REPRESENTATION OF FAMILIARITY IN THE FACE SPACE

A Thesis Proposal

Submitted to the Faculty of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Psychology

University of Regina

By

Yaren Koca

Regina, Saskatchewan

February 22, 2024

Table of Contents

List of Figures

Abstract

The visual system can successfully distinguish thousands of faces despite the fact that all faces share the same overall configuration. The face space theory suggests the way the visual system overcomes this homogeneity problem is by encoding faces in a multidimensional space based on their perceived characteristics. The face space is also able to unify many interesting phenomena in the literature pertaining to how recognition performance for faces may differ based on their orientation, distinctiveness, their membership to an other-group, and how caricatured they are. Recent literature shows that face familiarity is another important factor that creates remarkable recognition and perceptual differences in processing faces. Given the idiosyncratic nature of within-person variability under different viewing conditions, a curious problem is how the visual system can successfully achieve amalgamation of different views of the same individual into a single representation. The aim of the current research program is to understand how the face space can accommodate face familiarity. The first proposed experiment aims investigate how the face space would represent multiple ambient photos of familiar and unfamiliar faces. The second proposed experiment aims to examine how the face space changes as a face becomes familiar. The third proposed experiment explores a norm-based model of identity-specific subspaces. Together, I hope to reconcile these two aspects of face processing (differentiation and amalgamation) into a unified model of face recognition that updates certain assumptions of the face space theory based on the findings of the more recent literature.

1 Introduction

> You're so exactly like other people... the two eyes, so (marking their places in the air with his thumb) nose in the middle, mouth under. It's always the same. Now if you had the two eyes on the same side of the nose, for instance—or the mouth at the top—that would be some help. (Carroll, 1946)

As quoted in Rhodes (1996), Humpty-Dumpty's observation above highlights a remarkable capability of the visual system. All faces share the same configuration; they have the same basic components (i.e., two eyes, a nose, and a mouth) arranged in the same way. Yet, we are able to discriminate thousands of faces from one another. In fact, Jenkins et al. (2018) calculated that on average, a person knows about 5000 faces. That is 5000 faces that the visual system can individuate. Just as fascinating is our ability to recognize a familiar face across changes in viewpoint and context, such as lighting, expression, age, and makeup. How does the visual system overcome the homogeneity problem while being able to tolerate the variability across different views of a familiar face? We must assume that any model of face recognition should account for organizing faces in memory in a way that can serve to discriminate faces from one another, while encoding identity-specific information that can serve to recognize familiar individuals across several contexts. In this dissertation, I will propose a model of face recognition that reconciles a well-founded theory of discriminating faces with relatively newer findings in the literature concerning the importance of face familiarity.

The introduction section is organized as follows. First, the face space theory will be reviewed with findings in the literature that the face space theory serves to unify. Namely, I will discuss the inversion effect, the distinctiveness effect, the caricature effect, the other-race effect, and face-specific after-effects. These phenomena all provide insights into the architecture of face representations in memory that provide an account of how faces are discriminated. Second, I will argue that the face space theory has some theoretical and methodological limitations that have been discussed in the more recent literature. These limitations emphasize the importance of face familiarity, with the operational definition of face familiarity as tolerance to within-person variability, that is, our ability to recognize a familiar face across several views despite the fact that faces highly vary in unsystematic ways.

1.1 The Face Space: Telling Faces Apart

Models of representation, classification, and recognition of stimuli represented in a multidimensional space have been advanced for non-face stimuli (e.g., Nosofsky, 1986; see Valentine, 2001). In the early 90s, the face space was proposed as a theory for representing faces in a multidimensional representational space where faces are encoded according to their perceived characteristics (Valentine, 1991). According to the face space theory, each face corresponds to a location in the face space. It is unclear how many dimensions there are in the face space, and what constitutes those dimensions, but they serve to discriminate faces from one another. The face space has been widely accepted as a unifying concept that could account for several findings in the literature at the time, namely the inversion effect, the distinctiveness effect, the caricature effect, and the other-race effect (Valentine, 1991; Valentine et al., 2016).

An important property of the face space is that the origin of the face space is the central tendency of these dimensions, and previously encountered faces are normally distributed around the origin. Valentine proposed two models in which faces are encoded in the face space: the norm-based model and the exemplar-based model. The main distinction between these two models is the role of a prototype. The norm-based model suggests that an abstracted norm, or a "prototype" sits at the origin of the face space. This prototype is an idealized representation of all previously-encountered faces and does not correspond to any of the individual instances. Importantly, this model supposes that faces are encoded based on the relative distance from the prototype, so each face is an n-dimensional vector in an n-dimensional face space (Figure *1.1*A). In other words, each face is encoded based on how it differs from the prototypical face, which in turn helps overcome the homogeneity problem by representing faces as the way in which they deviate from a prototype face. According to the norm-based model, there are two steps that are assumed to be involved in the recognition of a face: (a) a stimulus is encoded as an n-dimensional vector, and (b) some decision process is used to determine whether or not this vector matches with the vector of a familiar face. This decision process will depend on the error associated with the vector, (i.e., poor viewing conditions will allow for more error) and a measure of similarity between the new vector and the nearest familiar face's vector. Although it is not known how this similarity is calculated, it is implicitly assumed to be a Euclidean distance.

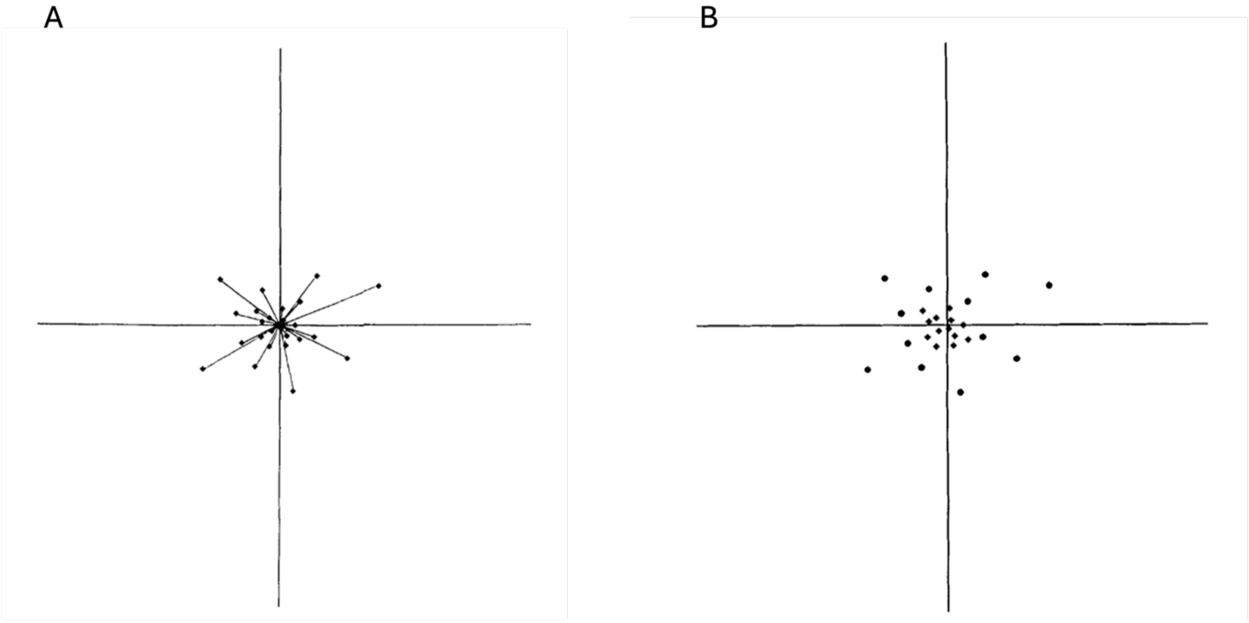

Figure 1.1. (A) A two-dimensional norm-based model of the face space taken from Valentine (1991). Each encoded face is represented as a vector. At the origin is the prototype face, which is assumed to be a central tendency of all faces. (B) A two-dimensional illustration of the exemplar-based model taken from Valentine and Endo (1992). In this model, each face is encoded as an absolute point as opposed the vector. There is no prototype at the origin, but since faces are assumed to be normally distributed, they are clustered more densely around the central tendency.

The exemplar-based model, on the other hand, does not necessitate the abstraction of a prototype and suggests that faces are located as individual points in the face space (Figure 1.1B).

The origin of the face space, then, is only where the highest density of faces is found and does not include a prototype that plays a role in encoding faces. The similarity between two points (i.e., faces) in the face space is simply a monotonic function of the distance between them. The decision process, then, will depend on the estimate of error associated with encoding (similar to the norm-based model), and the distance between the location of the encountered stimulus and the nearest familiar face.

Valentine (2001) distinguishes the norm-based model from the prototype models of semantic categorization in that the norm-based model still assumes that exemplars are used in recognition, it is only the way faces are encoded that differs from the exemplar-based model. Therefore, the exemplar-based models and the norm-based models often make similar predictions regarding how faces are discriminated, but they offer different underlying mechanisms for several phenomena in the literature. In the following sections, phenomena that are unified within the face space theory will be described, and their implications for the face space theory along with their relationship with the norm-based and the exemplar-based models will be discussed.

1.1.1 The Inversion Effect: Looking at Upside-Down Faces

Faces, compared to other classes of mono-oriented (i.e., viewed in one orientation) visual stimuli, suffer disproportionately from inversion (e.g., Diamond & Carey, 1986; Fraire et al., 2000; Valentine, 1988; Yin, 1969). Yin (1969) conducted a study where subjects studied images of faces, houses, airplanes, and stick figure men in motion. Following the study phase, subjects were given a two alternative force choice (2AFC) recognition test where they were asked to indicate which image appeared in the study list. For some stimuli, the orientation of the study and the test images matched (i.e., upright-upright, inverted-inverted), and for others the orientation of the study and the test images mismatched (i.e., upright-inverted, inverted-upright).

In all conditions that involved inversion, recognition performance was worse, but importantly, inversion was disproportionately detrimental to face recognition performance, despite all the image categories being mono-oriented. These results revealed that although all mono-oriented objects suffer from inversion, faces suffer from inversion even more.

Yin (1969) argued that the disproportionate inversion effect of faces indicates that faces are a special class of visual objects. This speculation raised the question regarding what property or properties of faces make them special and also vulnerable to the inversion effect. The most prevalent hypothesis is that faces suffer from inversion because they are processed holistically (e.g., Young et al., 1987) – all faces share the same configuration, so the ability to discriminate faces must rely on detecting subtle differences in the overall configuration of faces. This cannot be accomplished if faces are decomposed into parts, since faces with different configurations appear as different identities even if the individual parts are held constant (Figure 1.2). Inversion is assumed to disrupt holistic processing (Diamond & Carey, 1986), which makes faces more difficult to recognize when inverted.

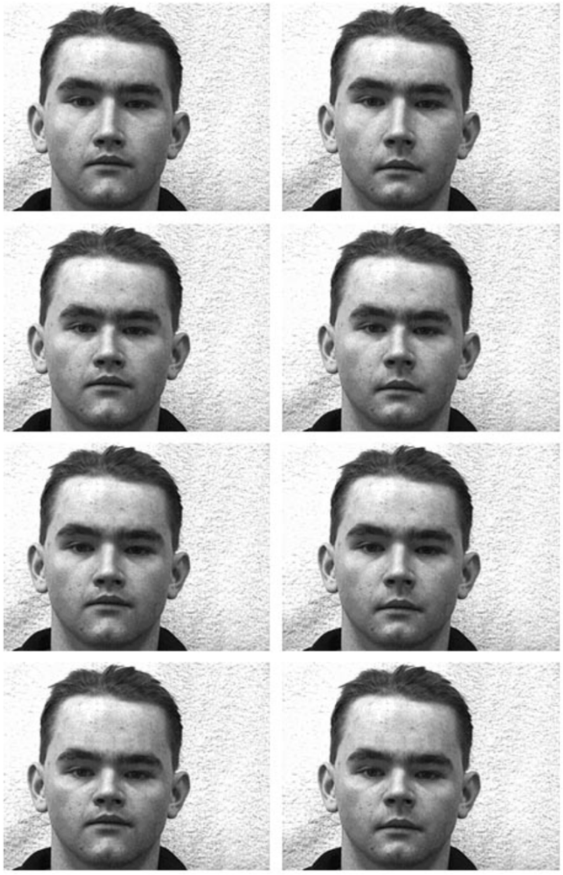

Figure 1.2. Faces that share different configuration with features constant, taken from Freire et al. (2000).

Evidence for holistic processing of faces came from Young et al.'s (1987) composite face illusion. When the top half and the bottom half of two different identities are used to create a composite face, the perception of a third identity emerges (Figure 1.3). Moreover, when subjects are shown a composite face and are asked to identify one of the constituent parts, recognition slows down because the top and the bottom halves interfere with each other. This impairment in recognition is eliminated when the top and the bottom halves are misaligned or when the faces are inverted, suggesting that faces are processed in a holistic manner, and inversion interferes with holistic processing.

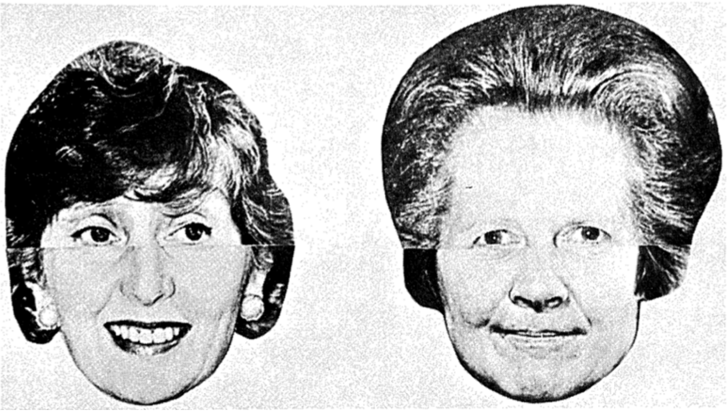

Figure 1.3. A composite face illusion created using Margaret Thatcher and Shirley Williams, taken from Young et al. (1987).

Farah et al. (1995) investigated whether holistic processing in general is impaired by inversion. In their experiment, they created dot patterns that only differed in subtle configuration. Subjects studied these dot patterns, some subjects saw mono-coloured dot patterns (i.e., holistic processing) and some saw dot patterns with multiple colours to encourage part decomposition by utilizing the Gestalt similarity principle. During the recognition test, these dot patterns were shown either upright or inverted. As predicted, recognition of mono-coloured dot patterns was disrupted by inversion to a greater extent than multi-coloured dot patterns. In another experiment, the authors applied this to faces by either presenting whole faces or faces decomposed into parts. When recognition was tested using whole faces (upright or inverted) they found that subjects who studied whole faces suffered from inversion more than the ones who studied parts of faces. These results suggest that inversion disrupts holistic processing in general, and since faces are processed holistically, they are particularly vulnerable to inversion.

Freire et al. (2000) provided additional evidence that inversion disrupts holistic processing of faces while also investigating whether the inversion effect is a perceptual phenomenon. In their study, they used faces that contained the same parts while varying configuration (Figure 1.2). Subjects were given a matching task where a pair of images were presented side by side and subjects were asked to indicate whether the pair of images depicted the same person. They found that performance was impaired by inversion. In another study, they employed the same matching task using faces with the same configuration but different parts by replacing individual features of the face with someone else's. This time inversion did not disrupt matching performance, suggesting that inversion impacts perception of faces as well as memory for faces by disrupting holistic processing.

Diamond and Carey (1986) examined whether there are other classes of visual stimuli that use holistic processing. They argued that expertise would mediate the ability to detect subtle differences in configuration (called second-order configuration), so any visual object would suffer from inversion by experts of that visual object. To test this prediction, they gathered dog experts and novices and gave them a recognition task using dog images. They found that dog experts' recognition of dogs was impaired by inversion to a larger extent than novices' recognition of inverted dogs. The authors suggested that expertise in any visual object class that shares the same overall configuration (e.g., two eyes above a nose above a mouth) suffers from inversion, as long as they differ based on their second-order configurational features. Therefore, they argued that faces are not inherently special, but rather our expertise of faces makes them a special class of stimuli.

1.1.1.1 Implications for the Face Space. The face space is agnostic towards the inversion effect (Valentine, 2016). Goldstein and Chance (1980; Valentine, 1988) argue that a face prototype develops with extensive experience to upright, own-race faces. This face prototype improves the ability to recognize upright own-race faces, but it comes at a cost of inflexibility in recognizing faces with transformations such as inversion or faces that are of another race. This prototype rigidity is similar to the face prototype proposed by the norm-based model. According to the norm-based model, the first step of recognition involves encoding the face as an n-dimensional vector, and the second step involves a decision process to determine whether the face is familiar. A factor that contributes to this decision process is the error associated with encoding the stimulus, which will be greater with more difficult viewing conditions. The exemplar-based account makes a similar assumption, where a face is first encoded as a point in the n-dimensional space, and the decision process is influenced by the error associated with encoding the stimulus. Inversion, then, is simply a transformation in the stimulus that increases the encoding error. Therefore, both models make a similar prediction that inverting a face will make it more difficult to recognize. Importantly, both models also predict that the effect of inversion will be greater for typical faces than for distinctive faces, because the increased error in encoding an inverted typical face will make it more likely to activate a nearby representation. This prediction was confirmed in Valentine (1991) who found that inversion caused a greater impairment in recognizing typical faces compared to distinctive faces.

1.1.2 The Distinctiveness Effect: The Role of Inter-Item Similarity

There is superior recognition performance observed for distinctive faces compared to typical faces (e.g., Bartlett et al., 1984; Cohen & Carr, 1975; Johnston et al., 1997; Shepherd et al., 1991; Valentine & Bruce, 1986a; 1986b). For example, Going and Read (1974) asked subjects to rate unfamiliar faces on “uniqueness” (compared to the general population; i.e., absolute distinctiveness), and later found recognition accuracy to be higher for the faces that were rated high in uniqueness compared to low in uniqueness. Cohen and Carr (1975) replicated this effect by asking subjects to rank faces based on distinctiveness (within the stimulus set; i.e., relative distinctiveness) and found lower misses and false alarms for recognizing distinctive faces. Light et al., (1979) found that “unusual” faces were more accurately and confidently recognized under both incidental and intentional learning conditions, and with presentation times varying between three seconds to 15 seconds per item. Using personally familiar and famous faces, Valentine and Bruce (1986a; 1986b) found that distinctive faces were accepted as familiar faster than typical faces. Shepherd et al (1991) also found that distinctive faces were recognized with higher hit rates, higher sensitivity (*d'*), and shorter reaction time.

It is challenging to define distinctiveness for faces compared to other classes of visual stimuli for which distinctiveness can be described on a single dimension (e.g., color; Valentine, 2001). Nevertheless, the effect seems to be robust regardless of how distinctiveness is operationalized in the literature. Some researchers adapted terms like "atypical" (e.g., Light et al., 1979; Vokey & Read, 1992), "unusual" (e.g., Bartlett et al., 1984), "unique" (e.g., Going & Read, 1992), or "distinctive" (e.g., Valentine & Bruce, 1986a; 1986b). Commonly, distinctiveness is measured through asking subjects to rate the faces based on how easy it is to spot them in a crowd (e.g., at a busy railroad station; Valentine & Bruce, 1986a; 1986b; Valentine & Endo, 1992), rate the faces that look more similar to a typical high school male student (Light et al., 1979), or simply asking them to rate or rank faces for distinctiveness (e.g., Cohen & Carr, 1975). Some present subjects with pairs of faces and ask them to rate the similarity of the pair (e.g., Johnston et al., 1997; Light et al., 1979), so items that are rated as most dissimilar are described as most distinctive. Although these ratings are subjective, they have high inter-rater agreement (Cohen & Carr, 1975; Valentine, 2001). In fact, Valentine and Endo (1992) found cross-cultural agreement on distinctiveness. British and Japanese subjects rated the distinctiveness of British and Japanese faces, and there was high agreement about which faces were more distinctive irrespective of the subjects' and the faces' race. Subjective ratings can also be predicted by more objective measures of distinctiveness. For example, Bruce et al. (1994) gathered eccentricity measures by computing distances between facial features (e.g., nose length, the beakiness of the nose) and found that subjective ratings of distinctiveness correlated with the eccentricity measures. Therefore, although it is difficult to pin down what exactly constitutes distinctiveness for faces, different ways of measuring and describing distinctiveness seem to refer to the same psychological construct of typicality.

