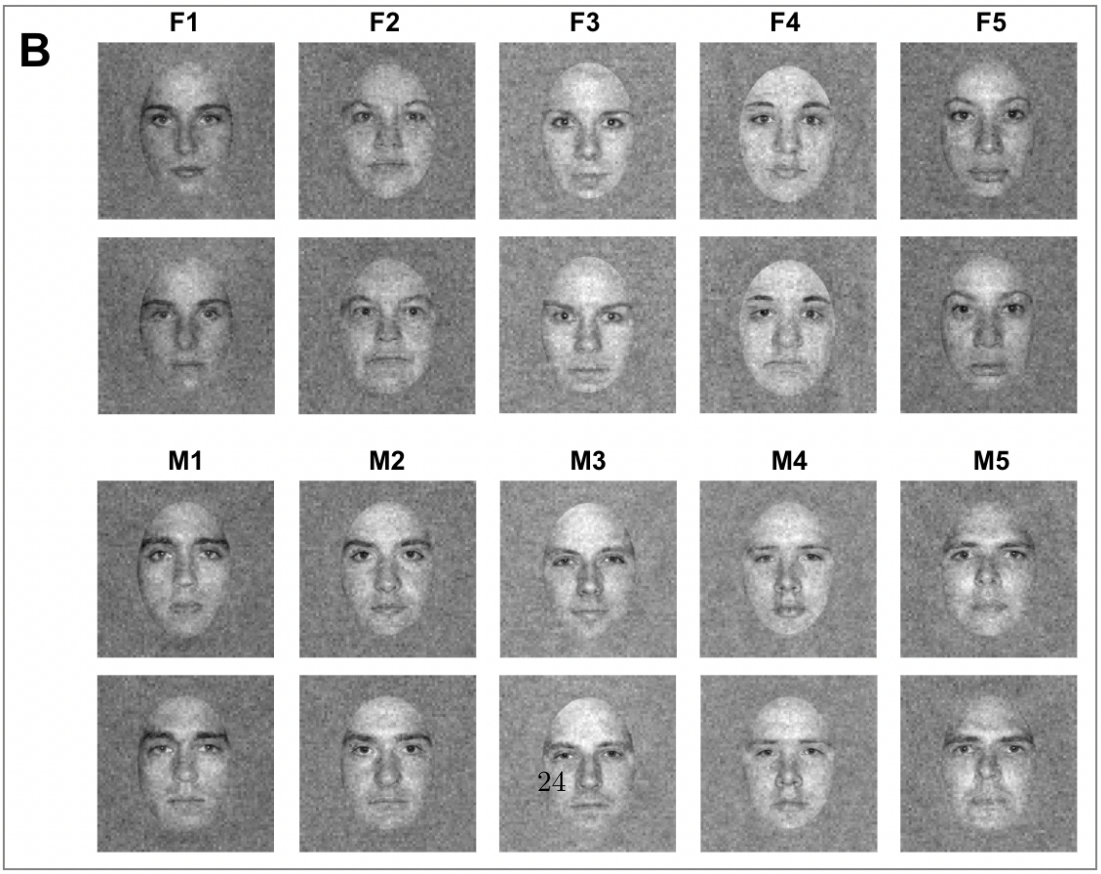
Visual analyses concluded that the difference between most CI/anti-CI pairs was easily visible when the base face was present. For the male and female judgments, the effects are in line with expectations of what we understand masculinity and femininity to look like from a Western lens. For example, faces judged more feminine had fuller lips(Brown and Perrett, 1993), more upturned mouths(Hess et al., 2009), and a smaller nose(Buchala et al., 2005), whereas faces judged more masculine had thinner lips, more neutral or negative expressions, and larger noses.

For the androgynous condition, the specific modulating effect of the CI depended at least somewhat on the perceived gender of the underlying face, and the effect of the CI (in cases where there was a discernible difference) seemed to androgenize the face by pulling it towards the other binary gender: a female face was made more androgynous by making it look more masculine, and masculine faces were made androgynous by making it more feminine.

We also found that the androgynous condition had a higher rate of indistinguishable CI/anti-CI pairs. If people are truly androgenizing faces by pulling them toward the opposite binary gender, it is possible that it is difficult for some faces to be androgenized if the original gender is more difficult to classify. It could also be that if we are to think as androgyny as being orthogonal to binary gender (male-female), people have an easy time "pulling" a male face to look more female and vice versa, but a face that already looks very androgynous might be harder to modulate. Further, it could be that participant results are canceling themselves out, either on an individual level (i.e. choosing a more feminine face as more androgynous on one trial but choosing a more masculine face as androgynous on another trial, due to the face that the base face is not providing sufficient information) or on a group level (i.e. some participants perceive a certain face in such a way that androgenizing it means making it more male, but others are inclined to make it more female).

Overall, participants agreed that the CI (+base face) was more prototypical than the anti-CI (+base face), which supports the conclusions of the visual analyses performed in Experiment 1. All three groups (androgynous, male, and female CIs) scored significantly above chance which indicates that participants could not only (in most cases) discern between the CI and anti-CI, but also agreed that the CI was more prototypical. While