What accounts for the distinctiveness effect? Light et al., (1979) investigated the possibility that the distinctiveness effect was mediated by greater attention being paid to distinctive faces, in line with the depth of processing account proposed by Craik and Tulving (1975). Accordingly, they presented subjects with distinctive and typical faces and asked them to judge their likeability (elaborate processing) or gaze direction (shallow processing) under both incidental and intentional learning conditions. They hypothesized that if the distinctiveness effect is mediated by depth of processing, there would be an interaction between typicality and the type of judgment (likeability or gaze direction) in that the recognition advantage for distinctive faces would be reduced when measured in likeability ratings. They did find such an interaction for false alarms, but it fell short of statistical significance and therefore was inconclusive.

Shepherd et al. (1991) explored whether the distinctiveness effect is a result of encoding distinctive features. They hypothesized that shortening the presentation time during study would prevent subjects from encoding distinctive features and in turn would diminish the effect. Although Light et al. (1979) also manipulated presentation durations, they allowed for about three seconds per item, which may have been long enough for distinctive features to be encoded. As such, Shepherd et al. (1991) shortened the presentation time to 1 second per item. They replicated the distinctiveness advantage for hits, false alarms, and reaction times. However, they did not find an interaction between distinctiveness and presentation time, thus, shortening the presentation duration did not reduce the distinctiveness effect. Further, Valentine and Bruce (1986b) presented subjects with distinctive and typical faces either intact or jumbled and asked them to indicate whether it was a face or not (i.e., face classification task). They found that intact typical faces were classified as faces faster than intact distinctive faces. If the distinctiveness effect was based on greater attention being paid to distinctive faces or distinctive features, they would be classified as faces faster than typical faces.

Bartlett et al. (1984) suggested that the distinctiveness effect was solely based on “context-free familiarity” information. They claimed that previously-unseen typical faces elicit a sense of familiarity due to their resemblance with most previously-encountered faces, and this in turn results in more false alarms for typical faces. Therefore, a pre-familiarization process with all faces should eliminate or even reverse the recognition advantage for distinctive faces. One set of subjects rated unfamiliar typical and distinctive faces based on how familiar they felt (e.g., may have seen before on a city street), and found that typical faces indeed received higher familiarity ratings. Another set of subjects were presented with unfamiliar typical and distinctive faces and were asked to rate them on friendliness followed by a recognition test on those faces. Some of these subjects were additionally exposed to all the faces prior to the recognition test (i.e., “familiarized”). They found that familiarizing subjects with the faces prior to the recognition test reduced the recognition advantage for distinctive faces. In other words, exposure to a distinctive face resulted in a higher increment of familiarity than exposure to a typical face. However, this finding was observed only for false positives.

Valentine and Bruce (1986b) further tackled the role of context-free familiarity on the distinctiveness effect by using a different experimental paradigm that looked into response latencies. They argued that if context-free familiarity forms the basis of the distinctiveness effect in that distinctive faces gain more familiarity from each additional exposure, already familiar distinctive faces should feel more familiar than familiar typical faces. This would be reflected as a positive correlation between distinctiveness and familiarity ratings. To test this prediction, they first collected distinctiveness ratings from unfamiliar subjects. They then presented those faces to familiar subjects and gave them a familiarity decision task (i.e., is this face familiar?) while measuring their reaction time. At the end of their experiment, the same familiar subjects were asked to make familiarity judgments on the faces (i.e., which of these faces are you most familiar with?). They found that distinctive familiar faces, and the faces that were rated as most familiar, were classified as familiar faster than typical or less familiar faces. Critically, however, distinctiveness ratings and familiarity ratings were not correlated. In other words, they found that the faces that were most distinctive were not necessarily the most familiar, disconfirming the context-free familiarity hypothesis. This was also demonstrated in Vokey and Read (1988), where general familiarity judgments and the distinctiveness of faces contributed to recognition performance, but familiarity and distinctiveness factors were not related to each other. Therefore, context-free familiarity does not seem to be the only factor underlying the distinctiveness effect.

In Light et al. (1979), inter-item similarity ratings (of pairs of faces) were correlated with distinctiveness ratings, so distinctiveness must have its structural basis in inter-item similarity. They explained this by proposing a two-component (i.e., schematic and specific) memory model. Typical faces activate the prototype face (i.e., schematic face) for faces due to their resemblance with most previously encountered faces. This schematic memory is in turn responsible for increased false alarm rates for typical faces. Distinctive faces, on the other hand, deviate from this schematic representation and thus access a “specific” memory component, resulting in higher hit rates and reduced false alarm rates for distinctive faces. In line with this theory, Vokey and Read (1992) proposed that distinctiveness was composed of two orthogonal components: general familiarity (i.e., context-free familiarity) that results in reduced discrimination for typical faces (increased false alarms), and memorability that results in increased discrimination for distinctive faces (decreased false alarms and increased hits). They asked subjects to provide ratings of attractiveness, typicality, likability, familiarity, and memorability on a set of faces. Applying Principal Components Analysis (PCA) on the ratings, they found that the familiarity and memorability factors completely overlapped with the variance explained by the typicality factor. In another experiment, they gave subjects a recognition test and found that the familiarity component was reflected exclusively in false alarms and not in hits, whereas the memorability component was mostly reflected in hits. Importantly, when familiarity and memorability were factored into the regression, adding typicality did not contribute to the model, suggesting that typicality could be described as a summation of familiarity and memorability.

In summary, distinctive faces are more readily recognized than typical faces. Distinctiveness of a face has its structural basis in the inter-item similarity in the population of face representations (Light et al., 1979), and is reflected by two orthogonal components: familiarity and memorability (Vokey & Read, 1992). The distinctiveness effect reveals a noteworthy property of face representations: similarity of faces plays an important role in encoding and recognizing faces. Typical faces are similar to one another, and they access a prototypical representation of faces. Distinctive faces are dissimilar to this schematic representation; therefore, they access a more specific memory component which in turn facilitates their recognition.

1.1.2.1 Implications for the Face Space. Distinctiveness, then, is a function of inter-item similarity within the population of face representations (Light et al., 1979) and is described by general familiarity (reflected by typical faces) and memorability (reflected by distinctive faces; Vokey & Read, 1992). In line with the face space theory, where faces are also encoded in dimensions that serve to discriminate them, typical faces are clustered near the origin where the highest exemplar density is found (i.e., high general familiarity), and distinctive faces are encoded in more isolated regions where the exemplar density is low (i.e., high memorability). Typical faces, then, have several neighbors, which make them more likely to be confused with one another, resulting in higher false alarm rates. Distinctive faces, on the other hand, are more isolated, which makes them more likely to be accurately recognized and less likely to be mistakenly recognized as having been seen before (lower false alarm rates).

Since the distinctiveness effect is assumed to be a result of exemplar density, both the norm-based and the exemplar-based models make similar predictions about the distinctiveness effect. When a distinctive familiar face is encountered, it falls closer to the representation of the target face than another neighbor, making it faster and more accurate to recognize the face. When a typical familiar face is encountered, it falls close to the target representation but also to a near neighbor, making it slower and less accurate to recognize. When an unfamiliar distinctive face is encountered, it will be encoded in a location where there are not many faces around, decreasing the possibility of false alarms. When a typical unfamiliar face is encountered, it will fall into a dense location, making a near face more likely to be falsely recognized. Exemplar density also provides an account for why typical faces are classified as faces faster (Valentine & Bruce, 1986b). Assuming that the classification task is done by calculating the similarity between the target and the central tendency of the category, when a typical face is encountered, it falls closer to the central tendency, making it faster to classify it as a face.

However, the mechanisms underlying the distinctiveness effect are described differently by the two models. According to the norm-based model, the similarity of two faces that are separated by the same distance depends on the distance from the norm. If two faces are closer to the norm, the angle that separates them is larger, whereas if two faces are further away from the norm, the angle that separates them is smaller (Figure 1.4). This leads to the prediction that two distinctive faces (that are further away from the norm) that are separated by the same distance from two typical faces would be more difficult to discriminate. Since this prediction goes against what would be predicted based on the distinctiveness effect, we can assume that the effect of exemplar density on recognition is larger than the opposing effect of the distance from the norm. Therefore, the exemplar-based model may provide a more parsimonious account for the distinctiveness effect, where the effect is based solely on the exemplar density in the face space.

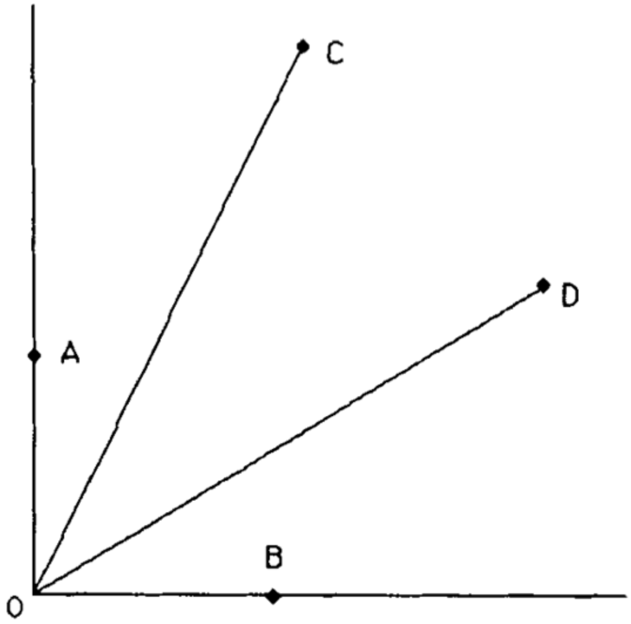

Figure 1.4. An example illustration taken from Valentine (1991). Two sets of faces that are equidistant but are not equally dissimilar. The distance between AB and CD are the same, but the OA and OB are separated by a larger angle than OC and OD.

1.1.3 The Caricature Effect: Not All Distortions Are the Same

Can someone's distorted image be better recognizable than an undistorted one? Caricatures - images where distinctive characteristics of a face are exaggerated - are recognized at least as well as undistorted images, and are better recognized than anti-caricatures, images where distinctive features of a face are de-emphasized. (e.g., Benson & Perret, 1991; Gibling & Bennett, 1994; Mauro & Kubovy, 1992; Rhodes et al., 1987). Caricatures are either artistic, which are created through caricature artist drawings, or computer-made, where the differences between the target's face and an average (i.e., norm) face are enhanced. For example, if a face has smaller than average eyes, they are made even smaller since the caricaturing process increases the difference between the prototype and the target face (Figure 1.5). For computer caricatures, the degree of enhancement is commonly varied in experiments, with greater percentages reflecting greater deviation from the norm. Anti-caricatures, then, are created by *decreasing* the difference between the image and the prototype, making the face look more typical in appearance. Therefore, caricatures (and anti-caricatures) represent a deviation from (or towards) a prototype, similar to distinctiveness of faces, with the important distinction that distinctiveness in caricatures is varied *within* a face instead of across faces (Rhodes et al., 1987).

It is unclear whether caricatures are better recognized than veridical images. For example, Hagen and Perkins (1983) failed to find a caricature advantage for unfamiliar faces. Subjects were shown images of unfamiliar faces to study, then picked out those individuals they have studied from a larger pool of images. These images were either artistic caricatures, three quarter view photographs, or profile photographs. In the same-medium condition the subjects saw the same type of images during study and test (e.g., caricature-caricature, profile-profile). In the different-medium condition they saw different types of images during study and test (e.g., caricature-profile, three quarter-profile). They found that subjects in the same-medium condition outperformed those in the different-medium condition. Within the different-medium condition, they found more false alarms for conditions involving caricatures compared to three quarter view and profile conditions. Hagen and Perkins (1983) argued that the caricature advantage may result from better access to a face's representation in long-term memory. Since unfamiliar faces do not have a stored representation, the caricature advantage could be exclusive to familiar faces.

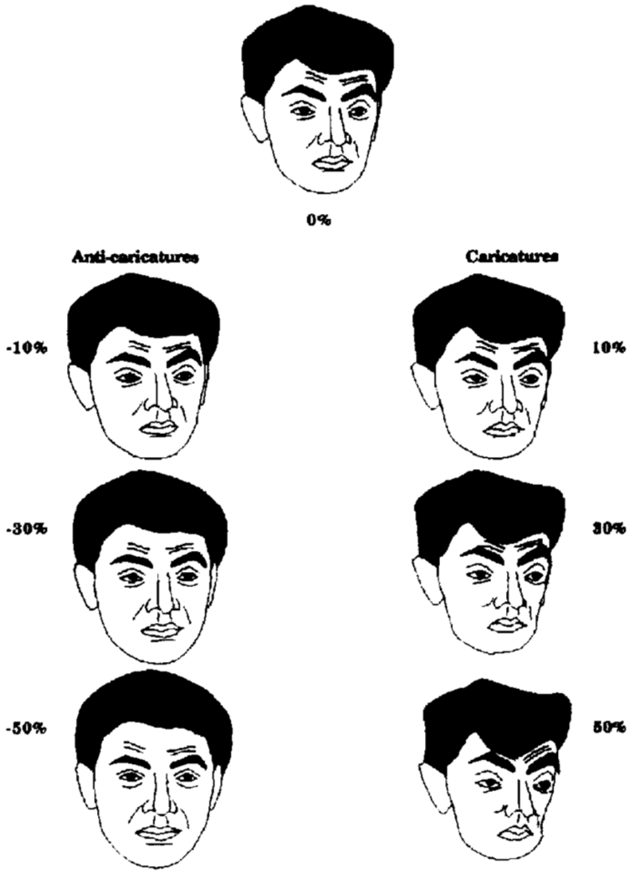

Figure 1.5. A veridical drawing of Rowan Atkinson (0%) and its varying levels of caricatures and anti-caricatures taken from Rhodes & Tremewan (1994).

However, Tversky and Baratz (1985) did not find a caricature advantage using familiar faces. They collected photographs of famous politicians and obtained artistic caricatures of them. Subjects saw either only caricatures, only photographs, or a mix of caricatures and photographs. In one experiment, they were asked to rate the images on goodness of likeness followed by a recall task. In another experiment, they were given a semantic verification task in which an image was paired with a name, and they indicated whether the image matched the name while their reaction time was measured. They found no caricature advantage for goodness of likeness ratings, recall performance, and semantic verification performance – photographs that were rated as better likenesses were recalled more accurately, and were identified more quickly.

One reason why Hagen and Perkins (1983) and Tversky and Baratz (1987) did not find a caricature advantage could be due to the fact that photographs simply contain more information than line drawings, indicating that photographs and line drawings may not yield a fair comparison. To address this, Rhodes et al. (1987) utilized a computer program to generate line drawings of veridical images, caricatures, and anti-caricatures (i.e., where the distinctive features are diminished instead of enhanced, so the image is made closer to the prototype) of familiar faces. In Experiment 1, they asked subjects to identify the people depicted in veridical line drawings, caricatures, and anti-caricatures and collected likeness ratings. They found no difference in accuracy, but caricatures were identified the fastest, followed by veridical line drawings, followed by anti-caricatures. Caricatures were also rated as better likenesses than veridical drawings (by interpolation, the best likeness caricature was at 16%). In Experiment 2, they obtained likeness judgments from unfamiliar viewers by first showing them the original photo upon which the drawings were based and asked them to rate the likeness of the photo to the drawings either by direct comparison to the photo or by asking them to rely on their immediate memory. They found that veridical line drawings were rated as the best likenesses, followed by caricatures and anti-caricatures which did not differ in likeness ratings.

Benson and Perrett (1991) used photographic caricatures instead of line drawings to investigate whether the caricature advantage is limited to line drawings, or whether the advantage could be enhanced by using photographic caricatures. They created caricatures, veridical images (created as a 0% caricature to eliminate bias due to pixel anomalies), and anti-caricatures of familiar faces. In Experiment 1, subjects were shown all images and were asked to provide familiarity ratings, choose the best likeness of the person among the images, and rate the goodness of likeness for the selected photo. They found that the goodness of likeness ratings peaked at the positive caricature (by interpolation, 4.4%), and caricatures were rated as better likenesses than anti-caricatures. When ratings were analyzed for each target (n = 7), they found that three targets received better likeness ratings for caricatures than veridical images, whereas for the remaining four targets the ratings did not differ for veridical images and caricatures. Importantly, they found a correlation between familiarity and the degree of caricaturing on the image chosen – the more familiar a face, the more likely that a caricature was the best likeness. In Experiment 2, they used a name-face matching task while measuring reaction time. The fastest RT was found for caricatures (except for the 48% caricature that did not differ from -48% anti-caricature). However, this RT advantage was limited to mismatch trials, meaning that subjects were quicker to reject a name that did not match the caricature.

Benson and Perret (1994) showed subjects line drawings of famous faces and asked them to adjust the image using a slider to achieve the truest likeness of the person. Since the slider was continuous, they were able to overcome the limitations of categorically increasing the caricature level. They found that the best likeness was about 42% over the veridical image. Interestingly, when analyzed by each face, they found that distinctive faces required less caricaturing to achieve the best likeness. In Experiment 2, they replicated Experiment 1 and found that the best caricature represented a 54% exaggeration. When the best likeness caricatures were given to a separate set of subjects in a naming task, they also found that caricatures were named faster and more accurately than veridical images. Curiously, this advantage was completely eliminated when the same task was given to subjects with the external features of the faces removed. It is unclear why that is the case, however the authors speculated that external and internal features may play different roles in identifying faces, as was found in Ellis et al., (1979), and recently confirmed by Kramer et al., (2018).

The interaction between the distinctiveness of a face and the degree of caricaturing needed to achieve best likeness found in Benson and Perrett (1994) is intriguing as it may provide an account for the mixed findings in the literature. Rhodes et al. (1996) aimed to further explore whether the caricature advantage is larger for typical faces than for distinctive faces. Subjects were trained to learn unfamiliar faces with varying distinctiveness followed by a recognition test of those faces with varying levels of caricaturing. They found that distinctive faces were recognized better than typical faces, and that caricatures were recognized better than anti-caricatures. However, they did not find an interaction between the degree of caricaturing and distinctiveness. The authors speculated that the amount of facilitation produced by a caricature can be described as a function of the relative difference between the *advantage* from caricaturing that results from moving the face to a less dense location (nominator) and the *disadvantage* from caricaturing that results from distorting the face (denominator). To cancel out the benefit of moving a typical face to a less dense location (the *advantage*), the degree of distortion needed would be relatively large (the *disadvantage*). For a highly distinctive face, the face is already in a less dense location (the *advantage*), and distorting the face (the *disadvantage*) to a large extent could result in a non-face-like appearance. Therefore, caricaturing would facilitate recognition for typical and highly distinctive faces to a lesser extent than faces that are moderately distinctive, indicating a curvilinear relationship between typicality and the most optimal degree of caricaturing to recognize a face.

Taken together, the literature suggests that caricatures of faces seem to be recognized at least as well as undistorted images, and indeed better than anti-caricatures. Although Hagen and Perkins (1983) and Tversky and Baratz (1985) did not find a caricature effect, they used photographs where textural information may override the benefits gained from exaggerating facial features. It also seems like not all faces benefit from caricaturing the same way. For example, Benson and Perret (1991) found that whether the caricature was rated as best likeness varied from one face to another. Similarly, Rhodes and Tremewan (1994) found the caricature advantage for familiar faces in a naming task, but the effect was not replicated with a different set of faces. Benson and Perret (1994) found that distinctiveness plays a role in the degree of caricaturing required for faces to achieve best likeness with typical faces needing higher levels of caricaturing. It is also evident that the effect is more reliably demonstrated with familiar faces, with mixed findings for unfamiliar faces.