A
F1
F2
F3
F4
F5
M1
M2
M3
M4
M5


B
F1
F2
F3
F4
F5
M1
M2
M3
M4
M5

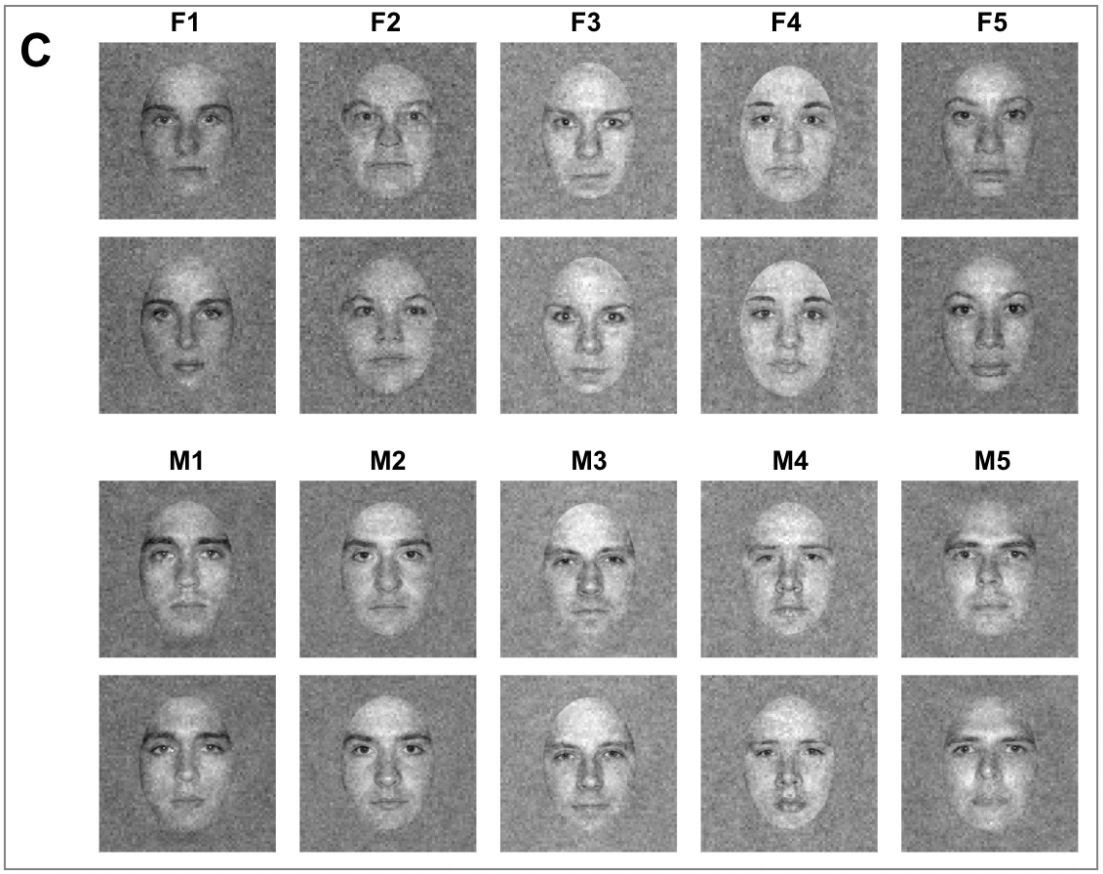


FIGURE 3.3: All CI and anti-CI pairs, with base faces, for each identity and judgment. A shows the androgyny judgment pairs, B shows the female judgment pairs, and C shows the male judgment pairs. In each pair, the CI is on top and the anti-CI is on the bottom.

performance for the gendered (male and female) CIs was near ceiling, scores for the androgynous CIs were significantly lower than either gendered group. While some androgynous CIs had similar agreement scores to the gendered CIs, 4 of 10 were at-chance. Removing these four faces did not change the relationship between the groups.

As previously discussed, this effect could be due to the qualities of our specific stimuli: maybe some of the base faces were already very androgynous and therefore participants from Experiment 1 had a difficult time making them look any more androgynous on a group level, as there is no single opposite of androgyny in the same way that many people would understand femininity as the opposite of masculinity. If this is the case, we would expect to see a relationship between the androgyny level of the base image and how easily the CI and anti-CI were discerned in the follow up, Experiment 2. An alternative to this theory is that participants do perceive androgyny in faces differently to binary gender.

3.3 Experiment 3

3.3.1 Introduction

The goal of Experiment 3 was to expand upon the results of Experiment 2. Given that some androgynous CIs were easily and consistently agreed upon but others were not, we theorized that this might have something to do with qualities of the base faces. Specifically, we sought to test the gender and non-gendered social qualities (trustworthniess, dominance, and attractiveness) associated with the faces as a means of explaining the variation seen in the CIs. In this Experiment, naive participants viewed the base faces only and were asked to judge them on various social qualities. By establishing a relationship between one or more of these qualities and agreement with CI, we can begin to explain why we saw such variation in the CIs. Further, we might be able to explain why this variation occurred for the androgynous CIs and not the gendered CIs.

3.3.2 Methods

Observers and Ethics

One hundred (38 female, 2 unknown; 18-64 years old, $M = 24.9$, $SD = 6.8$) participants were recruited through Prolific. All participants indicated that they were fluent in English prior to participating. The median completion time was 9.6 minutes. Participants

were compensated £7.50 per hour. Pavlovia users who participated in previous experiments were prevented from participating so as to ensure all participants were naive to the stimuli. All protocols were approved by the McMaster Research Ethics Board.

Stimuli and Experimental Design

The 10 face identity images were the stimuli for this study. These were the same images used in Experiments 1 and 2 but without any overlaid noise. The task was to rate each face on androgyny, femininity, masculinity, trustworthiness, dominance, and attractiveness. Each rating question was phrased as "How x is this face?" with the relevant quality filling in the blank (i.e. androgynous, female, male, trustworthy, dominant, attractive). The rating tasks were blocked, with an instruction screen before each block establishing that a new block was beginning and that there was a new judgment question being asked. The response scale was also consistent across all conditions, ranging from 1 ("Not at all") to 7 ("Extremely"), with 4 being labeled "Somewhat". Participants were told that there were no right or wrong selections, and to just answer based on their feelings. Each block was shown twice, shuffled randomly, yielding a total of 120 trials (10 faces x 6 conditions x 2). The two responses were averaged to create one rating for each face and judgment.

3.3.3 Results

The correlations between all six social judgments, are shown in Figure 3.4. The strongest ($\geq |0.7|$) correlations were: masculinity with femininity (-0.98), trustworthiness with attractiveness (0.8), attractiveness with dominance (-0.7), and androgyny with dominance (-0.7). Attractiveness has previously been positively linked to trust in faces (Zhao et al., 2015) and in voices (Rezlescu et al., 2015). Hoss et al. (2005) found that attractiveness can facilitate gender classification in binary faces, which may be related to our finding that androgyny (i.e. lack of a traditional binary gender) is negatively associated with attractiveness. Trustworthiness and dominance were weakly correlated (-0.2), a similar finding to Dotsch and Todorov (2012) who found a correlation of -0.27.