Rhodes et al. (1987) argued that since the caricature advantage was primarily observed with familiar faces, it must stem from a long-term memory mechanism. Accordingly, the caricature advantage could be evidence that faces are stored as caricatures, or, that faces are stored veridically, but caricatured inputs activate the stored representation better than veridical inputs because it does not activate as many similar competing representations as the veridical input. In the former case, the caricature advantage would occur because the caricature is a better match to the representation of the face in memory. In the latter case, caricatures provide better access to the mental representation of a face, because distinctive features of the face are exaggerated. In turn, although the caricature is not as similar to the veridical representation as similar to the veridical representation as the veridical input, it is less similar to competing representations in memory. Indeed, Benson and Perret (1991) found that subjects were quicker to reject a name that mismatched the caricature compared to the veridical image of the same person. These findings provide support for the interpretation that caricatures provide better access to the veridical representations due to their increased dissimilarity to other faces. Benson and Perret (1991) speculated that there could be multiple layers of representation, one level concerned with details representing faces in a metrically accurate way (i.e., veridical) that is useful for matching different inputs of a single identity, and another higher-level layer of representation concerned with discriminating faces that emphasizes distinctive features in comparison to a prototype (i.e., telling people apart).

Mauro and Kubovy (1992) argued that faces are represented as caricatures because distinctive features are emphasized as deviations from the prototype during encoding, indicating that a short-term memory mechanism was responsible for the caricature advantage. In their experiment, subjects were given a sequential matching task and were asked to indicate whether the two images were the same or not. They found that when a face preceded its caricature, subjects were slower to respond “different” compared to when a caricature preceded its face. Accordingly, the authors claimed that when a face is encountered it is encoded in a caricaturized form, making it slower to distinguish it from the caricature because it resembles the representation of the face in short-term memory.

It is unclear whether face representations are caricaturized or veridical. Mauro and Kubovy (1992) argued that there is no point in time where a face representation is veridical, because the face gets caricaturized during encoding. However, Rhodes et al. (1987) found that when subjects are asked to choose the best likeness image of an unfamiliar face, veridical line drawings were chosen over caricatures, even when subjects were asked to rely on their immediate memory. Had they been encoding the face as a caricature in short-term memory, it would be expected that they would choose the caricature as the best likeness because it would be a better match to the representation of the face in short-term memory. Moreover, some findings revealed similar performance for veridical and caricature images, making it more difficult to discern what the face representation in memory looks like. For example, Rhodes and Tremewan (1994) found that caricatures of familiar and unfamiliar faces were recognized as accurately as veridical images, and both were recognized better than anti-caricatures. However, they also found that at some caricaturing levels for some faces, caricatures were recognized better than veridical images. Lee et al. (2000) obtained similar results in a naming task where some target faces showed a caricature advantage and some did not, but they found an overall caricature advantage when data were collapsed across targets. Benson and Perret (1991) found a recognition advantage for caricatures, but their likeness results were mixed; some targets' caricatures received higher likeness ratings, whereas some targets' veridical images received higher likeness ratings. Therefore, even though Mauro and Kubovy's (1992) findings provide compelling evidence that face representations are caricaturized in memory, they do not completely rule out the possibility that caricatures simply provide better access to undistorted representations.

More recently, caricature effects have been tested using photographic caricatures rather than line drawings (e.g., Allen et al., 2009; Hancock & Little, 2011; Kauffman & Schweinberger, 2008). Although caricature effects are more pronounced and consistent for line drawings (e.g., Rhodes et al., 1987), they have been weaker and less consistent for photographic caricatures (e.g., Benson & Perrett, 1994), possibly due to the textural information provided by photographs.

In fact, more recent findings seem to demonstrate a preference towards the anti-caricature. For example, Allen et al. (2009) showed subjects images of themselves and of their close friends. Subjects were allowed to scroll through the images to manipulate the degree of caricaturing, then they reported the image they thought was the best likeness. The results revealed a preference for the anti-caricature (about -12%). Since subjects needed to discriminate different levels of caricaturing, the preference for the anti-caricature could be an artifact of discriminability. Due to distinctiveness, faces that deviate away from the norm (i.e., caricatures) will be easier to discriminate than faces that deviate towards the norm (i.e., anti-caricatures). Therefore, preference for the anti-caricature could be due to its closer resemblance to the veridical image. Attractiveness could also play a role in choosing the anti-caricature as the best likeness. Faces closer to the average are usually perceived as more attractive, so a person's perceptions of their own and their friends' attractiveness may have influenced the preference for the anti-caricature.

Kauffman and Schweinberger (2008) investigated the neural correlates of the caricature effect using ERP. Subjects viewed familiar (personally and celebrities) and unfamiliar faces while doing a familiarity classification task, as well as best likeness ratings for familiar faces. They found no effect of caricature for the familiarity classification task, reflecting a large range of tolerance for different levels of distortion. For the likeness ratings, however, caricatures received the worst likeness rating with no difference between veridical images and anti-caricatures. Their ERP data also showed no effect of caricaturing for familiar faces. For unfamiliar faces, interestingly, caricatures elicited larger N250 amplitudes (related to personal familiarity). This is possibly due to the fact that subjects were becoming familiar with caricaturized unfamiliar faces to a greater extent in the course of the experiment since they are more distinctive.

The caricature effect could also be an artifact of adaptation, since most procedures involving caricatures require several seconds of viewing a face. Adaptation is a recalibration of the visual system after continually viewing a stimulus that biases the appearance of a subsequent stimulus in the opposite direction of the adapted stimulus (e.g., adapting to a male face makes an androgynous face appear more feminine; Webster et al., 2004). Hancock and Little (2011) investigated the role of adaptation in the caricature effect. For example, if one adapts to an anti-caricature, it may influence the preference for the caricature as a veridical representation. In their experiment, subjects were given celebrity faces that they adjusted to match the best likeness. This was done before and after adapting to an average face. The results revealed a preference for the anti-caricature before adaptation, and this effect was strengthened with adaptation to the average face. Therefore, some findings in the literature that found a preference for the caricature can be due to adaptation effects.

In summary, findings on the caricature effect are quite inconsistent, which could be an artifact of several factors. First, caricature effects are stronger for line drawings. For example, Rhodes et al. (1987) and Benson and Perrett (1994) found that caricatures (around 50% and 42%, respectively) were chosen as the best likeness and were recognized faster and more accurately. However, when photographic caricatures are used, Benson and Perrett (1991) found the best likeness caricature to be at around 4.4%. Allen et al. (2009) found a preference for the anti-caricature using photographic stimuli. Importantly, introducing color and texture information with photographic stimuli can exaggerate identity-irrelevant factors, like lighting or facial rotation, which in turn can reduce the caricature effects as subject may choose a smaller degree of caricature or an anti-caricature to reduce those identity-irrelevant factors (Hancock & Little, 2011). However, in the studies by Hancock and Little (2011) and by Allen et al. (2009) such factors were controlled in making the caricatures, and they still found a preference for the anti-caricature. Second, the type of task used can influence the result. Some studies used a recognition task, but many used goodness of likeness ratings, which may tap into different memory mechanisms. Recent research shows that likeness ratings are idiosyncratic and highly dependent on specific experiences with individuals (Ritchie et al., 2018). Indeed, Allen et al.'s (2009) finding regarding the preference for the anti-caricature could be contaminated by perceptions of attractiveness. Therefore, likeness ratings may have too much variation to sufficiently capture caricature effects. Third, not every face may benefit from caricaturing to the same extent. For example, Benson and Perrett (1994) found that distinctive faces achieve their best likeness with smaller caricaturing than typical faces. Rhodes and Tremewan (1994) found the caricatures to be effective for some faces but not others. Nevertheless, the caricature effect poses an interesting question regarding how faces are represented in memory.

1.1.3.1 Implications for the Face Space. The caricature effect is often used as evidence for the norm-based model of the face space over the exemplar-based model because there is a clear role of the prototype face in producing the effect. Caricatures are created by increasing the differences between a face and a prototype face, which emphasizes features of a face that deviate from the norm. And caricatures, although not always better recognized than veridical images, are evidently more readily recognized than anti-caricatures where the face is altered to be closer to the prototype (e.g., Rhodes et al., 1987; Rhodes & Tremewan, 1994). This suggests that faces are represented according to their deviation from the norm. However, the exemplar model could also explain the caricature effect solely in terms of a distinctiveness advantage; exaggerating facial features makes a face more distinctive and carries it to a more isolated region in the face space. The challenge for the exemplar model then would be to explain how the identity is preserved when the absolute point of the representation is moved. Therefore, it is assumed that the advantage from increasing distinctiveness would offset the disadvantage from distorting the face, making the face more recognizable.

Lateral caricatures, images in which a face is shifted “sideways” in the face space (Rhodes, 1996; Figure 1.6; Figure 1.7), have been used to understand the construction of the face space as they provide a unique case in which the amount of distortion can be the same as caricatures and anti-caricatures, but the direction of distortion is now orthogonal to the identity vector. The exemplar-based model and the norm-based model can make opposing predictions for the recognition of lateral caricatures. The exemplar-based model predicts that lateral caricatures will be better recognized than anti-caricatures because lateral caricatures are located in a region where lower exemplar density is found compared to anti-caricatures. The norm-based model predicts that lateral caricatures will be recognized more poorly than anti-caricatures, because not only the degree of distortion but the *direction* of distortion is important for recognition. However, the norm-based model can still account for better recognition of lateral caricatures than anti-caricatures if the distinctiveness of laterals offsets the impact of orthogonal distortion from a face’s vector (Rhodes et al., 1998).

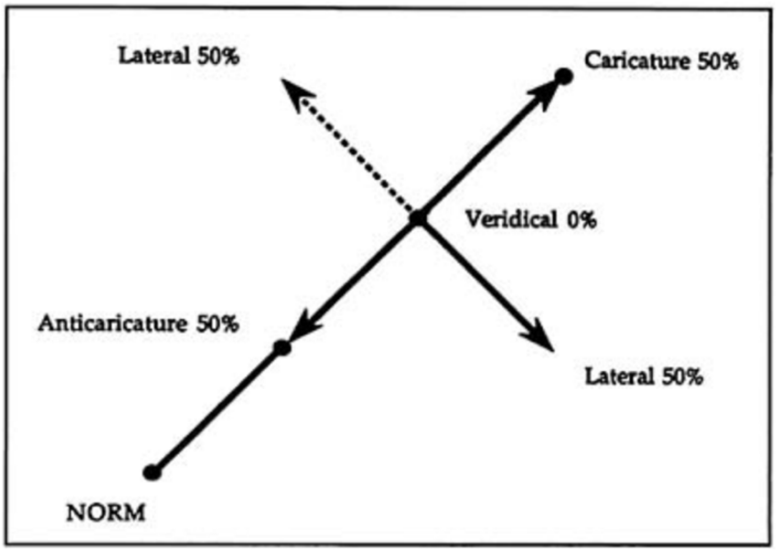

Figure 1.6. An illustration of the direction of vectors for caricatures, anti-caricatures, and lateral caricatures in a norm-based model of the face space, taken from Rhodes (1996).

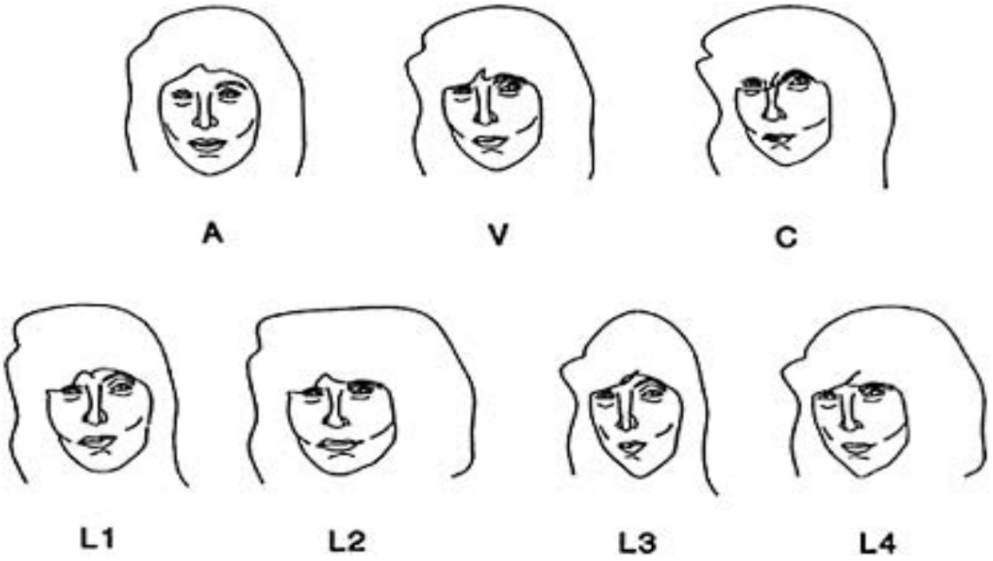

Figure 1.7. 50% distortions and a veridical line drawing of Cher taken from Rhodes (1996). A = Anti-caricature, V = Veridical, C = Caricature, L1-4 = Lateral Caricatures.

Unpublished data from Carey et al. (1992; cited in Carey, 1992; Rhodes et al., 1998) showed that lateral caricatures were recognized more poorly than anti-caricatures, providing evidence for the norm-based account. However, Rhodes et al. (1998) reported that the stimuli were not created properly, which may have led to an unnecessary distortion of the lateral caricatures, contributing to poor recognition. Therefore, they aimed to replicate whether recognition of lateral caricatures is worse than anti-caricatures. In their experiment, they created line drawings of veridical images, caricatures, anti-caricatures, and four lateral caricatures of familiar celebrities. Lateral caricatures were created by moving the points on the face perpendicular to the direction of the norm-deviation vector (i.e., the identity vector). Subjects were shown these images and were asked to name the celebrity depicted in the image while their accuracy and reaction time were measured. They found that laterals were more accurately recognized than anti-caricatures with no difference in speed of identification, which is not consistent with either of the models.

Lewis and Johnston (1998) also aimed to explore the recognition of lateral caricatures by creating photographic caricatures through morphing unfamiliar faces. They created an average (i.e., prototype) face by combining a set of faces which was then morphed with a target face in varying amounts to create veridical images and anti-caricatures (i.e, veridical: ¾ face – ¼ average, anti-caricature: ½ face – ½ average). The caricature was the original face image as it was the most distinctive due to not being morphed with an average face. The lateral caricature was created through morphing the veridical image with a reference face that was not used in the set of faces that were used in averaging, which distorted the face in an orthogonal direction (i.e., *oblique* caricatures). In their first experiment, to test the processing of these caricatures, they employed a method similar to Mauro and Kubovy (1982) where subjects were shown the veridical image followed by two test images that were either a caricature, an anti-caricature, or an oblique caricature. Subjects were asked to choose the image most similar to the target (veridical) face. They found that caricatures were chosen more often than anti-caricatures and oblique caricatures, with no difference between anti-caricatures and oblique caricatures.

Lewis and Johnston (1998) reasoned that oblique caricatures can be delineated into two distinct types: Positive oblique caricatures are those that are located somewhere between the caricature and the perfect lateral caricature, whereas negative oblique caricatures are located somewhere between the anti-caricature and the perfect lateral caricature (Figure 1.8). According to the norm-based model, caricatures are recognized better because the direction of distortion is the same as the direction of the identity vector, thus caricatures make the identity vector larger while maintaining its direction. For oblique caricatures to create better representations, the angle between the vector associated with the oblique caricature and the identity vector should be smaller (i.e., closer to each other). Therefore, the norm-based model predicts that more distinctive positive oblique caricatures and less distinctive negative obliques will produce representations closer to the veridical (Figure 1.8a). According to the exemplar-based model, recognition depends primarily on the exemplar density in the region where the representation is located. For both positive and negative oblique caricatures, increasing the distinctiveness will move the representation to a less dense region, producing better representations (Figure 1.8b). Therefore, the norm-based model and the exemplar-based model make different predictions for the recognition of negative oblique caricatures.

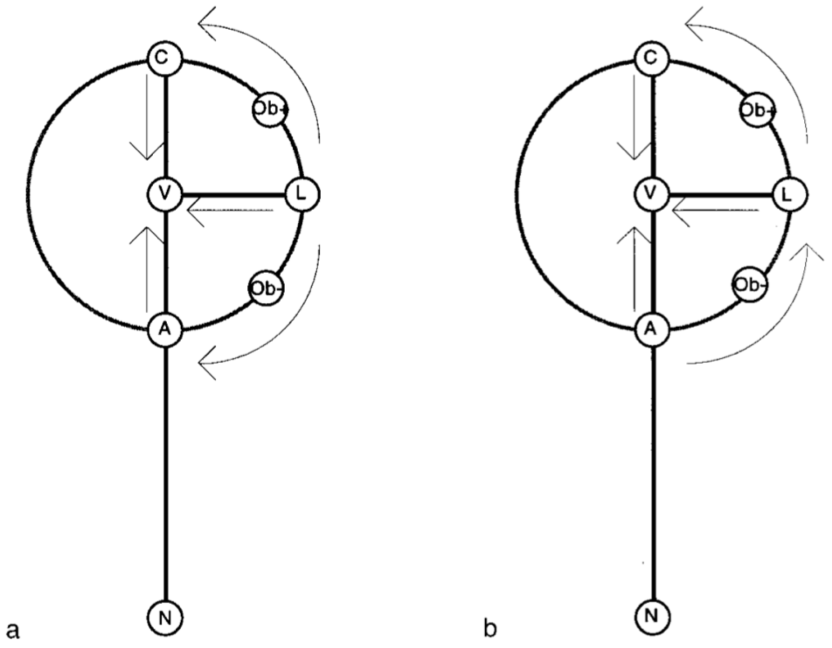

Figure 1.8. Different predictions made by (a) the norm-based model and (b) the exemplar-based model. The direction of arrows illustrates the direction where representations that are more similar to the veridical representation. The main difference is the direction of arrow for the negative oblique caricatures. N = Norm (prototype); V = Veridical face; C = Caricature; A = Anti-caricature, L = Lateral caricature; Ob+ = Positive oblique caricature; Ob- = Negative oblique caricature. Taken from Lewis & Johnston (1998).

In their second experiment, Lewis and Johnston (1998) aimed to test the perceived similarity between the different types of oblique caricatures. They followed the procedure of the first experiment (i.e., forced-choice similarity task) using the same stimuli while also including the veridical image displayed as the test face. They also collected distinctiveness judgments to determine the direction of the oblique caricatures, which revealed that all oblique caricatures were negative as they were rated less distinctive than the veridical images. The perceived similarity task showed that when the difference between the distinctiveness of the (negative) oblique caricature and the veridical image is decreased, the probability of choosing the oblique caricature over the veridical image increased, providing evidence for the exemplar-based model

(Figure 1.8b). In other words, more distinctive negative obliques correspond better with the representation in memory than less distinctive negative obliques.

Lewis and Johnston (1999) proposed the Voronoi-based model, an extension of the exemplar-based model, to provide a unifying account for the caricature effect. Voronoi diagrams are structures that divide a Euclidean space into discrete regions based on a certain number of locations. Each region is referred to as a "cell" that includes locations that are called "sites" (Figure 1.9). According to this model, each face is treated as an identity region (i.e., cell) rather than a single point in the face space as in the traditional exemplar-based model. A site within each region refers to the exact representation of a face with invariant physical properties across the dimensions of the face space. Therefore, although there is a single ideal representation for each face, recognition is achieved when a face corresponds to any representation within the identity region. The ubiquity of Voronoi diagrams in nature (e.g., giraffe skins) and the account for recognition across different instances of a face make this model particularly compelling.

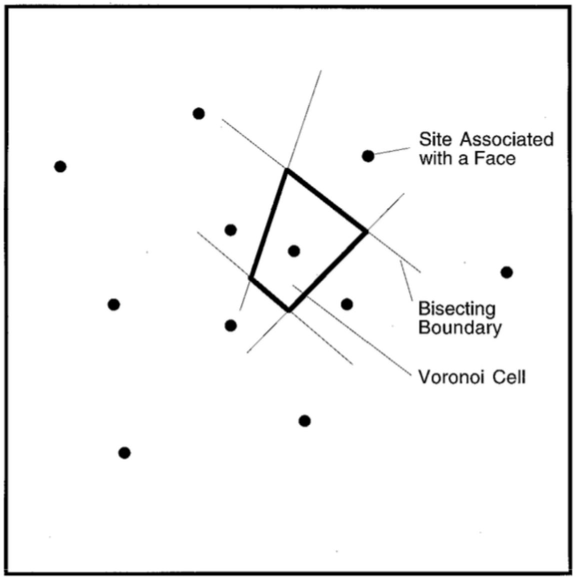

Figure 1.9. A Voronoi cell in a two-dimensional space. Bisecting boundaries are driven at the halfway-point between two neighbouring sites. Each site refers to the ideal representation of a face, whereas the cells correspond to the identity region. Taken from Lewis and Johnston (1999).

A characteristic feature of the Voronoi model is that it emerges from a normally distributed face space, which creates a predominance of sites that are off-centred[1]. The cells with off-centred sites have their sites towards the origin of the face space rather than at the centre of the cell, which means the centre of the cell is often further away from the origin. Recognition is assumed to improve when the representation is closer to the centre of the cell since the centre of the cell will be more distant from near neighbours. According to the Voronoi-based model, caricaturing does exactly this. The caricature corresponds to a representation that is closer to the centre of the cell, which makes it less similar to the competing representations.

The model can also account for the distinctiveness effect based on exemplar density. Due to the normally-distributed nature of the face space, typical faces that are clustered around the origin have smaller cells. In turn, the noise associated with encoding the input becomes more detrimental to recognition since a small amount of noise is sufficient for the input to fall outside of the cell. Lewis & Johnston (1999) created simple neural network simulations using the Voronoi-based model and were able to replicate the distinctiveness effect and the caricature effect using both two-dimensional and five-dimensional models. Interestingly, they found that more distinctive faces required less caricaturing for optimal recognition (see Benson & Perrett, 1994 above). In line with this model, faces are encoded veridically, and caricatures speed recognition during retrieval; and this blur between encoding and retrieval accounts for the blur regarding whether face representations in long-term memory are veridical or caricaturized. Therefore, according to the Voronoi-based model, face representations are neither veridical nor caricatures, but rather regions that encapsulate both and many other representations.

[1] To generate a Voronoi diagram for better illustration, navigate to https://cfbrasz.github.io/Voronoi.html.

In conclusion, the face space can provide an account for the finding that caricatures of faces are better recognized their corresponding anti-caricatures, despite that the amount of distortion is the same. It is unclear whether the norm-based model or the exemplar-based model provides a better description of the face space using the caricatures. The norm-based model suggests that caricatures are recognized better; since faces are represented based on how they deviate from the norm (i.e., distinctive characteristics), caricaturing enhances characteristics of a face that are distinctive. This, in turn, preserves the direction of the identity vector while enlarging its magnitude, which makes them recognized more readily. According to the exemplar-based model, the ease of recognition depends on the exemplar density around the region of the face representation. Caricaturing moves the face representation to a less dense location through enhancing distinctive features, which makes the face easily recognizable. The Voronoi-based model, which is a version of the exemplar-based model, suggests that caricatures are recognized better because caricaturing moves a face moves the representation to a more centred point within the identity region, which in turn reduces the competition from the nearing neighbours.

1.1.4 The Other-Race Effect: "They All Look Alike to Me"

A number of studies have demonstrated an effect of race in recognizing faces – the other-race effect (ORE) is the tendency to recognize faces of one's own race better than faces of other races (e.g., see Bothwell et al., 1989 for a meta-analysis)[2]. For example, in Shepherd et al. (1974), African and European subjects studied African and European faces. They were asked to recognize the study faces among distractors a day later. They found that European subjects

[2] The term "race" in this literature is rather loosely defined as major anthropological groups (Valentine et al., 2016), and it is often used interchangeably with ethnicity. Due to the context of the research discussed, race here refers to structural visual information derived from stimuli. Indeed, the *perceived* race of a face is found to influence face processing even if the visual features are left the same, suggesting an overall psychological out-group effect irrespective of physiognomic factors (see Lewis & Hills, 2018). The nomenclature I use to refer to different racial groups is taken verbatim from the papers I cited, which is why the terminology may seem inconsistent.

recognized European faces better than African faces, and African subjects recognized African faces better than European faces. Ellis et al. (1975) showed Black and White faces to Black and White subjects and asked them to describe the faces to a person who would need to pick them up from a nearby station that did not know them. They assumed that the features that people pay attention to would be reflected in their descriptions of faces, and they found that White subjects were using features that are more useful for discriminating White faces but not necessarily Black faces, such as eye color, hair color, and hair texture. Similarly, Black subjects used a greater number of useful features to discriminate Black faces compared to White subjects, such as hair position, eye size, whiteness of the eyes, eyebrows, and so on.

Although the effect is quite robust, the theoretical factors contributing to this phenomenon are not fully understood. It is not the case that any one race is inherently more difficult to recognize than a different race as anthropometric analyses provided no evidence for differences in facial heterogeneity based on race (e.g., Goldstein, 1979), and the effect is demonstrated across races as was shown in Shepherd et al. (1974) and Ellis et al. (1975; see also Ng & Lindsay, 1994; Shepherd et al., 1974; Valentine & Endo, 1992). Therefore, it is assumed that there are differences between the memorial representations, or perceptual processing (Megreya et al., 2011) of own- and other-race faces. The most prevalent hypothesis in the literature is the contact hypothesis, suggesting that the ORE occurs due to limited contact with other-race faces. If so, increasing contact should reduce or eliminate the effect.

In Cross et al. (1971), Black and White subjects studied a set of 12 Black and White faces and later were asked to recognize them among distractors. They found that White subjects recognized White faces better than they recognized Black faces. Black subjects recognized White and Black faces equally well. Interestingly, White subjects in segregated neighborhoods made more false alarms for Black faces compared to White subjects who lived in integrated neighborhoods, suggesting that the ORE can be reduced with increasing contact. Goldstein and Chance (1985) provided evidence that the ORE can be reduced with training. In their study, White subjects were given a paired-associate learning task during a training session to learn Japanese faces over multiple sessions. It was found that training performance improved over subsequent sessions, trained subjects performed better than the control group for the recognition task, and the effects of training persisted after five months.

Chiroro and Valentine (1995) provided further evidence for the contact hypothesis. They included Black and White subjects with varying self-reported contact with other-race faces. Subjects studied faces followed by a recognition test. The *A'* scores indicated that low-contact subjects, irrespective of the race of the subject, demonstrated an ORE. High-contact white subjects also demonstrated an ORE, but high-contact Black subjects recognized faces of both races equally accurately. Although the contact hypothesis is compelling, there is also evidence that it does not influence recognition performance. For example, Ng and Lindsay (1994) recruited East Asian and White subjects from Canada who reported the extent and the intimacy of their relationships with own- and other-race faces. Subjects were shown Eastern and White faces along with four comparison images and they were asked to report the similarity between the target face and the comparison faces. Later, they were given a recognition task on those faces. The authors found that both groups of subjects demonstrated an ORE in the *d'* measure that was driven by false alarms. There was no correlation between contact and recognition performance, which was replicated by incorporating subjects from Singapore who had equivalent other-race contact with White subjects.

It should be noted that testing for the contact hypothesis is quite challenging because many contemporary societies are multi-racial, there can be social desirability effects at play when reporting contact with other races (Chiroro & Valentine, 1995), and there is increasing influence of visual media where one can be exposed to a diverse range of races. Further, contact can take many shapes and forms. For example, one can be surrounded by other-race faces but may not feel the need to discriminate among them (e.g., an exchange student only hanging out with a few people of their own ethnicity). Therefore, quality of interactions may be more important than the quantity of interactions, and quality of interactions may be difficult to capture.

As is found in distinctive faces, there is an asymmetry in the classification and identification performance of other-race faces. While recognition of other-race faces is poor, classifying an other-race face as belonging to that race is better than for own-race faces. For example, Ge et al. (2009) recruited Chinese and Caucasian subjects and showed them own- and other-race faces to recognize later. They were also given a race classification task where they categorized the faces as either Chinese or Caucasian. The reaction time results revealed a three-way interaction between the subject race, race of face, and the type of task. Caucasian subjects recognized Caucasian faces faster and classified Chinese faces faster, and Chinese subjects recognized Chinese faces faster and classified Caucasian faces slower. They also found a negative correlation between recognition and classification performance, suggesting that expertise with a particular race comes at a cost in categorization.

1.1.4.1 Implications for the Face Space. There is some evidence that dimensions used to differentiate other-race faces differ from own-race faces. For example, Ellis et al. (1975) found that Black and White subjects differed in their descriptions of facial features of Black and White faces. In Shepherd and Deregowski (1981), Black and White subjects used different attributes to describe the similarity between two Black faces and two White faces. Therefore, according to the face space framework, it is assumed that the face space is constructed most optimally to discriminate among own-race faces, therefore, other-race faces are encoded more densely than own-race faces (Figure 1.10; Valentine & Endo, 1992; Valentine et al., 2016). Both the exemplar-based and the norm-based model suggest that the ORE is due to the use of inappropriate dimensions when encoding other-race faces since the dimensions of the face space are most appropriate for encoding own-race faces. The two models, however, differ in the encoding processes and how the similarity between two faces is calculated. According to the exemplar-based model, the similarity between two faces is calculated as the Euclidean distance between the two faces, for both own-race and other-race faces. How the similarity is calculated for the norm-based model is not as straightforward. Valentine (1991) and Valentine and Endo (1992) suggest the similarity is determined by the angular separation or the dot product of two vectors. However, these similarity metrics fail to meet the assumptions that collinear representations are not necessarily identical (even though their angular separation is 0°), and collinear representations will be more similar to each other than any other two noncollinear representations that are equidistant (Byatt & Rhodes, 1998). Therefore, according to Byatt and Rhodes (1998), the similarity metric for the norm-based model is the cosine of the angle between two vectors divided by the simple distance between two faces.

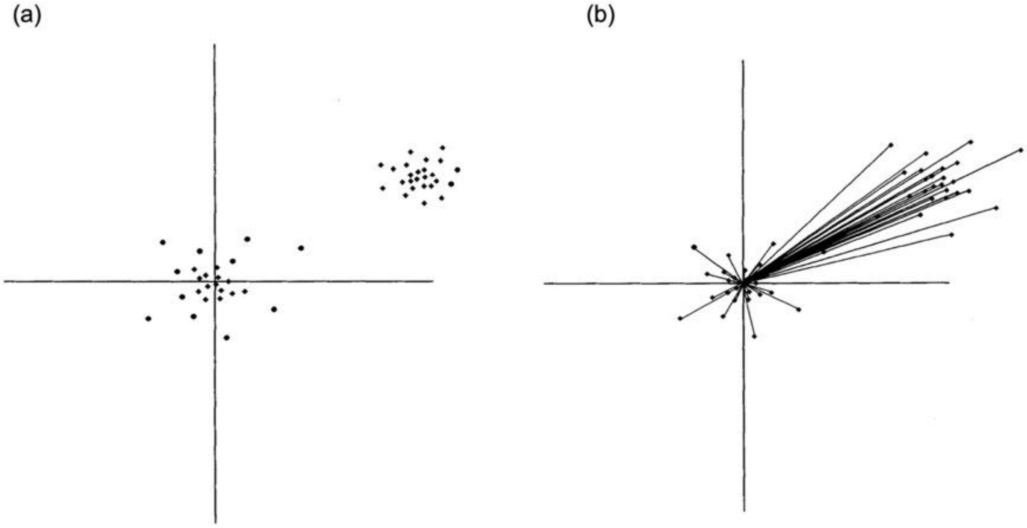

Figure 1.10. The own- and other-race faces in a two-dimensional hypothetical face space. (a) A purely exemplar-based model. Around the origin is where own-race faces are encoded. Other-race faces are encoded in a denser cluster as they are encoded using inappropriate dimensions. (b) A norm-based model. Other-race faces are encoded using the own-race norm, which is not optimal for coding other-race faces. Taken from Valentine and Endo (1992).

Although other-race faces are more densely clustered in the face space, there is no *a priori* reason to assume that the *difference* in exemplar density around typical and distinctive faces would be less for other-race faces. Valentine and Endo (1992) aimed to explore whether or how distinctiveness interacts with the other-race effect. They first collected distinctiveness ratings from Japanese and British subjects using Japanese and British faces and found that subjects of both races agreed on which faces were distinctive. British faces also were rated as more distinctive than Japanese faces for both groups of subjects. When subjects were given a recognition task on these faces, they found that both Japanese and British subjects demonstrated an ORE in *A'* scores, both groups of subjects had more hits for British faces, British subjects made more false alarms for Japanese faces with no main effect of race of face for Japanese subjects in false alarms, and Japanese subjects correctly rejected Japanese distractors faster than British distractors with no main effect of race of face for British subjects in correct rejections. In all five measures an effect of distinctiveness was obtained, but there was no interaction among distinctiveness, race of face, and race of subjects. In other words, the effect of distinctiveness on recognition was equivalent for both own-race and other-race faces.

In Experiment 3, subjects were given a face classification task. Although typical faces were classified faster than distinctive faces consistent with the previous literature, other-race faces were not classified faster than own-race faces, which would be predicted by the exemplar-based model due to the high exemplar density of other-race faces, similar to typical faces. Valentine (1991) found other-race faces to be classified faster; however, that study was not cross-cultural. The authors reasoned that a face classification task may be an inappropriate task to demonstrate an interaction between race of face and race of subject since the face classification task requires a judgment of a superordinate category (face irrespective of race). Therefore, In Experiment 4, subjects were given a race classification task with the rationale that a race classification task would require judgment of a subordinate category (race of face). They found that British subjects classified Japanese faces faster than British faces, but Japanese subjects classified both races of faces equally. Additionally, British subjects classified typical British faces faster than distinctive British faces, and Japanese subjects classified typical Japanese faces faster than distinctive Japanese faces. Neither group demonstrated a distinctiveness effect for the other-race faces, demonstrating a greater distinctiveness effect for own- but not other-race faces.

These findings are supported by the exemplar-based model but not the norm-based model. The exemplar-based model suggests that recognition of other-race faces would demonstrate an equivalent distinctiveness effect to own-race faces, because the *relative* difference in exemplar density for own- and other-race faces is assumed to be similar. In other words, even though other-race faces are more densely clustered, distinctive other-race faces are still encoded in a relatively more isolated region compared to typical other-race faces. This was demonstrated by the first experiment where there was an agreement in both groups of subjects on which faces were distinctive irrespective of the race of face, and the lack of interaction of distinctiveness with race effects. The exemplar-based model's prediction for race classification performance is not clearly outlined, however. Since classification relies on a between-category judgment, an effect of distinctiveness in classification of own-race but not other-race faces can be a result of greater exemplar density of other-race faces. Therefore, the exemplar-based model can account for the results obtained from Valentine and Endo (1992). On the other hand, the norm-based model would predict that in a recognition task, the distinctiveness effect would be smaller or absent for other-race faces because the angular separation of other-race faces is smaller than own-race faces, assuming angular separation is an appropriate similarity metric for the norm-based model. This is due to other-race faces being encoded with the own-race norm (see Figure 1.10b).

The norm-based model could account for these findings if other-race faces are encoded with their own norm which reduces parsimony and adds a requirement for the appropriate norm to be selected before deriving individuating information about faces. According to the face recognition model proposed by Bruce and Young (1986), faces are encoded with semantic and structural codes, with (visually derived) semantic codes corresponding to properties like sex, age, and race, and with structural codes corresponding to identity information. Bruce and Young (1986) argue that these visually derived semantic codes are independent from the recognition of an identity. If an appropriate norm (e.g., someone's race) needs to be selected before accessing structural information, it would mean there is a degree of dependence between visually derived semantics and structural information. Nevertheless, Ge et al. (2009) found a correlation between race classification and recognition of other-race faces so there could be some sort of dependency between visually derived semantics and structural information.

Chiroro and Valentine (1995) aimed to investigate the effect of distinctiveness in recognizing other-race faces while including contact as a factor. They found that level of contact with other-race faces interacted with distinctiveness, such that low-contact subjects in both race groups (Black and White) demonstrated an effect of distinctiveness in recognition of own-race faces but not other-race faces. High-contact subjects in both groups demonstrated an effect of distinctiveness in recognition of both own-race and other-race faces. The authors provided evidence that the distinctiveness effect is a result of experience with a particular population of faces. This is also in line with the exemplar-based model; if the population of faces encoded in the other-race space is increased, the relative exemplar density can be more accurately established. According to Valentine and Endo (1992), the norm-based model fails to provide an account for the effect of contact; since other-race faces are encoded with an own-race norm, the angular separation among other-race faces will always be limited.

Additional evidence for the exemplar-based model came from Byatt and Rhodes (1998) using caricatures. Since the norm-based model assumes other-race faces to be encoded using the own-race norm, other-race caricatures using an own-race norm (i.e., wrong-norm) should be easier to recognize than caricatures created with the race’s own norm (i.e., right-norm). The exemplar-based model makes the opposite prediction, since the caricature advantage stems from reduced exemplar density and other-race faces are encoded with inappropriate dimensions, right-norm caricatures of other-race faces should be easier to recognize than wrong-norm caricatures. To test these predictions, European subjects were shown Chinese and European faces and were asked to rate them on distinctiveness and learn their names. Later, they were tested on their recognition of the veridical images, right-norm caricatures, wrong-norm caricatures, right-norm anti-caricatures, and wrong-norm anti-caricatures. They found that own-race veridical faces were recognized as well as own-race caricatures, and that both were recognized better than anti-caricatures. For other-race faces, all images were recognized equally well except for the -50% anti-caricature which was recognized poorly. Importantly, in all conditions, right-norm distortions were more accurately recognized than wrong-norm distortions, providing evidence for the exemplar-based model. The authors argued that the norm-based model could account for these results if other-race faces are encoded with their own norm, but then the norm-based model would fall short in explaining the other-race effect, because the norm-based model's explanation for the other-race effect is that other-race faces being encoded with an inappropriate norm.

This interpretation can be problematic, because it assumes that the other-race norm would be optimal for distinguishing other-race faces. However, limited exposure to other-race faces can form an other-race norm that is not stable enough to allow for effective discrimination of other-race faces. Burton et al. (2016) showed that averages of faces stabilize after about 20 images, which could mean that it takes about 20 instances of other-race encounters to form a stable norm of an other-race face. However, this account would suggest that each encounter with a face is included and weighed equally in the formation of the average. How a facial norm emerges, to my knowledge, is not exactly known and is a potential avenue for future research. For now, there does not seem to be clear evidence favoring either of the models, with seemingly more evidence in line with the exemplar-based model (e.g., Byatt & Rhodes, 1998; Valentine & Endo, 1992). The norm-based model seems to be able to account for many findings with the assumption that each category has its own norm, but then the norm-based model becomes indistinguishable from the exemplar-based model.