A linear model for the masculinity and femininity ratings provided an R^2 of 0.98. Given that these two ratings were so highly correlated, we transformed them into a single "Gendered" score, which describes how much a face was associated with a binary gender (i.e., male or female). This new score was created by converting the male and female ratings into a single gendered rating by calculating the principle component of the two ratings. The gendered score is shown in Figure 3.5, plotted against the androgyny

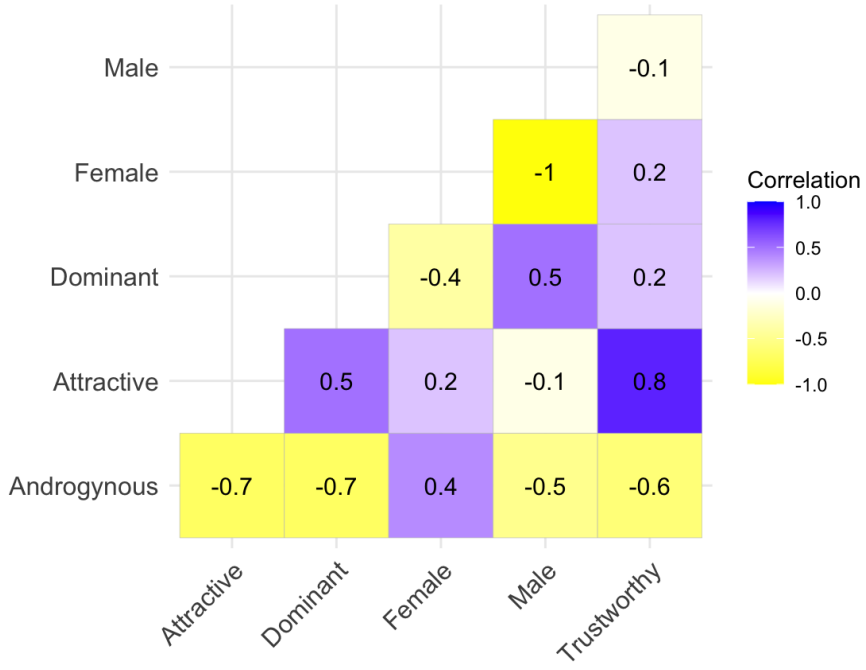


Figure 3.4: Correlations for all six judgments, rounded to the nearest tenth. Purple represents positive correlations and yellow represents negative correlations.

ratings for the same faces. The R^2 for the model involving these two scores was 0.77. Although it appears that the female faces tend to have overall higher androgyny scores, this effect was not significant ($F(1,10)=1.89$, $p>0.05$).

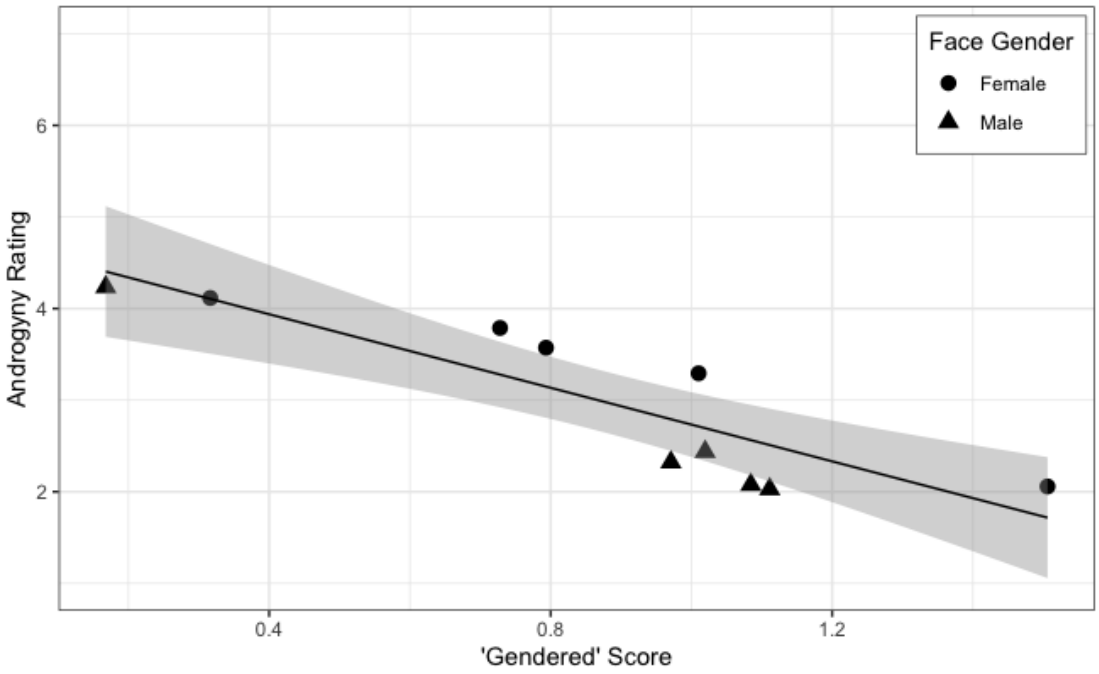


Figure 3.5: Gendered scores and ratings of androgyny for all 10 face identities. Circles represent female faces and triangles represent male faces. The 95% confidence interval for the linear regression line is represented by the grey shading.

Finally, we examined whether gender ratings of the base faces were associated with the agreement scores from Experiment 2. We were specifically interested in the androgynous judgments as this was the only condition in which participants in Experiment 2 performed at-chance in identifying the CI over the anti-CI. Figure 3.6 shows a very strong correlation ($R^2=0.89$) between base androgyny rating of an identity and the agreement score for the CI/anti-CI of that identity in Experiment 2.