The face space framework also fits with the mirror effect of recognition and classification of other-race faces (e.g., Chiroro & Valentine, 1995; Ge et al., 2009). In the mirror effect, distinctive faces are recognized faster but classified (as faces) slower. Other-race faces are recognized more slowly but classified (as belonging to which race) faster, which is due to the high exemplar density of other-race faces. Accordingly, recognition of other-race faces is impaired due to the close proximity of near neighbors, whereas classification performance of other-race faces is better because the overall activation of the region is high. An alternative account comes from Levin's (1996; 2000) race-coding hypothesis. Put simply, the race-coding hypothesis suggests that other-race faces are encoded explicitly with their race information at the expense of individuation information. Therefore, the other-race effect is not due to an expert coding of individuation information of own-race faces per se, but rather an expert coding of race information of other-race faces. Consequently, classifying other-race faces would not be influenced by distinctiveness, because the classification performance is controlled by a unifying race-feature rather than individuating information like typicality. Levin also hypothesized that whereas other-race faces are encoded with a present (positive) race feature, own-race faces are encoded with an absence (negative) of the race feature. In line with Treisman and Gormican's (1988) visual search asymmetry, targets with present features are found faster than targets with absent features. As such, in a visual search task where an other-race face is the target among own-race distractors, the target would be found faster than if it is an own-race face among other-race distractors.

In Levin (1996), subjects were given a race classification task on famous and unfamiliar own- and other-race faces where distinctiveness was systematically manipulated through caricaturing using both line drawings (Experiment 1) and photographic stimuli (Experiment 2).

The results revealed that subjects were faster at classifying other-race faces than own-race faces, but only when the faces were very distinctive (caricatured). There was no classification advantage for typical faces. In another experiment, inverted faces were also used as stimuli to investigate whether disruption in configural processing (associated with expert processing of individuating information; Diamond & Carey, 1986) would account for the classification performance. Similar to Valentine's (1991) finding, inversion did not reduce the classification advantage of other-race faces, indicating that the ORE is not necessarily a result of disruption in configural processing. Lastly, subjects were asked to either search for an own-race target among other-race distractors, or for an other-race target among own-race distractors to test for the visual search asymmetry. The results revealed that subjects were quicker in locating an other-race face than an own-race face, and that increasing set size made them disproportionately slower to locate an own-race face than an other-race face. The search asymmetry suggests that other-race faces are encoded with a present feature, whereas own-race faces are encoded with the absence of the race feature. However, when tested cross-culturally, the search asymmetry was replicated for White subjects, but Black subjects did not demonstrate an asymmetry.

Levin argued that the race-coding hypothesis is an *alternative* to the face space framework. However, Byatt and Rhodes (2004) suggest they do not have to be mutually exclusive, as the dimensions used in encoding race information may receive more weight for encoding other-race faces than own-race faces, which would provide support for both the race-coding hypothesis and the face space theory. To investigate whether the ORE stems from higher exemplar density, Byatt and Rhodes (2004) collected similarity ratings for pairs of own- and other-race faces from one set of subjects, and gave a recognition task to another set of subjects. The similarity ratings showed that, despite being instructed to focus on individuating information rather than race, own-race faces were rated as more similar to each other than to other-race faces and vice versa. Importantly, other-race faces were more similar to each other than own-race faces were to each other. A multidimensional scaling analysis by Lee et al. (2000) confirmed that own- and other-race faces formed distinct clusters, with other-race faces forming a denser cluster than own-race faces. This clustering was an accurate predictor of the other-race effect.

Rhodes et al. (2009) further explored the race-coding hypothesis as an underlying effect of other-race recognition performance. They reasoned that if coding race information is at the expense of individuating information, then paying attention to race information should be detrimental to identification performance irrespective of race of face. Subjects were asked to either passively view a set of own- and other-race faces, rate the attractiveness of the faces, or rate typicality of the faces based on race. They found that when given a recognition task on these faces, subjects in all groups demonstrated an ORE. In fact, subjects who rated the faces demonstrated an even better recognition of own-race faces than other-race faces compared to the subjects who passively viewed them. The ORE was replicated in a task where subjects were asked to classify races of faces, suggesting that directing attention to race information does not eliminate the ORE.

In summary, according to the face space framework, the ORE occurs due to the higher exemplar density of other-race faces compared to own-race faces. This was supported by Byatt and Rhodes (2004) who found that other-race faces form distinct and denser clusters than own-race faces which accurately predicts recognition performance. The face space can also predict how the ORE can interact with phenomena like distinctiveness (Valentine and Endo, 1992), inversion (Valentine, 1991), and caricature (Byatt & Rhodes, 1998). Most findings seem to be in line with the exemplar-based model but the norm-based model could also account for the results if it is assumed that different face categories have their own norm. However, that would make the norm-based model less parsimonious and theoretically indistinguishable from the exemplar-based model.

1.1.5 The Adaptation Effect: Changing the Norm

Visual adaptation is a process in which the visual system recalibrates systematically after constant viewing of a particular stimulus property (Strobach & Carbon, 2013; Valentine, 1996; Webster & MacLin, 1999). Adapting to a stimulus changes the appearance of a test stimulus in the opposite direction of the adapted stimulus property. For example, adapting to downwards motion makes a stationary object appear like it is going upwards, commonly known as the waterfall illusion (Addams, 1834). Adaptation effects can manifest in a multitude of visual dimensions, such as orientation (Gibson & Radner, 1937), spatial frequency (Blakemore & Sutton, 1969), and color (Webster & MacLeod, 2011). More recently, adaptation effects have been demonstrated for face stimuli (e.g., Carbon & Ditye, 2012; Leopold et al., 2001; Strobach & Carbon, 2013; Webster & MacLin, 1999). The multidimensional nature of faces as a visual category means that face adaptation can take many forms depending on the adapting information such as age, gender, configuration, distortion, attractiveness, identity, and so on (Strobach & Carbon, 2013). Face adaptation is also considered a high-level process due to its transference across different sizes and retinal locations (Rhodes & Jeffery, 2006; Valentine et al., 2016).

One of the first demonstrations of face adaptation came from Lewis and Ellis (2000). They showed that repeated presentation of a name or different views of a face lead to semantic satiation (i.e., repeating a semantic concept results in a loss of the concept’s meaning). In their experiment, subjects were asked to repeat a celebrity's name out loud for either three times or 30 times, followed by a face image that either matched or mismatched with the name. It was found that subjects were slower to indicate whether the face matched the name when they repeated the name 30 times. Similarly, when they were repeatedly shown different images of a celebrity face followed by a test face that was either related (an associate) or unrelated (a non-associate), subjects were slower in their judgments of identifying the test face when the images were repeated 30 times compared to three times, suggesting that satiation can occur for non-verbal semantic categories, like facial identity.

Additional evidence in a more visual domain came from Webster and MacLin (1999) who examined the face distortion aftereffects (FDAEs). In their matching task, subjects learned an undistorted face, and later were asked to adjust a test face to match the original face, both before and after adapting to a face. In the ratings task, subjects were simply asked to rate whether a displayed image was normal or distorted, also both before and after adapting to a face. They found that when subjects adapted to a distorted face, they matched and rated the original face to be distorted in the opposite direction. For example, if the adapting face was expanded, the original face appeared contracted. Additionally, the level of distortion of the adapting image systematically influenced the appearance of the test image. Adapting to a distorted face greatly influenced the appearance of the original face, but adapting to the original face had little effect on the appearance of the distorted face. This suggests that adaptation is not simply a shift in perception but rather a renormalization process. Some faces serve as a "neutral point" (i.e., the norm), and adaptation to these faces does not influence the appearance of distorted images because the adapted stimuli reinforce the existing norm. On the other hand, adapting to a distorted face changes the appearance of a test face in the *opposite* direction, meaning that the norm may have shifted towards the adapted stimulus.

Face adaptation is assumed to be a high-level process since it transcends low-level information like retinal location and size of stimuli (Valentine et al., 2016). Nevertheless, evidence suggests that face adaptation shares dynamic properties similar to low-level adaptation processes like orientation and shape. Leopold et al. (2005) examined these dynamic properties of face identity aftereffects (FIAE). Specifically, they examined the relationship between the temporal presentation of the adapting and test stimuli and the strength of adaptation, which is a relationship previously found for other classes of visual stimuli like orientation and shape. They created faces and anti-faces (i.e., faces that lie on two opposite ends of an identity trajectory passing through the norm, see Figure 1.10) assuming a norm-based model of the face space. After being trained on discriminating the identities used in the study, subjects were given an adaptation paradigm where they were first shown a name cue, then adapted to an anti-face in varying durations followed by a test face presented in varying durations. The test face was always the average face. Subjects were asked to rate the strength of the apparent identity of the test face that matched the name cue. The results revealed an adaptation effect where subjects' ratings of apparent identity were greater and in the opposite direction of the anti-face. Importantly, adaptation was stronger for longer adaptation durations and shorter test durations, mimicking dynamic properties of adaptation to lower-level stimuli. These results indicate that adaptation could be a feature of the visual system in general to adapt to novelties in the environment, and that adaptation could serve as a tool to understand categorical representations of visual stimuli in general (see Webster & MacLeod, 2011 for a review of face adaptation paralleled with color adaptation).

Carbon and Ditye (2012) examined whether adaptation effects are transferable over different environmental settings, as it could be that being in the same laboratory environment for adaptation and test phases activates episodic representations which in turn activates adapted representations. In their study, all subjects went through an adaptation phase in the formal experimental laboratory to either strongly compressed, strongly extended, or veridical celebrity faces. After a 7-10 day delay, half of the subjects were tested in the lab, and half were tested in an informal leisure room of the department where the environmental and social properties of the experimental setting were different. The test included a 2AFC task where subjects were shown a veridical face paired with either a slightly extended or contracted face. The veridical face could be the same picture, a different picture of the same person, or a novel identity. The task was to choose the veridical face. The authors replicated the adaptation effect in that adapting to a distorted face made subjects more likely to choose the face distorted in the opposite direction (e.g., extended to contracted) as the veridical face. Adaptation was strongest for the same picture, followed by a different picture of the same identity, and weakest for a novel face. Importantly, the stability of the adapting and the testing environments did not contribute to the adaptation effects, showing that sustainable adaptation is not dependent on episodic representations. Given that adaptation effects are present after a week and in different settings, adaptation may be a continually normalizing mechanism that tunes the entire face space towards newly encountered visual information.

1.1.5.1 Implications for the Face Space. Face adaptation effects have been used to understand the architecture of face representations, specifically the nature of the face space. Adaptation effects are often used as a support for the norm-based model as they serve as a renormalization process (e.g., Webster & MacLin, 1999), indicating that the norm plays an important role in how faces are encoded. Influential evidence in favor of the norm-based model came from Leopold et al. (2001), who tested FIAEs using realistic faces. In their experiment, a face space was constructed with an average face at the origin. Importantly, they created anti-faces by generating faces through calculating the difference between the face and the average so that anti-faces and their corresponding original faces would lie on the same axis, passing through the average face (Figure 1.11). They found that adapting to an anti-face increased sensitivity in identifying the original face, i.e., adapting to anti-Adam made it easier to identify Adam. Interestingly, adapting to an anti-face made the average face appear as a new identity in accordance with the identity trajectory. In other words, adapting to anti-Adam made the average face look like Adam. This was not true for mismatch trials where the adaptation face and the test face were on different identity trajectories, for example, adapting to anti-Adam made it more difficult to identify Jim. These findings highlight the importance of a norm face in differentiating faces since adaptation systematically influenced processing a face on the same identity trajectory but not others.

Rhodes and Jeffery (2006) investigated whether the FIAEs were selective to adapting to opposite faces, or whether an equally dissimilar non-opposite face would induce similar aftereffects. As in Leopold et al., (2001), the authors created faces and anti-faces which lay in opposite ends of the face space while passing through the average face. Non-opposite faces were also created that were rated to be equally dissimilar to the target as the opposite face. Subjects were first trained to identify target faces. To collect baseline identification thresholds, subjects were asked to identify "weaker" versions of target faces that were morphed with an anti-face or a non-opposite face. Subjects were then given an adaptation paradigm where they adapted to either an anti-face or a non-opposite face. The test face was a 50/50 morph of the adapting face with the target face, and subjects were asked to identify the test face. They found that adaptation reduced identification thresholds for target faces; in other words, adapting to a face made the morph test face appear more like the target. However, importantly, this effect was significantly larger for anti-faces than non-opposite faces. Interestingly, they also found that identification thresholds were overall lower for faces morphed with their opposite compared to their non-opposite. This was confirmed with a second experiment where subjects were asked to rate the similarity between a morph face and one of its constituents. Morphs created with opposite faces were rated as less similar to their constituents than morphs created with non-opposite (but equally dissimilar) faces. These results show that the monotonic distance between two identities in the face space is not sufficient to explain the adaptation effect. Rather the *direction* of the dissimilarity plays an important role, providing support for the norm-based model.

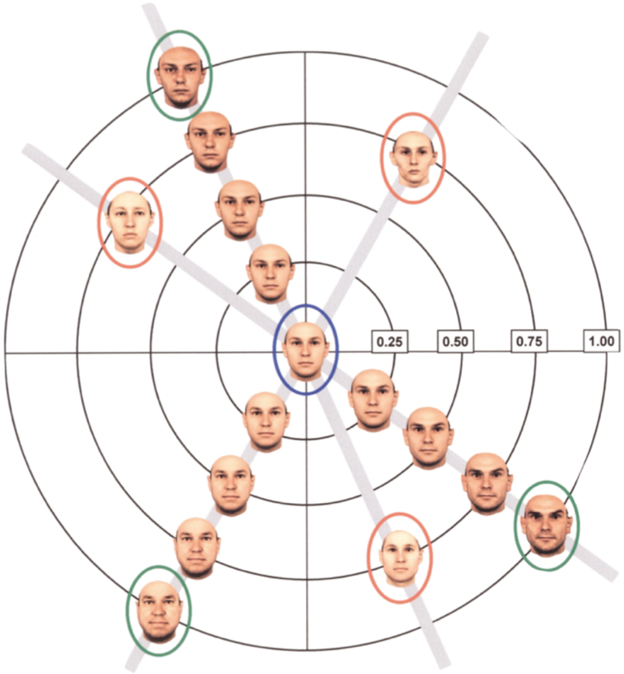

Figure 1.11. The faces and their anti-faces, taken from Leopold et al. (2001). Faces in green circles correspond to the faces (e.g., Jim, Adam) and the faces marked in red circles that are on the same line as the faces correspond to their anti-faces (e.g., anti-Jim, anti-Adam). The face at the center is the average face. 0.25, 0.50, 0.75, and 1.00 reflect identity strength, manipulated by morphing the average face to the target faces.

Ross et al. (2014), however, provided evidence that exemplar-based models can also account for the FIAEs. They developed computational models using a Gaussian distribution and PCA, where they constructed a face space assuming either an exemplar-based model, a traditional norm-based model, or a two-pool norm-based model. The difference between the traditional and the two-pool norm-based models is that in the traditional norm-based model, the norm is explicitly represented, and faces are encoded as deviations from the norm; whereas in the two-pool norm-based model the representation of the norm is implicit, and the faces are encoded as opponents based on their perceived characteristics. For example, male and female faces can be encoded as opposites in the gender dimension, or Adam and anti-Adam can be encoded as opposites in some identity dimension (Leopold et al., 2001; Rhodes & Jeffery, 2006). Additionally, the exemplar-based model was created in a way that newly learned identities were represented by distributions based on their similarity to previously learned faces as opposed to the more traditional demonstrations of the exemplar-based representations where faces are represented as unique points. The authors simulated the anti-face adaptation in Leopold et al. (2001), adaptation to opposite and non-opposite faces in Rhodes and Jeffery (2006), and the weaker adaptation effects found in adapting to the norm (e.g., Webster & MacLin, 1999). They found that the exemplar-based model and the two-pool norm-based model predicted these findings in the literature, however, the traditional norm-based model fell short in predicting some of the findings.

In summary, adaptation effects have been widely used to support the norm-based model of the face space. Specifically, adaptation to the norm does not create strong adaptation effects (Webster & MacLin, 1999), suggesting that adaptation serves as a renormalization process. Adaptation for face identity is also selective for the opposite faces that have their identity trajectory passing through the norm (Leopold et al., 2001; Leopold et al., 2005; Rhodes & Jeffery, 2006), which highlights the role of the norm in representing faces in the face space. However, simulations suggest that exemplar-based models make similar predictions about FIAEs as well as norm-based models where the norm is implicitly represented (Ross et al., 2014).

1.2 Face Familiarity: Telling Faces Together

Valentine's (1991) face space theory provides an account for explaining how the visual system overcomes the homogeneity problem (i.e., all faces share the same basic configuration) when discriminating different faces from one another. In a multidimensional representational space, each face is encoded in a location relative to other previously encountered faces based on their perceptual similarity to one another. This theory is especially appealing considering the prevalence of such multidimensional spaces to represent and organize other categories of stimuli (e.g., color; Shepard, 1962). Nevertheless, in the case of face perception and memory, it has some theoretical and experimental limitations, as the more recent literature has demonstrated. Such limitations boil down to the qualitative differences between processing familiar and unfamiliar faces. While the face space theory has greatly focused on distinguishing faces from one another, the other side of the coin is the ability to "tell faces together", that is, the visual system's ability to amalgamate different instances of a face into a single representation of that person's identity.

1.2.1 Theoretical Limitations of the Face Space: Beyond a Single Point or Vector

Earlier models of the face space claim that each face is represented as a single point (the exemplar-based model) or a single vector deviating from a norm face (the norm-based model). Such an approach to representing faces does not explain how a familiar face can be recognized across several viewing conditions as familiar face recognition is robust to changes in viewpoint, expression, age, angle, lighting, even image distortions. For example, Hole et al. (2002) showed that a familiar face can be successfully recognized even when it is vertically stretched, suggesting that familiar face recognition is robust to even distortions in the facial configuration. Bruce et al. (2001) found that familiar faces can be recognized from low-quality CCTV videos.

The ability of the visual system to tolerate such variability for familiar faces is unlikely to stem from a single memorial representation. The assumptions that there is a single “true” representation for a face, and that any change from this true representation is an encoding error or a change in identity, fall short in explaining how we can recognize familiar faces despite the variability in viewing conditions (Burton, 2013).

More recent models of the face space suggested the idea that faces are represented as regions encompassing a range of instances that can be recognized. One such model is the Voronoi model discussed above (Lewis & Johnston, 1999), where a normally distributed Euclidean-based space is divided into discrete regions (i.e., cells; Figure 1.9). Each cell contains a “site” that corresponds to an ideal representation of an identity based on their invariant properties, so the Voronoi model suggests that faces are represented as points but the region surrounding the point allows a face to be recognized across different viewing conditions. The size of the region surrounding a representation is related to how well a face can be recognized, thus, faces with large regions surrounding them can tolerate more variability. Although simulations of this model successfully replicated the distinctiveness and the caricature effects found in the literature, an important limitation of this model is that it assumes unfamiliar faces will be falsely recognized as familiar faces because the entire volume of the space is filled with identity regions. Thus, in contrast to humans who can simultaneously fail to identify a face and not mistake it for some other face, the model cannot give an “unknown” response (Lewis, 2004; Lewis & Johnston, 1999).

Tanaka et al. (1998) suggested that face representations are surrounded by “attractor fields” that can activate a face’s representation through multiple inputs (Figure 1.12). The size of the attractor field of a face is related to the density of other exemplars surrounding the representation in the face space. Since distinctive faces are located in a region with lower exemplar density, it allows their attractor fields to be larger, allowing for more instances to fall within the identity’s region. Evidence for attractor fields came from 50/50 morphs of a typical and a distinctive face. A traditional model of the face space would predict that a 50/50 morph should bear equal resemblance to its parent faces because similarity between two faces is determined by the Euclidean distance of their coordinates in the face space. The attractor field model, on the other hand, would predict that a 50/50 morph would bear more resemblance to the distinctive parent, because it would activate the distinctive face more strongly than the typical face due to the larger size of the distinctive face’s attractor field. In a series of experiments, the authors found that when subjects are asked to make a similarity judgment of the 50/50 morph between the typical and the distinctive parent, they choose the distinctive parent, providing evidence for the attractor field model.