In the masculinity condition, we saw a moderate correlation ($R^2=0.41$) and a significant effect ($p<0.05$), but most faces were still near ceiling (all above 80% agreement). Base femininity rating was not strongly correlated ($R^2=0.13$) with whether participants agreed that the CI was more feminine than the anti-CI, likely because all of the faces scored near ceiling. A linear model did not produce a significant effect of base femininity.

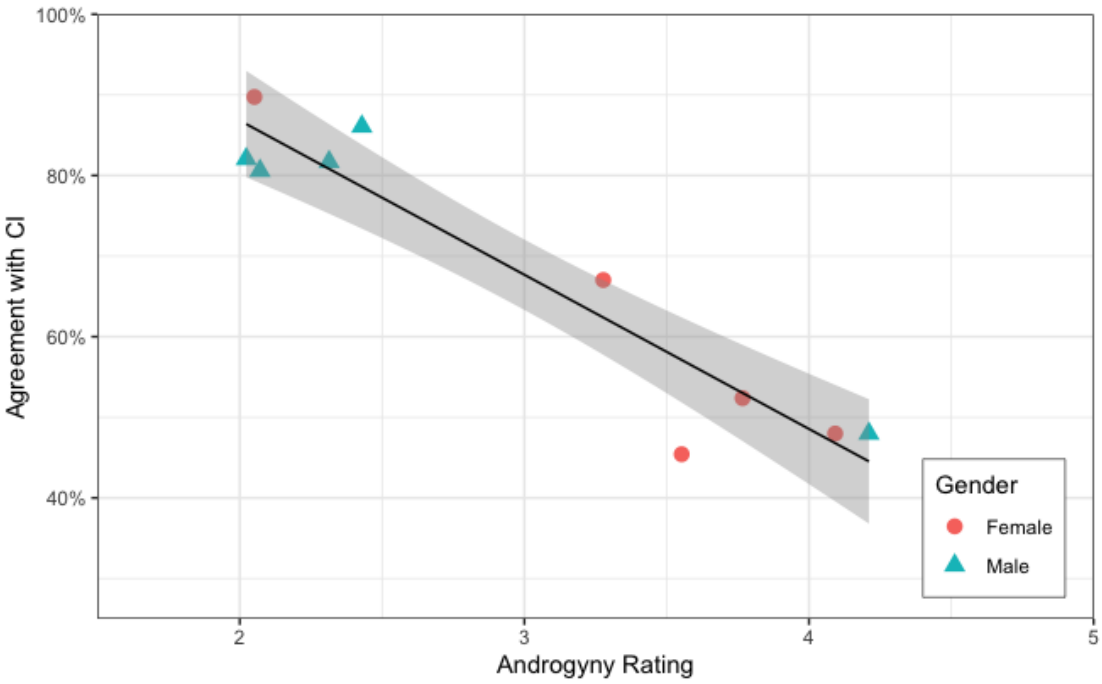


Figure 3.6: Base androgyny ratings compared to the agreement score for androgynous CIs from Experiment 2 using all ten identities. Colour and shape indicate the gender of the base face. Significance of the base rating is displayed as well as the adjusted R^2. Axes have been adjusted to best visualize the range of data.

3.3.4 Discussion

As in Chapter 1, we saw strong correlation between masculinity and femininity judgments. When these two judgments were transformed into one factor, gender, we saw a strong correlation with androgyny judgments. Androgyny was also negatively correlated with most judgments (aside from femininity), providing support for the idea that people associate binary gender with non-gendered judgments like attractiveness, dominance, trust.

We saw a strong negative correlation between the androgyny rating of a base image and CI agreement. CI agreement measured in Experiment 2 presumably depends on the difference between a CI and anti-CI, as similar pairs would be hard to consistently distinguish and more salient pairs would be easier. The current results suggest that a more androgynous base face is more difficult to make even more androgynous, resulting in little meaningful difference between the CI and anti-CI. Conversely, faces rated as less androgynous could be made *more* androgynous through the reverse correlation study and therefore the CIs and anti-CIs would be more distinguishable from each other. This result could be useful for future studies in determining which face stimuli will produce the most distinctive CIs. Conclusions about the male and female conditions may be less meaningful given that agreement was very high overall for both tasks.

3.4 Experiment 4

3.4.1 Introduction

In Experiment 2 we were able to see that the CIs we produced were generally distinct and representative of the intended category. However, as we could not find any significant clusters in Experiment 1, we are still not able to associate specific features with androgyny. This is important to understanding how to apply our results to future studies and the results of Experiment 4 could contribute to creating ecologically valid androgynous faces in future studies. In Experiment 4, we asked participants to view CI and anti-CI pairs (added to their respective base face) and make judgments about specific face parts. The goal of this was to establish which face areas are associated with androgyny and binary gender. Further, it will allow us to understand if there are certain fundamental differences between androgyny and gender.

3.4.2 Methods

Observers and Ethics

Two hundred seventy participants (128 female, 2 unknown; 18-67 years old, $M = 31.8$, $SD = 11.4$) were recruited through Prolific. All participants indicated that they were fluent in English prior to participating. Participants were compensated £7.50 per hour for participating. The median completion time was 12.7 minutes. Pavlovia users who participated in previous experiments were prevented from participating so as to ensure all participants were naive to the stimuli. All protocols were approved by the McMaster Research Ethics Board.

Stimuli and Experimental Design

Participants were shown CI and anti-CI pairs and asked to judge face parts based on varying gender judgments (depending on condition). This was a 3x3x6 task where there were 3 levels of CI (androgynous pairs, female pairs, and male pairs, generated from Experiment 1), 3 levels of gender judgment (androgyny, femininity, masculinity), and 6 levels of face area (forehead, eyebrows, eyes, nose, mouth, and jaw) Each participant performed one judgement (androgyny, femininity, masculinity) on one set of CI stimuli. Participants were assigned randomly to one of nine conditions with the constraint that there were 30 participants per group. Face part was a within-subject variable: the questions about different parts were blocked and presented in a random order. Each block was shown twice for a total of 120 trials (10 face pairs x 6 levels x 2).