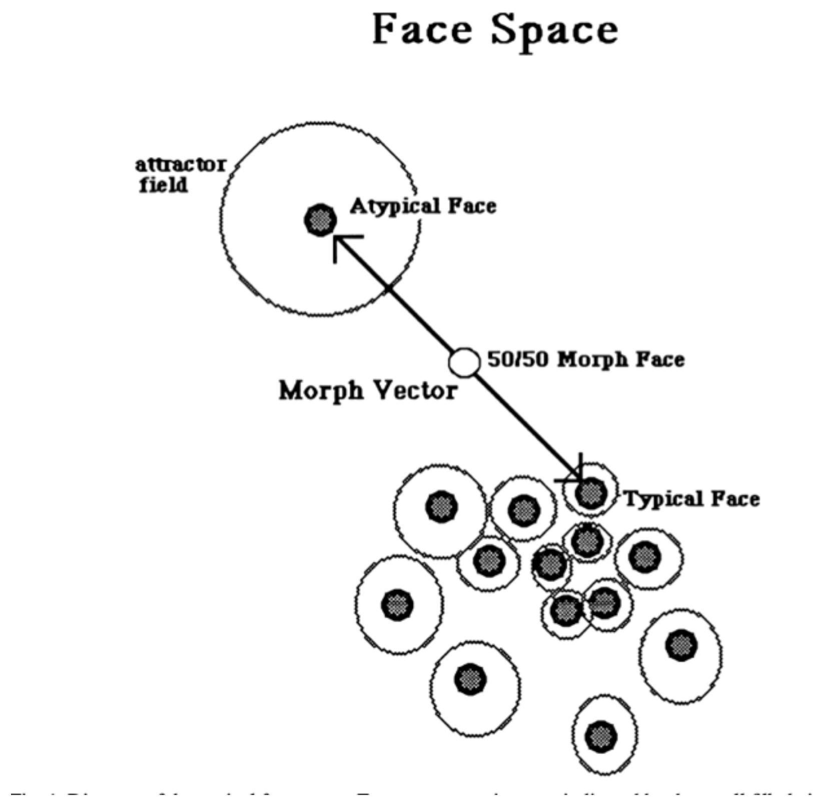

Figure 1.12. The attractor field model taken from Tanaka et al. (1998). Atypical faces have larger attractor fields due to low exemplar density, whereas typical faces have smaller attractor fields. A 50/50 morph of a typical and an atypical face is equidistant from its constituents, yet it bears stronger resemblance to the atypical face due to a stronger activation of the atypical face's large attractor field.

In another study, Tanaka and Corneille (2007) created morphs of typical and atypical faces with varying ratios to manipulate their Euclidean distance from their atypical and typical parents. Subjects were then given a sequential and a simultaneous matching task where they were asked to make a same/different judgment on a morph face paired with its typical or distinctive parent face. In line with the attractor field model, they found that morphs that were closer to a distinctive face (i.e., distinctive morphs) were less likely to be successfully discriminated from the distinctive face, whereas morphs that were closer to a typical face (i.e., typical morphs) were more likely to be discriminated from the typical face. According to the attractor field model, since distinctive faces have larger attractor fields, a morph that is *x* units away from a distinctive face will be more likely to fall within the attractor field bounds compared to a morph that is *x* units away from a typical face (Figure 1.13). This makes the perceived similarity between the distinctive morph and its distinctive parent face larger than the perceived similarity between the typical morph and its typical parent face.

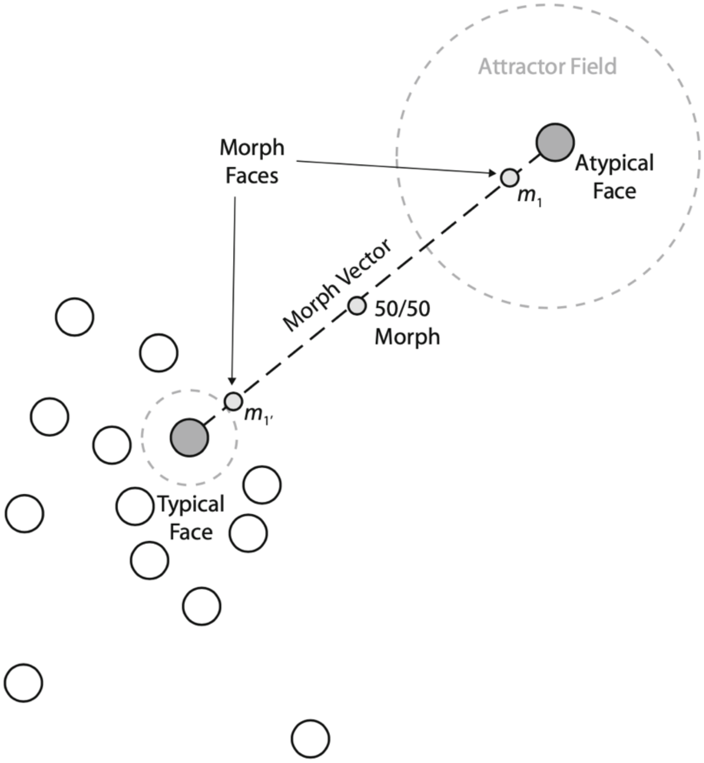

Figure 1.13. Attractor field model taken from Tanaka and Corneille's study (2007). The distinctive (atypical) and typical faces are represented in gray circles, surrounded by their attractor fields that are represented as dashed lines. Morphs are represented as lighter gray. Although the two m_1 morphs are the same distance away from the typical or the distinctive face, the morph closer to the typical face is more likely to fall outside of the bounds of its attractor field, yielding easier discrimination, whereas the morph closer to the distinctive face is more likely to fall inside of the bounds of its attractor field, yielding more difficult discrimination.

More recent evidence for the attractor fields came from Laurence et al. (2016) investigating the ORE. Previous research discussed above framed the ORE as a problem of discriminating other-race faces with the assumption that other-race faces are encoded using dimensions best suited to discriminate own-race faces, which are inappropriate for encoding other-race faces. This leads other-race faces to be encoded in a region with high exemplar density (Figure 1.10), making them more confusable with one another. The greater density of exemplars surrounding other-race faces implies that they will have smaller attractive fields, making a smaller range of within-person variability to be tolerated. In Laurence et al.’s study, subjects were given a card sorting task to test for tolerance to within-person variability. The cards included multiple ambient photos belonging to two unfamiliar own- or other-race identities, and the subjects were asked to sort these cards into piles so that each pile corresponded to an identity. The authors found that while the subjects created about 7 piles for own-race faces, they created about 9-10 piles for other-race faces, indicating that subjects perceived there to be more identities in the other-race set. Importantly, subjects made few identification errors, meaning that they rarely put two different identities in a single pile for both own- and other-race faces, which did not differ significantly. These results confirmed that attractor fields are smaller for other-race faces, and that it is not necessarily the case that other-race faces “all look alike”, but rather they “all look different”.

1.2.2 Theoretical Limitations of the Face Space: Differences Between Familiar and Unfamiliar Face Processing

When we assume faces are represented as single points or vectors, it becomes difficult to explain how faces can be recognized across different viewing conditions. Although the Voronoi model (Lewis & Johnston, 1999) or the attractor field model (Tanaka et al., 1998) can provide an account for many views of a face activating the face’s representation, there are certain limitations with the underlying assumptions of these models. Both models assume that the size of a region surrounding a face representation is determined by its distinctiveness. In a normally distributed Voronoi diagram, regions that are clustered at the centre (i.e., typical faces) are necessarily smaller than regions that have their sites on the “edges” of the space (i.e., distinctive faces; see Footnote 1). Similarly, the size of an attractor field is determined by the density of surrounding representations, making distinctive faces have larger attractor fields than typical faces. A larger region surrounding a representation is indicative of how well different instances of a face can be recognized, because a larger region will map more different inputs onto the same identity than a smaller region. However, there are likely factors other than distinctiveness that determine how well within-person variability can be tolerated, and one such factor is familiarity with a face. In fact, there is evidence that processing identity of familiar and unfamiliar faces is qualitatively different (e.g., Megreya & Burton, 2006; 2007), therefore, conflating familiar and unfamiliar faces in a single space fails to account for these differences.

Although unfamiliarity disadvantages have been demonstrated previously for recognition paradigms (e.g., Ellis et al., 1979), more recent research shows that there are also unfamiliarity disadvantages in *perception* tasks such as simultaneous eye-witness lineup paradigms (e.g., Bruce et al., 1999; Megreya & Burton, 2006; 2007). For example, in Bruce et al.'s study (1999) subjects were shown lineups as shown in Figure 1.14 below where the target could be present or absent in the lineup, and the subjects were asked to choose the target face (at the top) from the lineup. Subjects only reached about 70% accuracy in both target-absent and target-present lineups. This is particularly interesting given that the task was untimed, it had no demands on memory (i.e., simultaneous matching), subjects knew the target was absent in about half of the trials, and viewpoint and expression were matched across faces. Further, performance got worse as viewpoint and expression differed between the target and the lineups, suggesting that unfamiliar face matching is highly image-dependent.

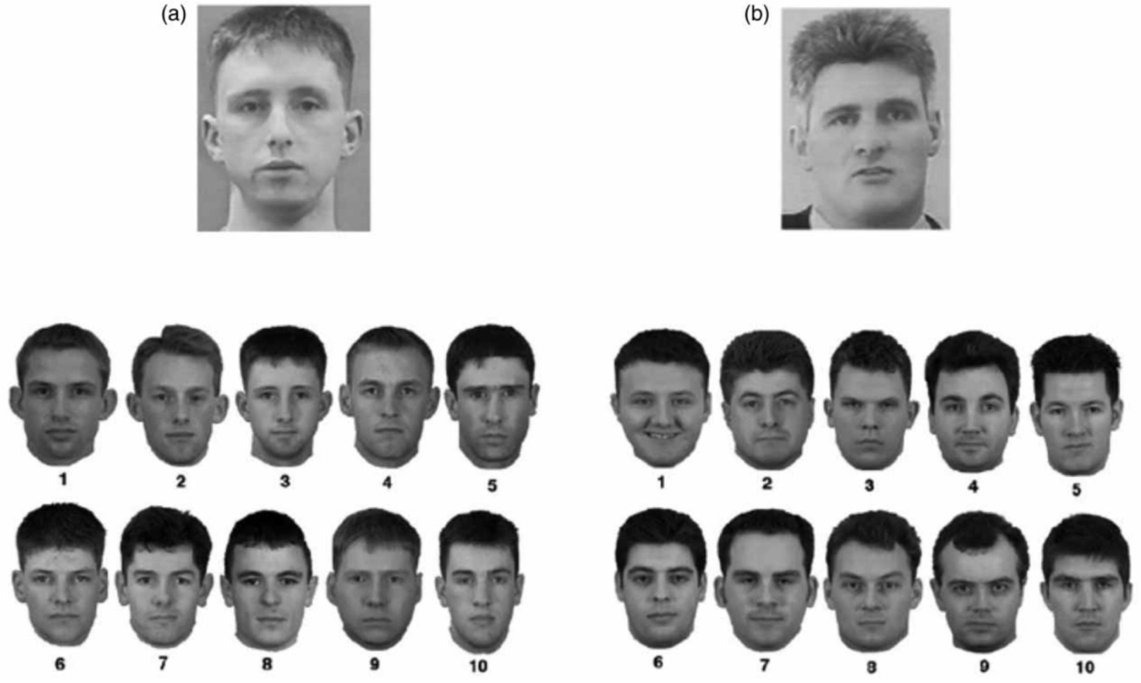

Figure 1.14. Example lineup task used in Bruce et al. (1999), screenshot obtained from Burton (2013). Can you match the target (on the top) with one of the faces in the lineup?

Answers: a) 3, (b) target-absent

Low accuracy observed in this task is unlikely to be due to task demands. In Megreya and Burton's (2007) study, subjects were only shown two images in a same/different judgment task (i.e., a matching task) where two images either depicted the same person or different people. They found the same pattern of accuracy (about 70% correct). Therefore, the issue seems to be not task demands, but rather encoding unfamiliar faces. Familiar face matching, on the other hand, leads to almost perfect performance. In Megreya and Burton's (2006) study, even when subjects were briefly familiarized with some faces their accuracy reached above 87%. In Bruce et al.'s study (2001), subjects were able to match familiar faces with above 90% accuracy, even from low quality CCTV videos. This suggests that familiar and unfamiliar face matching rely on qualitatively different processes. In fact, Megreya and Burton (2006), in their paper titled "*Unfamiliar faces are not faces*", showed that unfamiliar face matching accuracy is highly associated with *inverted* familiar face matching accuracy. Since inversion disrupts configural processing (e.g., Diamond & Carey, 1986), this suggests that configural information cannot be extracted from unfamiliar faces to the same extent. Therefore, while familiar face matching utilizes a stored memory representation, unfamiliar face matching is highly dependent on pictorial cues.

Why do these differences arise when processing familiar and unfamiliar faces? The answer seems to be variability across different views of the same face (Figure 1.15). In a very influential study by Jenkins et al. (2011), subjects were given a card sorting task, including two different identities with 20 photos each. It is important to note that these photos were *ambient*, meaning that viewing conditions were not controlled; lighting, expression, viewing angle, and so on were unsystematically varied. Subjects were asked to sort these cards into piles where each pile represents one identity. They found that unfamiliar observers created a median of 7.5 piles, significantly higher than the correct answer which was two piles. Importantly, misidentification errors were quite low - subjects almost never placed two different identities in the same pile (fewer than 1% of trials). When the same task was given to a set of observers that were familiar with the identities in the card decks, performance was almost perfect. These results suggest that different views of a face involve high variability, and when a face is unfamiliar, variability is not effectively tolerated. This leads observers to attribute changes in pictures of a face to changes in perceived *identity* of the face, similar to what was found in Laurence et al.'s (2016) study with other-race faces using the same task.

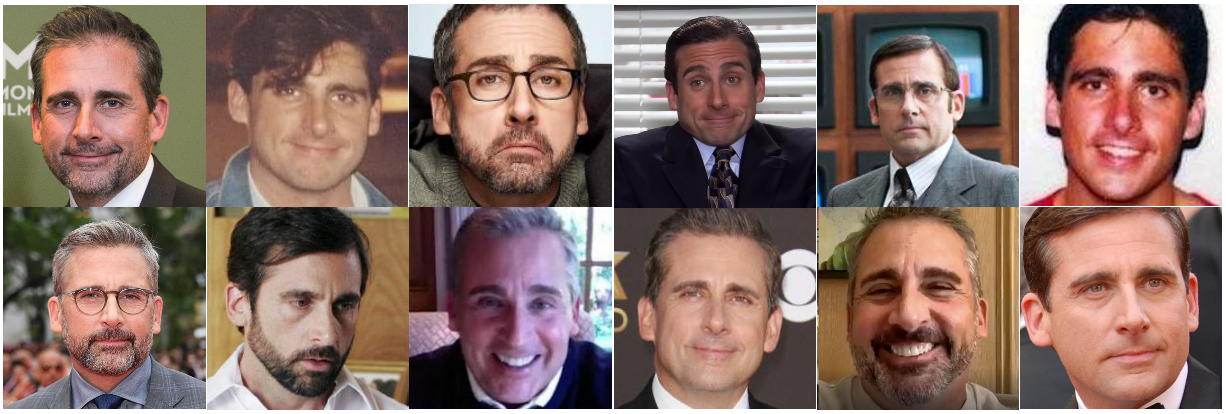

Figure 1.15. Different images of Steve Carrell across different years, emotional expressions, camera, viewing angle, hairstyles and so on. One can easily observe that photos depicting the same person can highly vary. Yet, for familiar faces, this variability is readily tolerated and the face can be successfully recognized. For an unfamiliar viewer, however, it is with great difficulty these photos are judged as depicting the same person.

Familiarity then, can be defined as tolerance to within-person variability. Given that familiarity with a face does not transfer to familiarity to another face (Burton, 2013), the ways in which a face varies are a *unique* property of that face. Computational support for this idiosyncratic variability came from Burton et al.'s (2016) study. In their study, they employed a PCA on multiple ambient photos belonging to the same person. A PCA returns components arranged by how much variance is accounted for by each component in a descending order, so that the first component accounts for most of the variance in the data, the second component accounts for the second-most variance, and so on. This allows for observing the variability captured in a set of high-dimensional data, in this case the variability captured by different views of a person. Using PCA, the authors found that while the first three components for each identity commonly captured dimensions like head angle and lighting, starting from the fourth component of the PCA, every identity varied in a unique way. For example, if the fourth component captured a smile for someone, it captured eye movement for another person. Even when there were overlaps in the components between individuals, the *range* over which these components varied was idiosyncratic; for example, a 10° head turn for someone with a longer nose would cast different shadows than someone with a smaller nose.

It should be noted that within-person variability is not *entirely* idiosyncratic. In a recent paper, Kramer et al. (2018) showed that familiarity with a face can partly benefit unfamiliar face matching performance[3]. Additionally, Burton et al. (2016) found that the first three components of the PCA were similar across different faces, although how a face varied within those components still reflected idiosyncrasy, the most variance captured by the PCA was common across different faces. This is also consistent with Megreya and Burton's (2006) finding that inversion also disrupts unfamiliar face matching, suggesting that some configural information is being extracted from unfamiliar faces. If unfamiliar face matching was entirely feature-based, accuracy matching upright and inverted unfamiliar faces would be the same. Therefore, it is not surprising that some variance would perhaps not be shared across *all* faces to the same extent, but rather across some faces to some extent. This raises an interesting question regarding whether and to what extent the encoding dimensions are shared in the face space. It could be that a layer of the face space represents common variances across different faces that serve to discriminate faces, and that a layer with identity subspaces serves to tolerate within-person variability. A similar mechanism was suggested by Benson and Perret (1991) discussed above (p. 23).

Since there is a large discrepancy between processing familiar and unfamiliar faces due to idiosyncratic variability, some researchers argued that we are not experts at processing faces, only *some* faces, the ones that are familiar to us (Burton, 2013). Whether we are experts at face

[3] In their study, they used PCA combined with Linear Discriminant Analysis (LDA), which combines a more statistical, bottom-up approach (PCA) with a more high-level, top-down approach (LDA) through labeling different "classes" (i.e., identities).

processing or whether faces are special has been a long debate in the literature (e.g., Diamond & Carey, 1986; Farah et al., 1998; Rossion, 2018). On one hand, faces, despite their homogeneity, can be discriminated fairly well. Importantly, faces are more vulnerable to inversion than other classes of perceptual stimuli (e.g., Yin, 1969), suggesting that faces are "special". Diamond and Carey (1986) argued that expertise with any homogeneous class of perceptual stimulus makes it vulnerable to inversion, suggesting that faces are special only in the sense that we are more likely to become experts in face processing than experts in other classes of stimuli. On the other hand, more recent literature shows that although faces share homogeneity across different identities based on their general configuration, one person's face involves a high degree of variability that is at least to some extent idiosyncratic (Burton et al., 2016; Jenkins et al., 2011). This variability makes processing unfamiliar faces error-prone, because the unique variability of an unfamiliar face is not learned and thus cannot be tolerated across viewing conditions.

Rossion (2018) argued that face expertise should be considered through comparing typical human observers with populations or species that are not experts at face recognition. For example, patients with prosopagnosia perform worse than typical subjects in unfamiliar face matching tasks, suggesting that the ability to recognize faces assists with unfamiliar face matching. This is similar to what was found in Kramer et al. (2018) using PCA and LDA to model face processing. A recent paper by Blauch et al. (2021) using Deep Convoluted Neural Networks (DCNN) to create object- and face-expert models found that face-expert models outperformed object-expert models in unfamiliar face matching tasks, suggesting that familiar and unfamiliar face processing both tap into a shared expertise. However, such benefits to unfamiliar face matching, although useful, are nonetheless a result of learning familiar identities, and not an ability developed to deal with unfamiliar faces (Young & Burton, 2021).