Participants were shown a CI and anti-CI pair (as all blocks were completed twice, the CI was shown on the left once and on the right once, similarly to Experiments 2) and asked to rate the androgyny, femininity, or masculinity about a specific face part. Judgments were phrased as being relative to the other face. For example, participants judging the jaw in the femininity condition were shown a response scale that had the options 1) Left has a much more female jaw; 2) Left has a slightly more female jaw; 3) They are equal; 4) Right has a slightly more female jaw; and 5) Right has a much more female jaw. New instructions were shown before each block to ensure participants were clear on what they were being asked to judge.

The ratings were coded as integers ranging from -2 to 2, and were transformed such that positive values indicated that the participant rated the CI (rather than the anti-CI) as having the face part that was more androgynous, masculine, or feminine. The two responses for each face identity and face part were averaged to create a single score.

As there were no right or wrong answers for this task, it is likely that some participants were not paying full attention to the task. To discourage participants from clicking rapidly through the experiment, we designed the experiment such that participants had to click on a judgment and then click on a submit button to move to the next trial. Although it is still possible that participants were providing low quality data, we were not concerned about this as a whole given that in Experiments 1 and 2, we were able to see an effect in our composite classification images. If a sizeable portion of participants were providing nonsense data, we would not see such a strong effect.

3.4.3 Results

Androgynous CIs

Judgments of masculinity, femininity, and androgyny of various face parts for androgynous CIs were evaluated with separate 2-tailed t tests. Judgments of masculinity and femininity did not differ from zero for any face part ($p \geq 0.05$ in all cases). In the androgyny condition there were three face areas that were significantly more androgynous in the CI than the anti-CI: the eyebrows, eyes, and jaw (all $p<0.05$). All other areas did not differ significantly from 0 (Figure 3.7A). Removing the face identities which had low agreement scores in Experiment 2 (i.e., faces F2, F3, F4, & M4 in Figure 3.1) did not change the results (Figure 3.7B).

Ratings for androgynous CIs and anti-CIs, sorted by androgyny ratings obtained in Experiment 3, are shown in Figure 3.8. The specific order of face identities is sorted by the baseline androgyny rating. We can see that there is a great deal of variation between base faces. Specifically, there seems to be more variation for faces that were rated as less androgynous to begin with. We can also see that in most cases the gendered ratings align with the perceived gender of the base face, i.e. the male faces have more feminine CIs and more masculine anti-CIs, and vice versa for the female faces. The least androgynous base face, M3, had its androgyny CI rated as having a feminine jaw, mouth, nose, and eyes, with the anti-CI being more masculine for all of those parts. For F1, the next least androgynous CI, observers rate the androgyny CI as having a masculine jaw, mouth, nose, eyes, and eyebrows and the anti-CI as being more feminine for all of those parts. These results suggest that, starting with these strongly gendered base faces, changing any or all of the face parts in a way that makes them less gendered resulted in a face that was perceived as more androgynous. The results obtained with faces M1, M2 suggest that changes to the forehead can also contribute to androgyny. Hence, these

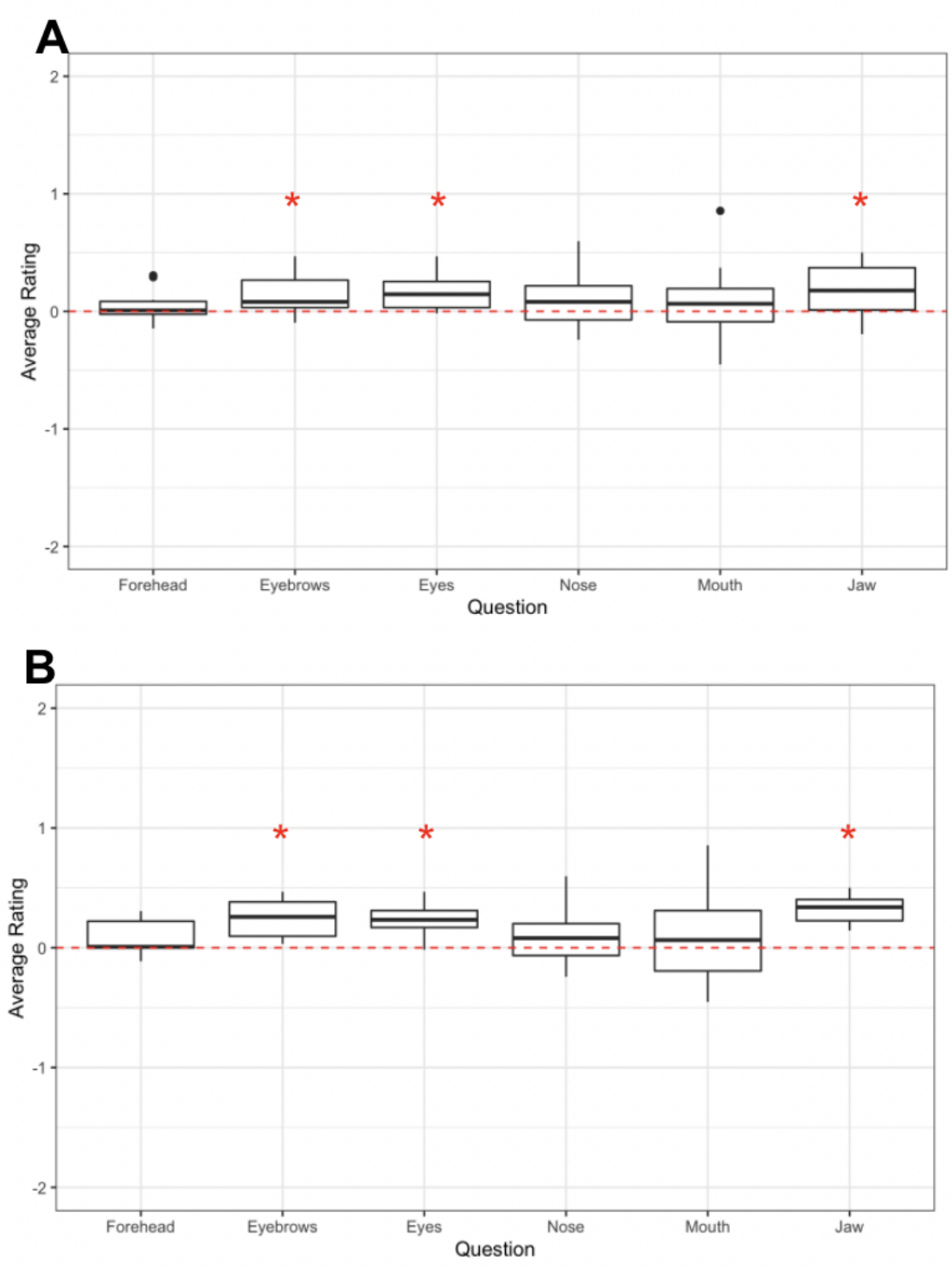


FIGURE 3.7: Androgyny ratings of androgynous CIs and anti-CIs, with (A) and without (B) the CIs that scored at-chance on androgyny judgments in Experiment 2. Positive scores indicate that the CI was more androgynous and negative scored indicate that the anti-CI was more androgynous.

results suggest that perceived androgyny was influenced by many faces parts and that the influence of specific parts varied across faces.

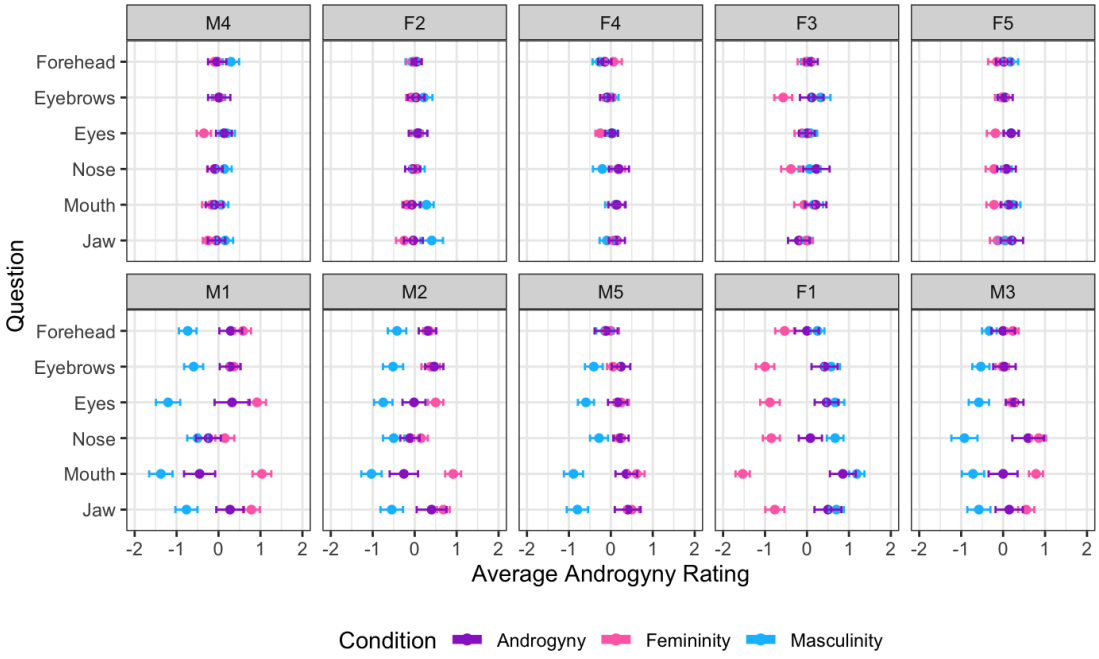


FIGURE 3.8: Ratings of androgynous CIs across all judgment for each individual identity. Identities are sorted by the androgyny ratings from Experiment 3, descending upper left to lower right.

Male CIs

In the male CI pairs, shown in Figure 3.9, all areas of the CI were rated as significantly more male than the anti-CI and all areas of the anti-CI were rated as significantly more female than the CI (all $p<0.05$). No areas significantly differed from 0 for the androgyny ratings. Similarly to the androgynous faces, we see that as masculinity increases, so does the range of values.

Female CIs

In the gendered judgments, the female condition showed similar results to the male condition: all areas of the CI were rated as significantly more female than the anti-CI and all areas of the anti-CI were rated as significantly more male than in the CI (all $p<0.05$) (Figure 3.10. However, there were areas of the female CI that were rated as significantly ($p<0.05$) more androgynous than the anti-CI: the eyebrows, forehead, and nose. No parts were rated as significantly more androgynous in the anti-CI.

FIGURE 3.9: Ratings of male CIS across all judgments for each individual identity. Identities are sorted by the masculinity ratings from Experiment 3, descending upper left to lower right.

The visual relationship between range and base rating seen in the androgynous and male CIs does not appear to persist in the female CIs Figure 3.10.