So, are we face experts? The question itself is somewhat problematic. Treating faces as a special case of object recognition, or treating *all* faces as a unified class of perceptual stimulus does not sufficiently capture the differences between familiar and unfamiliar face processing (Burton, 2013). This has important implications for the face space, which assumes all familiar and unfamiliar faces are encoded in a single space. However, if we assume variability to be idiosyncratic, different identities should be encoded using *idiosyncratic* dimensions. Burton et al. (2016) showed that a PCA model trained with a person's face made fewer errors reconstructing an untrained image of that person's face than reconstructing an untrained identity. In other words, "learning" the variability of someone's face, while useful for recognizing that face, led to making more errors recognizing someone else's face. Therefore, it is not merely the case that faces are represented as points, vectors, or even regions, but rather that faces have their very own subspaces with their own set of dimensions. Indeed, it has been found that computational models that utilise face-specific subspaces outperform traditional models where a single space is shared by all trained faces (Aishwarya & Marcus, 2010; Shan et al., 2003).

1.2.3 Methodological Limitations of the Face Space: Exemplar-dependent Recognition

Until the last decade or so, most work done on face processing, including the face space framework, has overlooked the importance of within-person variability. In fact, such variability was purposefully excluded as "noise" (Jenkins et al., 2011). This led to researchers to use images taken under highly controlled settings, with limited variability in the person (e.g., expression, weight, age), and in the image-dependent contexts (e.g., viewpoint, lighting, camera; Burton et al., 2011). This is dissimilar to how we process faces in the real world, which involves encountering a range of variability in both person- and context-dependent aspects. Since this variability is a key aspect of face processing, minimizing variability in experimental settings ends up treating face processing as a special case of image processing (Burton, 2013). However, many recognition studies used the same image during study and test (e.g., Bartlett et al., 1984; Cohen & Carr, 1975; Cross et al., 1971; Ge et al., 2009; Goldstein & Chance, 1985; Hancock & Rhodes, 2008; Light et al., 1979; Ng & Lindsay, 1994), or an image that varies in only a single dimension (expression, e.g., Byatt & Rhodes, 2004; Chiroro & Valentine, 1995; Valentine, 1991, viewpoint, e.g., Hagen & Perkins, 1983; Newell et al., 1999). Since unfamiliar face recognition relies on unsophisticated image-dependent comparisons (e.g., Megreya & Burton, 2006), it is imperative to utilize within-person variability through incorporating natural (ambient; Jenkins et al., 2011) images that we encounter in the real world to have a more complete understanding of how we process faces.

Another limitation from previous research comes from the assumption that a single photo is a good representation of a person's face. For example, when collecting ratings to measure facial properties like distinctiveness, only one photo of a person is used. However, when variability is incorporated, different views of a face may elicit different judgments. For example, Jenkins et al. (2011) collected attractiveness ratings for multiple ambient photos of unfamiliar celebrities, and found that variability in attractiveness ratings for one person was greater than variability in attractiveness ratings between people. Similarly, they found that likeness ratings, commonly used as a dependent variable to measure the caricature effect (e.g., Benson & Perrett, 1991; Rhodes et al., 1987), also varied greatly across different photos of a familiar person's face. These results suggest that properties that we may assume are inherent to a face may be captured differently across different photos.

The aim of the current research program is to address the above-mentioned theoretical and methodological limitations in hopes of understanding how familiarity is represented in the face space. Research on the face space was primarily concerned with how faces are discriminated, which is a curious problem given that faces share great similarities in their general configuration. Such research was motivated by consideration of faces as a homogeneous category of stimuli while focusing on the nuances among different members of this category. However, while there are similarities across faces, there are great variances within different views of a single face. Therefore, another very curious problem in face recognition is how different views of a face are amalgamated into a robust, stable mental representation that can be used to successfully recognize that face despite such variability. I have argued above that familiarity with a face can be defined as tolerance to its within-person variability. Thus, to investigate how familiarity is represented in the face space, I will first explore how within-person variability is represented in the face space for familiar and unfamiliar faces (Chapter 2). Then, I will examine how a face's representation in the face space might change as a face becomes familiar (Chapter 3). Lastly, I will argue for representing familiar faces as identity-specific subspaces while exploring whether these subspaces use norm-based coding (Chapter 4). Together, this research program aims to reconcile the face space theory with newer approaches to face familiarity that emphasize the importance of exposure to within-person variability. In doing so, I hope to propose a comprehensive theory of face recognition that can both explain how the visual system discriminates different faces while successfully tolerating within-person variability for familiar faces.

2 How is Within-person Variability Represented in the Face Space?

2.1 Introduction

In the traditional models of the multidimensional face space, each face is encoded in a location in the space either as a point or a vector according to its perceived characteristics

(Valentine, 1991). Since the dimensions of the face space are unknown, an important factor that provides information about where a face is encoded is the face's distinctiveness. Distinctiveness, by definition, is determined by the inter-item similarity of the faces in the face space. Distinctive faces are assumed to be *dissimilar* to other faces, and typical faces are assumed to be *similar* to other faces and to each other. Indeed, Light et al. (1979) showed that inter-item similarity ratings for pairs of faces agreed with distinctiveness ratings; that is, faces that were rated as dissimilar when paired with other faces were also rated as more distinctive. Therefore, distinctiveness, which is structurally based on inter-item similarity, is a fundamental property of the face space (Valentine, 2001), especially given that the face space, like other multidimensional perceptual spaces (e.g., color; Shepard, 1962), is similarity-based.

The distinctiveness of a face can be measured in many ways. For example, Valentine & Bruce (1986a) asked subjects to rate how easy it would be to pick out the person from a crowd, Light et al. (1979) asked subjects to rate the person's similarity to a typical high school male student, and Cohen and Carr (1975) asked subjects to simply rank the faces by distinctiveness. More indirect measures of distinctiveness can be collected through (context-free) familiarity and memorability ratings, which have been shown to capture two orthogonal components of distinctiveness Specifically, lower familiarity and higher memorability predicts higher distinctiveness and vice versa (Vokey & Read, 1992). Another more indirect way to measure distinctiveness is to collect inter-item similarity ratings on pairs of faces (e.g., Johnston et al., 1997; Lee et al., 2000; Light et al., 1979). Faces that are rated as more dissimilar to their pairs are more distinctive. These studies consistently found superior recognition performance for faces that are more distinctive, indicating that they all measure the same psychological construct of typicality.

One limitation of these studies, however, is that distinctiveness is measured using a single photo of a person, typically taken in highly controlled conditions. For example, Valentine (1991) collected distinctiveness ratings on photos of faces displaying a neutral expression from a frontal view, and Light et al. (1979) and Vokey & Read (1992) used high school yearbook photos. However, one photo is rarely, if ever, a good representation of a person, because a person's face varies greatly across different photos (Jenkins et al., 2011). In line with within-person variability, attributes like attractiveness (Jenkins et al., 2011) or judgments of likeness (Jenkins et al., 2011; Ritchie et al., 2018) also vary across different photos of a person. In fact, Jenkins et al. (2011) found that variability in attractiveness judgments within a person was greater than variability in attractiveness judgments between people, suggesting that attractiveness is as much a photo-centred property as it is a person-centred property (Burton, 2013).

There is no a priori reason to assume that the same would not hold for distinctiveness judgments. Aspects that make a face distinctive are likely to be present to different extents across different viewing conditions, which is especially intriguing considering that most faces are likely distinctive in at least one dimension. Evidence for this came from Burton and Vokey (1998) who explored the typicality paradox of the face space. Usually, the face space is illustrated as a two-dimensional Cartesian space for ease of interpretation (e.g., see Figure *1.1*, Figure 1.4, Figure 1.6, Figure 1.10). In a normally distributed two-dimensional Cartesian space, most faces are clustered near the central tendency, which is described as containing typical faces that resemble most previously encountered faces. However, in studies where typicality ratings are collected, most faces do not end up being rated as "highly typical", which is counterintuitive considering that most faces are assumed to be typical. Burton and Vokey (1998), however, showed that this is an artefact of conceptualizing the face space in two dimensions. When the normal distribution of exemplars in a space with more than two dimensions is modelled mathematically, it actually appears that the face space is shaped like a "donut"[4], where highly typical faces are quite rare. This makes more intuitive sense considering the possibly high number of dimensions of the face space (e.g., Lewis, 2004 speculated there to be 22 dimensions), where most faces would likely be distinctive in at least one dimension instead of most faces being highly typical in all dimensions.

Since the face space is multidimensional and the dimensions are unknown, it is unclear what exactly makes a face distinctive. Although the idea that the distinctiveness of a face stems from an isolated distinctive feature or a set of features is intuitive, this idea received little support from the literature. For example, Valentine and Bruce (1986b) found that subjects are slower when classifying distinctive faces as faces. If a distinctive feature made a face distinctive, distinctive faces would be classified as faces faster, because the salient distinctive feature would signify to the cognitive system that the stimulus is indeed a face. Similarly, Tanaka et al. (1998) found that distinctive faces are vulnerable to inversion, which would not be expected if distinctive faces were encoded by an isolated feature. Assuming inversion disrupts second-order configural processing (i.e., relations between features; Diamond & Carey, 1986), the distinctiveness of a face may be an artefact of its configuration. However, the second-order configuration of a face is vulnerable to within-person variability, considering even a simple change like camera lens distortion can greatly alter a face's configuration (Figure 2.1; Burton, 2013; Sandford & Burton, 2014). Therefore, assuming most faces are somewhat distinctive in at least one aspect (Burton & Vokey, 1998), what constitutes distinctiveness likely varies across different faces. Because faces themselves vary across different views, it becomes highly likely that the perceived distinctiveness would differ across different views of a face. For example, a

[4] From my conversations with Dr. R. Kramer.

face with a long nose (i.e., distinctive by an isolated feature) may not appear as distinctive from a frontal view but may appear quite distinctive from a ¾ view, or a face that has a long inter-eye distance (i.e., distinctive by the second-order relational features) may appear distinctive from a frontal view but not as much from a ¾ view.

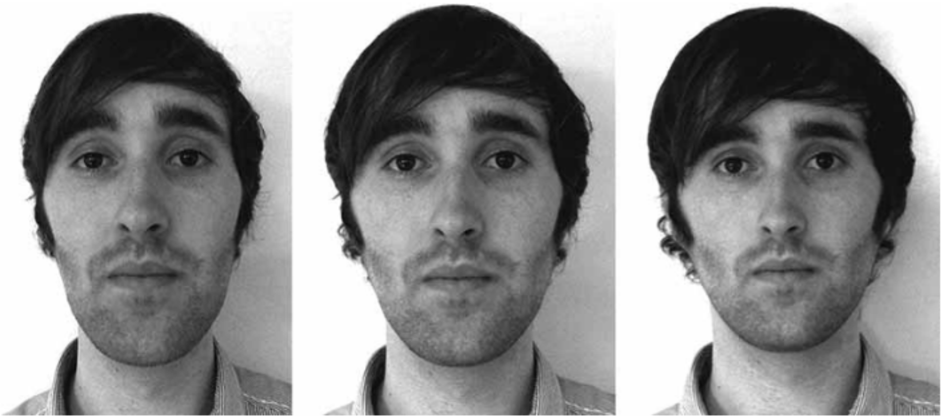

Figure 2.1. Face images taken from different distances, taken from Burton (2013). As can be seen, the second-order configuration of a face, often expressed as the metric distance between features, changes depending on how much the camera lens distorts the face.

The fact that each face varies idiosyncratically leads to the prediction that familiar and unfamiliar faces are processed differently, because the range of idiosyncratic variability across different views of a face is not learned for unfamiliar faces. Therefore, while familiar faces can be successfully recognized across a range of viewing conditions, unfamiliar face recognition is error-prone (e.g., Bruce et al., 1999; 2001; Jenkins et al., 2011; Megreya & Burton, 2006; 2007). This is because familiar face processing relies on an abstract, stable representation in memory, which makes familiar face processing robust to changes in viewing conditions because a novel encounter can be compared to a memorial representation. On the other hand, unfamiliar face processing is highly image-dependent, which makes it vulnerable to such changes. Besides influencing performance in a range of recognition and perception tasks, familiarity also influences social judgments. For example, judgments like attractiveness, trustworthiness, and dominance are more variable across different photos of an unfamiliar face compared to a familiar face (Mileva et al., 2019), and attractiveness judgments become more consistent as familiarity with a face increases (Koca & Oriet, 2023), suggesting a shift from rating "photos" to rating "people".

The perceived distinctiveness of a face is also influenced by familiarity. Faerber et al. (2016) asked observers to rate familiar and unfamiliar faces and their corresponding anti-faces, which are equidistant from the norm and therefore assumed to be equal in typicality. They found that while unfamiliar faces and their anti-faces were rated to be equally typical, familiar faces were perceived to be more distinctive than their anti-faces. These findings suggest that the face space increases the distance of a familiar face from other faces as it becomes familiar, which in turn increases its distinctiveness[5]. Relatedly, Kramer et al. (2018) found that as a computational model became more familiar with a face (i.e., increased number of training images), the distance between the target face and the nearest non-target face increased. It is unclear, however, how the distinctiveness ratings would *vary* across different photos of a face as a function of familiarity, since Faerber et al. (2016) collected their ratings on a single image per target identity.

The aim of the current study outlined in this chapter is to investigate how within-person variability is represented in the face space. Exposure to within-person variability is imperative to become familiar with a newly encountered face (e.g., Burton et al., 2016; Corpuz & Oriet, 2022; Ritchie & Burton, 2017), therefore, it is important to understand how the face space represents

[5] Valentine and Bruce (1986b) found that familiar faces were classified as faces slower than unfamiliar faces, similar to how distinctive faces are classified as faces slower than typical faces. Speculatively, the main effect of familiarity found in this experiment could be a result of increased distinctiveness of familiar faces.

within-person variability as a first step towards understanding how familiarity is accommodated in the face space. Since distinctiveness can be considered as an index of a face's location in the face space, representation of within-person variability will be explored by measuring distinctiveness through collecting inter-item similarity ratings on pairs of images containing different views of faces.

Collecting inter-item similarity ratings is beneficial for a few reasons. First, inter-item similarity ratings can provide information about a face's distinctiveness (Lee et al., 2000; Light et al., 1979), for example, if a face is repeatedly rated as highly similar to other faces in the stimulus set, it is considered typical, and if it is repeatedly rated as highly dissimilar to other faces in the stimulus set, it is considered distinctive. Second, inter-item similarity ratings can be subjected to a Multidimensional Scaling (MDS) procedure, which is a statistical method for mapping out the stimuli as locations in a multidimensional space based on their perceived similarity such that similar items are located near each other, and dissimilar items are located further apart (Hout et al., 2013; Jaworska & Chupetlovska-Anastasova, 2009). MDS has been previously used to construct face spaces to investigate the distinctiveness and caricature effects (Lee et al. 2000), and the other-race effect (Byatt & Rhodes, 2004).

The following sections of this chapter will outline the proposed methods and analyses to understand how within-person variability is represented for familiar and unfamiliar faces in the face space, along with specific predictions. I believe this work will provide important insights into how the inter-item similarity of different faces, and different views of a single face form the structural basis of the face space. Importantly, these insights will be evaluated in the context of face familiarity, which has been robustly shown to be an essential aspect of how faces are processed.

2.2 Proposed Methods

2.2.1 Participants

To my knowledge, there is no specific sample size recommendation to carry out an MDS procedure. Therefore, I ran a simple simulation to determine the number of subjects to be recruited for the study using R (R Core Team, 2021). The rationale is to obtain the number of ratings required for a given pair in the stimulus set (i.e., a single trial) to achieve a reasonable consistency across participants (i.e., the point of saturation). The simulation represents a worst-case scenario in which every subject gives a random rating to a given trial on a scale of 1-7, which is what will be used in the study. After the second subject, a standard deviation of the two ratings is calculated and stored in a list. After the third subject, a standard deviation of all three ratings is calculated and stored in the same list, and so on for 200 subjects. The standard deviation of the accumulated ratings for 200 subjects were then plotted, and this simulation was run 1000 times, each representing a single experiment. The resulting plot is shown in Figure 2.2.

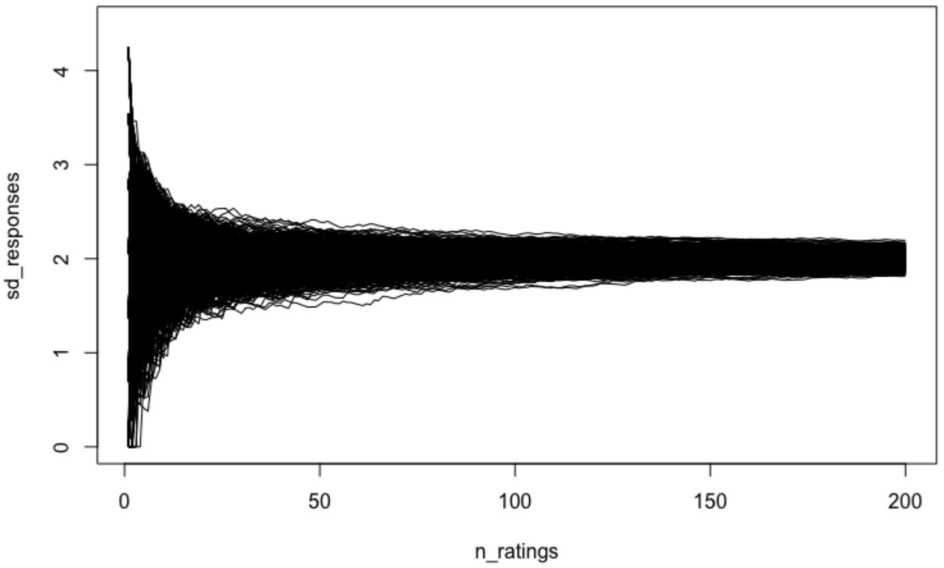

Figure 2.2. The standard deviation of random responses with each additional rating. Each line on the graph represents one iteration of the simulation.

To find the point of saturation, I then calculated the range of standard deviation (across 1000 experiments) at each number of subjects along the x-axis, and then plotted these ranges (Figure 2.3). Since the goal is to obtain a number where an additional rating would not significantly change the overall consistency of the ratings, I calculated the elbow point of this scree plot using the smerc package (version 1.8.2; French, 2023). For the simulation demonstrated here, the elbow point was at 21, suggesting 21 subjects would be sufficient to achieve the point of saturation assuming all observers give a random response. Therefore, I reasoned that 25 ratings per trial would be adequate.[6] Since the experiment will be run in two subsets (see below), I aim to recruit 50 subjects.

[6] I ran the simulation a few times, and the elbow points were often in between 20-30, therefore 25 seemed like a reasonable number. It is also similar to what was used in Lee et al.'s (24 subjects; 2000) and Byatt & Rhodes' (22 subjects; 2004) papers that used MDS.

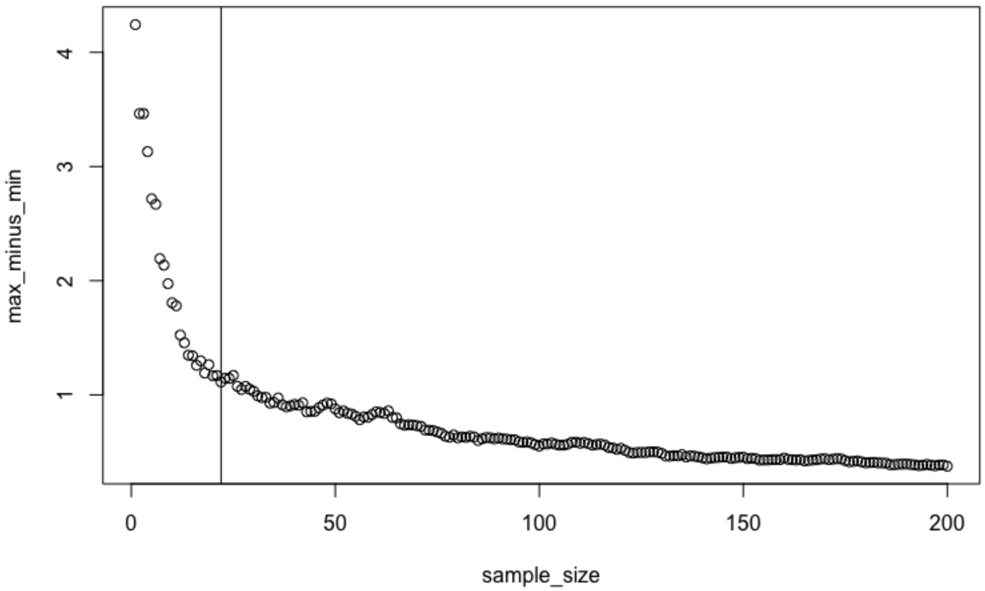

Figure 2.3. The ranges of the standard deviation (i.e., the thickness of the graph illustrated in Figure 2.2) as a function of each number of ratings that are added. The line vertical line indicates the elbow point of the curve.