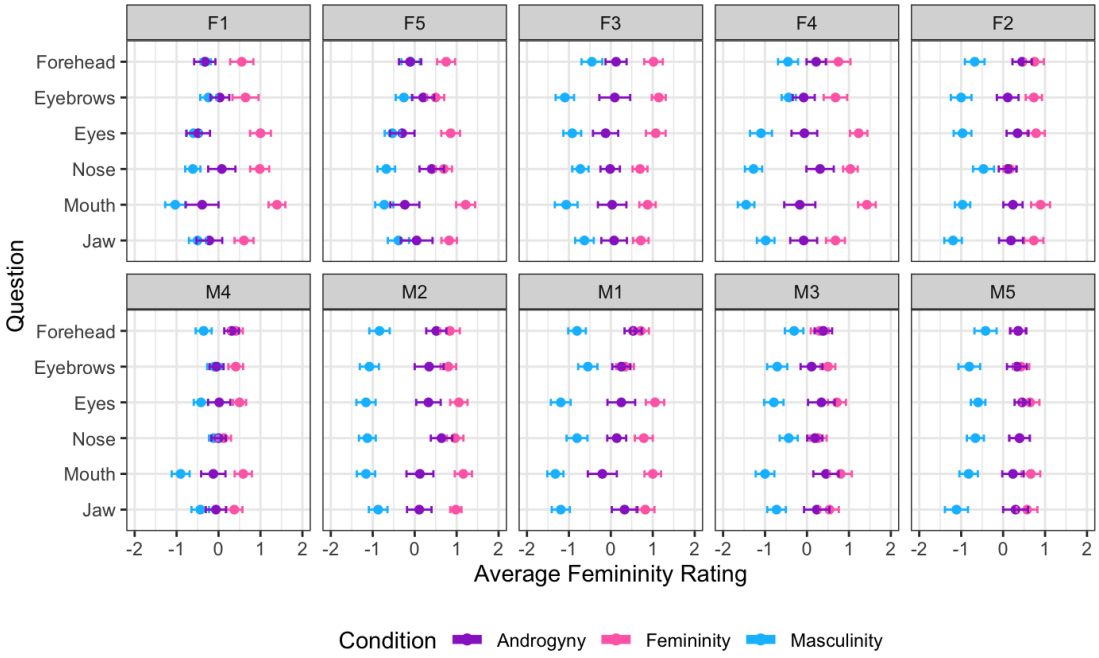


FIGURE 3.10: Ratings of female CIS across all judgments for each individual identity. Identities are sorted by the femininity ratings from Experiment 3, descending upper left to lower right.

3.4.4 Discussion

With this study we were able to establish which face areas are associated with judgments of androgyny, masculinity, and femininity. Whereas in the female and male conditions all areas were significantly different from the anti-CI, for the androgynous condition the eyes, eyebrows, and jaw seemed to be the only areas that were significantly transformed by the reverse correlation procedure. Given that the male and female CIs produced significant effects for all areas, the unique results for the androgynous CIs is more likely attributed to a genuine difference in perception and not a feature of the reverse correlation task. It may be that judging androgyny is a qualitatively different process than judging binary genders.

Further, when judging androgyny in male and female CIs, only the female CIs had any significantly androgynous areas. It is possible that there was something specific about the female faces used in this chapter, however, when these results are taken together with

Chapter 1, which used different face stimuli, we see a consistent tendency of participants to see androgyny more easily in female-identified faces than in male-identified faces.

3.5 Discussion

In this set of experiments we set out to visualize and analyze data-driven androgynous and gendered faces. Using a reverse correlation approach, we created classification images from just 300 trials. Although traditional quantitative analysis could not establish any effect, follow-up behavioural testing showed that most CIs were easily perceived as the intended category, showing that our reverse correlation technique worked appropriately, even when done as an online experiment with fewer degrees of control over the testing environment. When visualizing androgyny we found that several CIs were not discernible from the anti-CI. We conducted further testing and found that there was a significant effect of base androgyny rating: faces that were very androgynous to begin with were difficult to make more androgynous through the reverse correlation process, and the resulting CIs were hard to categorize. Unlike masculinity and femininity which have widely understood social opposites, a more androgynous face may be difficult to make "less" androgynous through a reverse correlation approach, as there is no single understanding of what an anti-androgynous face might look like. In Experiment 4, we sought to evaluate how face parts were perceived in the CIs we created. We found that in the gendered conditions, all face parts were strongly indicative of the relevant gender category, but when judging androgynous CIs, only some of the tested face parts were associated with androgyny. This finding might imply that androgynous faces are perceived in a unique way that cannot be summarized by how masculine or feminine the face is.

References

Brown, E. and Perrett, D. I. (1993). What gives a face its gender? *Perception*, 22(7):829–840.

Buchala, S., Davey, N., Gale, T. M., and Frank, R. (2005). Principal component analysis of gender, ethnicity, age and identity of face images. *Procs of IEEE ICMI 2005:*.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., and Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of vision*, 5(9):1–1.

Creighton, S. E., Bennett, P. J., and Sekuler, A. B. (2019). Classification images characterize age-related deficits in face discrimination. *Vision Research*, 157:97–104.

Dotsch, R. and Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, 3(5):562–571.

Gold, J. M., Sekuler, A. B., and Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cognitive Science*, 28(2):167–207.

Hess, U., Adams, R. B., Grammer, K., and Kleck, R. E. (2009). Face gender and emotion expression: Are angry women more like men? *Journal of Vision*, 9(12):19–19.

Hoss, R. A., Ramsey, J. L., Griffin, A. M., and Langlois, J. H. (2005). The role of facial attractiveness and facial masculinity/femininity in sex classification of faces. *Perception*, 34(12):1459–1474.

Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., and Banissy, M. J. (2015). Dominant voices and attractive faces: The contribution of visual and auditory information to integrated person impressions. *Journal of Nonverbal Behavior*, 39(4):355–370.

Sekuler, A. B., Gaspar, C. M., Gold, J. M., and Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology*, 14(5):391–396.

Zhao, N., Zhou, M., Shi, Y., and Zhang, J. (2015). Face attractiveness in building trust: Evidence from measurement of implicit and explicit responses. *Social Behavior and Personality: an international journal*, 43(5):855–866.

Chapter 4

General Discussion

This thesis sought to evaluate the assumptions that are often made about androgynous faces in perceptual research. Some assumptions were validated by our data collection, whereas other results could not be sufficiently explained by these assumptions, implying that these ideas were overly simple to begin with. The data presented in the current thesis will allow for more nuance in the area of androgynous face research.

4.1 Summary of key findings

Our experiments examined two key assumptions: First, that androgynous faces are equally masculine and feminine, and second, that morphed androgynous faces are a good substitute for natural androgynous faces. These assumptions were derived from the most common methodology for creating androgynous face stimuli: morphing together masculine and feminine faces (e.g. Webster et al., 2004; Kloth et al., 2010). In Chapter 2, we found that masculinity and femininity were strongly negatively correlated (replicated in Chapter 3) and that naturally androgynous faces were generally not rated as strongly masculine or feminine, which aligns with prior assumptions. However, when morphed stimuli were added, we saw that the two types of androgynous faces were not similar in their femininity ratings despite being similar in androgyny ratings. This finding shows that a) having equal masculinity and femininity is not a sufficient predictor of an androgynous face, and b) that morphed and natural faces are not always perceived in the same way. The former conclusion is supported by the documented mismatch between androgynous faces and their perception in Freeman et al. (2010), which found that linearly changing androgynous morphs were perceived non-linearly. The latter conclusion has been cross-validated by Ma's 2022 study on morphed and natural multiracial faces.

Given that both assumptions tested in Chapter 2 could not be validated, the work in Chapter 3 focused on studying androgynous face perception while minimizing assumptions. We chose to do this using a reverse correlation approach based on that of Dotsch and Todorov (2012). In Experiments 1 and 2, we found that participants were able to create distinct classification images (CIs) of masculine, feminine, and androgynous faces. Naive participants were able to discern the CI from the anti-CI in most cases. However, not all base faces produced distinct CIs: 4/10 androgynous CIs performed at chance in Experiment 2. In Experiment 3, we found that this effect could be explained by the androgyny rating of the base face: the base faces which were rated higher in androgyny were the same faces that were difficult for naive participants to distinguish when viewing androgynous CIs/anti-CIs. Further, we found that androgyny was moderately to strongly negatively correlated with attractiveness, dominance, and trustworthiness. Attractiveness has been previously linked to androgyny: Yang et al. (2015) found that participants preferred gendered to (morphed) androgynous faces even when they were engineered to be equally attractive, regardless of sexual preference. Conversely, Hester et al. (2020) found that male faces which were high in masculinity and femininity were rated as highly attractive and trustworthy. However, it is difficult to directly compare these results as this study did not use any androgynous face stimuli. It is possible for all these results to exist simultaneously: the effect of having masculine and feminine traits might be socially beneficial for male-coded faces but does not persist when the gender of the face is female-coded or cannot be coded as binary male or female.

In Experiment 4, we found that all face parts tested (forehead, eyebrows, eyes, nose, mouth, and jaw) were associated with masculinity and femininity in gendered CIs, whereas only some areas (eyebrows, eyes, and jaw) were associated with androgyny in androgynous CIs. Further, even with the at-chance faces removed, the androgynous CIs had lower agreement (Experiment 2) than the male or female CIs. This result suggests that androgyny and/or androgynous faces might be perceptually distinct from masculinity and femininity, as opposed to being a linear combination of the two. If the latter were the case, we might expect all tested face areas to be significantly more androgynous for the androgynous CIs. The results could be also explained by the overall lower agreement for androgynous faces: there were fewer significant diagnostic regions for androgynous CIs because participants were overall less sure. An alternate explanation for our findings might be that the androgynous CIs were perceptually distinct for a reason other than their androgyny, and the difference in significant face parts could be attributed to this third variable. Without further testing, it is not possible to conclusively say which is the case. However, with the information available, it seems more plausible that the difference

has to do with the lower agreement, which can be interpreted as lower distinctiveness, of the androgynous CIs.

4.2 Remaining questions

The current thesis did not thoroughly investigate which non-gendered factors might contribute to facial androgyny. Most of our work focused on relating androgyny to femininity and masculinity, but we concluded that these two judgments are insufficient for predicting androgyny, meaning that there are other unstudied factors which could be added to the model of androgyny. As discussed, social factors such as attractiveness could be relevant, but this would also raise the question of directionality; it is more likely that androgynous faces were judged as less attractive, trustworthy, and dominant because of their androgyny, and not the other way around. A good model of androgyny would isolate the factors which contribute to androgyny judgments, not simply anything that is correlated with androgyny. Hester et al. (2020) propose a model of gender perception that, while not specifically mentioning androgyny, could serve as a jumping off point for future work. This model proposes that facial traits (bottom-up) and societal beliefs (top-down) combine to lead to a gender judgment. They also propose breaking up the classic male-female continuum (on which androgyny would fall at the midpoint) into two dimensions: masculinity and femininity. Our findings established support for a version of the original continuum, as masculinity and femininity were strongly negatively correlated. Although androgyny fell somewhat towards the centre, this one dimension was not sufficient in predicting androgynous faces. Because of this, based on the current thesis, there may be one or more additional dimensions that contribute to androgynous face perception. To fully flesh out a model of gendered and androgynous face perception, more research will need to be done.

4.3 Conclusion

The current thesis investigated both established assumptions of androgynous faces and a data-driven approach of visualizing androgynous faces. The findings suggest that a revised conceptualization of androgyny is required to accurately reflect the nuances of androgynous faces. Given that androgynous faces are often used as a means to an end rather than a topic of interest, it is necessary for researchers to understand the tools they are using to investigate other effects. The current thesis contributes a deepened understanding of androgynous faces, which in turn will improve the quality of both

gender-related studies as well as studies in other topics that use androgynous faces as an investigative tool.

References

Dotsch, R. and Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, 3(5):562–571.

Freeman, J. B., Rule, N. O., Adams Jr, R. B., and Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, 20(6):1314–1322.

Hester, N., Jones, B. C., and Hehman, E. (2020). Perceived femininity and masculinity contribute independently to facial impressions. *Journal of Experimental Psychology: General*.

Kloth, N., Schweinberger, S. R., and Kovács, G. (2010). Neural correlates of generic versus gender-specific face adaptation. *Journal of Cognitive Neuroscience*, 22(10):2345–2356.

Webster, M. A., Kaping, D., Mizokami, Y., and Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, 428(6982):557–561.

Yang, T., Chen, H., Hu, Y., Zheng, Y., and Wang, W. (2015). Preferences for sexual dimorphism on attractiveness levels: An eye-tracking study. *Personality and Individual Differences*, 77:179–185.