Subjects will be recruited through the University of Regina Participant Pool. All subjects will have normal or corrected-to-normal vision. Subjects who do not meet the familiarity criteria will be excluded from the analysis. All procedures will be carried out in accordance with the Canadian Tri-Council Policy Statement on the ethical treatment of research participants and will be approved by the University of Regina Research Ethics Board. All subjects will sign a consent form prior to their participation, and will receive compensation.

2.2.2 Stimuli and Apparatus

Six photos will be collected for each of the eight target identities (half of them familiar, and half of them female) from the Internet. All images will belong to celebrities from abroad to ensure that multiple high-quality ambient photos of the targets are available (i.e., high within-person variability). Identities will be chosen to have similar demographics (e.g., age, ethnicity) to prevent similarity judgments to be made solely on demographics. None of the faces will be obstructed with objects like hands or glasses. All images will be cropped to include only the face. For each identity, an average face will be created using webmorphR (DeBruine, 2022), which is an R package that is used for landmarking and transforming faces, yielding a total of seven images per identity. To carry out an MDS procedure, it is recommended that each stimulus to be shown to the participants is paired with every other stimulus in the set without repetition. However, the pairs will only contain male or female faces to prevent subjects from making their judgments on the basis of gender. Previous studies that used inter-item similarity ratings (two of which used MDS; Byatt & Rhodes, 2004; Lee et al., 2000; Light et al., 1979) only used male faces, however, in this study both male and female faces will be used. Therefore, for each gender, 28*27/2 pairs will be used, which equals to 378 pairs, yielding a total of 756 pairs. Since presenting subjects with 756 pairs is not feasible in one session, the pairs will be randomly divided into 2 subsets, yielding 378 trials per subset. Each subject will view one of the subsets, and this will be counterbalanced across participants. These subsets will then be concatenated for MDS. The experiment will be programmed using jsPsych (de Leeuw, 2015) and the data will be collected online using Pavlovia. All data handling will be carried out using R (R Core Team, 2021), including data cleaning, analyses, and plotting.

2.2.3 Procedure

Subjects will be shown two images at a time, equidistant from the centre of the screen, and will be asked to rate the similarity between the two faces on a Likert scale ranging from 1-7 (1- strongly dissimilar; 7- strongly similar; Figure 2.4). The order of image pairs will be randomized across participants, and the order of the two subsets (see above) will be counterbalanced. Subjects will be asked to not base their ratings solely on identity, to prevent them from giving really low ratings for pairs that depict different identities, and really high ratings for pairs that depict the same identity. The instructions will be as follows:

> *"Please rate the similarity between the faces on the screen. Sometimes the faces will belong to the same person and sometimes they will belong to different people. Keep in mind that two photos of the same person can look very dissimilar, and two photos of different people can look very similar. Therefore, do not try to base your ratings solely on whether the photos seem to be of the same person, but rather on how similar the photos of faces look."*

At the end of the experiment, subjects will be asked whether they were familiar with any of the faces used in the study. Subjects who are unfamiliar with the familiar targets, and familiar with the unfamiliar targets will be removed from the analysis.

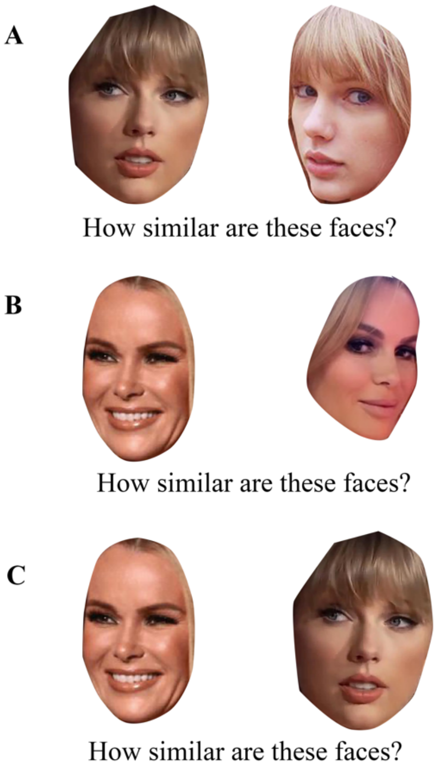
A
How similar are these faces?
B
How similar are these faces?
C
How similar are these faces?

Figure 2.4. Example pairs of (A) two familiar faces, (B) two unfamiliar faces, (C) a familiar and an unfamiliar face. The slider is not included in the illustration.

2.3 Proposed Analyses and Hypotheses

2.3.1 Multi-Dimensional Scaling

Ratings will be averaged across participants for each image pair, and then a distance score will be calculated by subtracting the ratings from 8 (i.e., 8 - $M_{\text{pair similarity}}$), which will then be used to create a distance matrix. A distance matrix is a symmetrical matrix that contains the name of each stimulus (i.e., photo) in rows and columns where their intersections indicate their perceived dissimilarity. It is symmetrical in the sense that rows and columns that contain the same stimulus will have a distance of "0" which fall diagonally along the matrix with the same values mirrored on each side. This dissimilarity matrix will be included in the MDS analysis using the ALSCAL algorithm (Takane et al., 1977) using the smacofx package in R (Rusch et al., 2023).

The primary objective of an MDS solution is to find the best fit using the smallest number of dimensions (Hout et al., 2013; Jaworska & Chupetlovska-Anastasova, 2009). The number of dimensions to include in the analysis will be selected by creating a scree plot of stress[7] values for each possible number of dimensions. Hout et al. (2013) recommend selecting the number of dimensions that correspond to the "elbow" of the plot, which often creates a balance between low stress (i.e., conflict) and the interpretability of the MDS solution. Kruskal and Wish (1978) recommend that stress should not exceed 0.15. The interpretability of the MDS solution is subjective in nature; however, the coordinates of the items from the MDS solution can be analyzed using more traditional analyses like regression.

[7] Collecting inter-item similarity ratings may create certain conflicts. For example, if a subject rates images A and B as highly dissimilar, and images A and C as highly similar, in a two-dimensional space B and C should also be highly dissimilar since B is similar to A and A is dissimilar to C. However, since subjects see two images at a time, they may rate B and C as similar as well, which creates a conflict that can be measured with a stress function (Hout et al., 2013).

2.3.1.1 Within-person Clusters. I will investigate the exemplar density of different views of the same face, and to what extent this exemplar density is modulated by familiarity. To do so, the MDS solution will be visually inspected to see how different exemplars of the same faces are clustered. Additionally, a mean distance of each face to all other exemplars of the same identity will be calculated, with higher numbers indicating higher distance and therefore lower exemplar density. Using familiarity as an independent variable and the mean distance as a dependent variable, a linear model will be fit to the data to observe the exemplar density for familiar and unfamiliar faces. I expect that exemplar density will differ as a function of familiarity.

One possibility is that inter-item similarity will be less variable for familiar faces than for unfamiliar faces, creating higher exemplar density clusters for familiar faces. I will call this the *representation anchoring hypothesis*. According to this hypothesis, since familiar face processing relies on an abstracted representation in memory, this representation would serve as an anchor for making inter-item similarity judgments such that identity-specific information would be prioritized over image-specific information. This was consistently shown for matching tasks (i.e., same/different tasks), where two photos of a familiar face are matched with almost perfect performance, whereas unfamiliar face matching is highly error-prone (e.g., Clutterbuck & Johnston, 2004; Megreya & Burton, 2006; 2007). In turn, any two photos of a familiar face would bear higher resemblance to each other than any two photos of an unfamiliar face. This was found for social judgments like attractiveness, trustworthiness, and dominance (Mileva et al., 2019), where familiar faces received more consistent ratings than unfamiliar faces. Therefore, the overall inter-item similarity might be higher for familiar faces, creating denser clusters in the MDS space. Further, tolerance to within-person variability is worse for unfamiliar faces (e.g., Jenkins et al., 2011); that is, two photos of an unfamiliar face can look like entirely different people, leading any two photos of an unfamiliar face to be rated as highly dissimilar, creating lower exemplar density clusters for unfamiliar identities.

Another possibility is that familiar faces will have lower exemplar density clusters. I call this the *image anchoring hypothesis.* Since familiar faces rely on an abstract representation in memory, subjects would be better able to discern identity-specific and image-specific characteristics of familiar image pairs. Since identity-specific characteristics of familiar faces will appear constant across two photos of the same person, image-specific differences would be more easily discriminable. In turn, two photos of a familiar face may appear more dissimilar than two photos of an unfamiliar face, where image-specific characteristics are not as effectively extracted. This is similar to the notion of "face likeness", where different photos of a person capture that person's appearance to varying degrees (Jenkins et al., 2011), and likeness judgments vary idiosyncratically across viewers (Ritchie et al., 2018). In turn, two photos of a familiar face may be judged as more dissimilar than two photos of an unfamiliar face because they likely do not capture the person's likeness to the same extent. Naturally, two photos of an unfamiliar face may also capture that face's likeness to a different extent. However, since face likeness, by definition, relies on an internal representation to be used as a reference (Ritchie et al., 2018), it would not be relevant to unfamiliar observers (i.e., observers cannot tell whether a photo is a good or bad likeness if they are unfamiliar with the person shown).

2.3.1.2 Between-people Clusters. I will also observe how clusters for each identity are located in relation to one another as a function of familiarity. To do so, the MDS solution will be visually inspected. Additionally, the mean distance of each exemplar to all other exemplars of different identities will be calculated, and will be analyzed with a linear model with familiarity as the independent variable. I expect that familiar faces will form more distinct clusters than unfamiliar faces. In other words, familiar faces will be clustered further away from each other and other unfamiliar faces. This is in line with the findings of Faerber et al. (2016) who found that familiar faces were rated as more distinctive than their corresponding unfamiliar anti-faces. Since faces and anti-faces located in regions with similar exemplar density have similar distinctiveness, this finding shows that familiarity shapes the face space so that familiar faces move further away from their near neighbors. A similar finding came from Kramer et al. (2018) in their a PCA + LDA model for face recognition, where the distance between a target face and the nearest non-target centroid increased as a function of familiarity. This is also a prediction that is in accordance with Tanaka et al. (1998)'s attractor field model. In this model of the face space, an attractor field of a face increases with decreasing exemplar density surrounding a face's representation. A larger attractor field indicates that more exemplars can be accepted as that face, allowing for higher tolerance to within-person variability (Laurence et al., 2016). Assuming faces become more distinctive as they become familiar (Faerber et al., 2016), their attractor fields would also get larger, suggesting that they would be located in a region with lower exemplar density. Therefore, I expect that identity clusters of familiar faces will be located further away from each other and from other unfamiliar faces.

2.3.1.3 Where is the Average Face? As mentioned above, an average face will be created for each identity for exploratory reasons to investigate where the average would be located in relation to other exemplars of the same identity. This will be done by visually inspecting the MDS solution, and by calculating the distance between the averages and the centroid of their corresponding identity clusters. This average-centroid distance will be analyzed by fitting a linear model with familiarity as the predictor.

The location of the average was explored in previous studies that used inter-item similarity ratings (Byatt & Rhodes, 2004; Lee et al. 2000) to investigate the norm-based model of the face space, and it was found that the average face was not at the origin. However, the procedure used in these studies created averages that were smoother in texture than exemplar images, possibly contaminating the inter-item similarity judgments. In this study, an average face will be created using a software that has a feature that can average faces without sacrificing skin-like texture (see Materials and Stimuli). Additionally, the averages will be created by averaging within-person images, and their relationship will be investigated in relation to the exemplars of the same identity as an attempt to explore a norm-based model of identity-specific sub-spaces. Previous research has shown that an average representation is extracted from different views of a face (e.g., Burton et al., 2005; Kramer et al., 2015), and refined with increased familiarity (Koca & Oriet, 2023), suggesting averaging can be an underlying mechanism for face familiarity as it can yield a stable representation that can tolerate variability across different views of a face (Bruce, 1994). This raises the question of whether an average face would be at the origin of an identity-specific subspace, perhaps as a reference point for exemplars to be encoded. In that case, I would expect the average-centroid distance to be smaller for familiar faces. Since an average representation would not be present for unfamiliar faces, I have no specific predictions for where the average would be located for unfamiliar faces.

In summary, the inter-item similarity ratings will be analyzed using MDS, and the coordinates returned from the MDS solution will be analyzed using a linear model. Of interest are the exemplar density within identity clusters, the distance between clusters of different identities, and the distance between each identity average and the centroid of the corresponding identity's exemplars. These variables will be evaluated within the context of familiarity to provide insights into how within-person variability is represented in the face space.

2.3.2 Rank-Image and Rank-Identity Correlation

Inter-item similarity ratings have been shown to agree with distinctiveness ratings (Light et al., 1979), such that items that are consistently rated as similar to other items are also rated as more typical, and items that are consistently rated as dissimilar to other items are rated as more distinctive. An average "image-distinctiveness score" will be calculated for each image by subtracting each rating from 8, and then averaging them for each image, so that images that were consistently rated as dissimilar to their pairs will receive a higher image-distinctiveness score, and images that were consistently rated as similar to their pairs will receive a lower image-distinctiveness score. For each identity, an identity-distinctiveness score will be obtained by averaging over the image-distinctiveness scores of each identity. These scores will be used to explore whether the variability in the image-distinctiveness scores contributes to the variability in the identity-distinctiveness scores. To do so, I will follow the procedure of Jenkins et al. (2011), which was used for likeness and attractiveness ratings. The procedure involves comparing two correlations: Rank-Identity correlation and Rank-Image correlation. To calculate the Rank-Identity correlation, the identity-distinctiveness scores (ranging from 1-7) will be arranged in an ascending order, and each identity will receive an identity-distinctiveness rank (from 1-8). The Rank-Identity correlation will be a Pearson correlation between the identity-distinctiveness scores and the identity-distinctiveness ranks. To calculate the Rank-Image correlation, the image-distinctiveness scores (ranging from 1-7) will be correlated with the identity-distinctiveness rank (each image of the person will have the same rank, ranging from 1-8 as above). These two correlations will be compared using Fisher's *z* test to investigate whether they are meaningfully different.

The rationale of this procedure is as follows. The Rank-Identity correlation will naturally yield a very high coefficient, because the rank of an identity is determined by arranging identity-distinctiveness scores. If the Rank-Image correlation is not significantly different from the Rank-Identity correlation (i.e., it also yields a very high coefficient), this suggests that the variability in the perceived image-distinctiveness scores is accounted for by changes in identity, and not necessarily changes in the perceived distinctiveness of different photos of the same person. On the other hand, if the Rank-Image and the Rank-Identity correlations are statistically different, this suggests that the variability in the image-distinctiveness scores is *not* accounted for by changes in identity. Therefore, it can be concluded that different photos of a person capture distinctiveness differently.

Given that people's faces vary highly and idiosyncratically across different viewpoints (e.g., Jenkins et al., 2011), and that most faces are somewhat distinctive on at least one dimension (e.g., Burton & Vokey, 1998), I expect that perceived distinctiveness will vary across different photos of a person as different photos would capture the distinctiveness of a face differently. Therefore, the Rank-Image correlation and the Rank-Identity correlation are expected to be statistically different, suggesting that the variability across the image-distinctiveness scores is not solely explained by identity-specific distinctiveness. In other words, for example, an overall typical face may have a distinctive exemplar and vice versa, emphasizing the importance of within-person variability across different views of a face.

I also expect that some faces will be overall more distinctive than others, as was found in many behavioral studies (e.g., Valentine & Bruce, 1986a; 1986b). This has also been shown more recently using computational modeling where the model recognized some faces better than others (Kramer et al., 2018), suggesting that there are aspects of some faces that make them more memorable than others. Therefore, I expect that identity-distinctiveness scores will vary across different identities. It is difficult, however, to make specific predictions regarding how overall distinctiveness would be predicted by familiarity in this case since familiarity is not manipulated in a cross design. Literature has found that familiar faces are rated as more distinctive (Faerber et al., 2016). However, this was compared to their corresponding unfamiliar anti-faces where the exemplar density of the surrounding region is assumed to be comparable. Although a familiar face might be more distinctive than an unfamiliar face that is equidistant from the norm as the familiar face, it is not clear whether *any* familiar face would be more distinctive than *any* unfamiliar face. Since the initial exemplar density is not known or directly manipulated in this study, I do not have specific predictions about the differences in identity-distinctiveness scores as a function of familiarity. The change in the exemplar density surrounding a face with increasing familiarity will be investigated in the study outlined in Chapter 3.

3 How Does the Face Space Change with Familiarity?

3.1 Introduction

The multidimensional face space theory (Valentine, 1991) provides a model of face recognition that is structurally based in inter-item similarity; that is, faces that are similar to each other are encoded closer together, and faces that are dissimilar are encoded further apart. Accordingly, typical faces look similar to each other and are encoded in a region with high exemplar density, and distinctive faces look dissimilar to other faces and are encoded in a region with low exemplar density. In turn, while typical faces are easily confusable with similar-looking others and hence are recognized slower, distinctive faces do not have many competing representations surrounding them, and therefore are recognized faster (e.g., Valentine & Bruce, 1986a; 1986b). Although this model provides a unifying approach for many phenomena in the literature, like the distinctiveness, caricature, and the other-race effects, as discussed previously, it is limited in its assumption that each face is represented as a single point or vector in the face space. This raises the question of how the visual system can recognize faces across several viewpoints.

More recent models of the face space have incorporated regions surrounding a face that can allow for multiple inputs to activate the face's representation. One such model is Tanaka et al.'s (1998) attractor field model in which a face's attractor field determines the range of inputs that can be accepted as that face. Therefore, as the size of the attractor field increases, more exemplars can be accepted as that face. In this model, the size of an attractor field is assumed to be determined by the exemplar density of the region in which the face representation is located, so that less exemplar density yields a face representation that is surrounded by a larger attractor field and more exemplar density yields a face representation that is surrounded by a smaller attractor field. Accordingly, distinctive faces have larger attractor fields and typical faces have smaller attractor fields.

Tanaka et al. (1998) provided evidence for the attractor field model by presenting participants with 50-50 morphs of typical and distinctive faces, and found that the 50-50 morph was psychologically more similar to its distinctive parent. Since a 50-50 morph is equidistant from its constituents, a traditional model of the face space would predict that the morph would bear equal resemblance to the distinctive and the typical parent. However, since the distinctive parent's attractor field is larger, the morph is closer to the attractor field of the distinctive parent

(Figure 1.12). In turn, the 50-50 morph of a typical and a distinctive face appears more similar to the distinctive face.

More recent evidence, using multiple ambient photos for each target, showed that the attractor field model can account for tolerance to within-person variability. In Laurence et al.'s (2016) study, subjects were given a card-sorting task of own- and other-race faces and were asked to create a separate pile for each identity. They found that subjects incorrectly created more piles for other-race faces than for own-race faces, suggesting that there is less tolerance to within-person variability for other-race faces. According to the attractor field model, since other-race faces are encoded in a region with high exemplar density, they have smaller attractor fields, so two images of the same other-race face are less likely to fall within the same attractor field, and are therefore mistakenly identified as different people. Further, Koca et al. (2023) trained subjects to become familiar with faces that share a varying degree of similarity with non-target faces to create clusters of identities in the face space in varying densities. They found that faces that shared high similarity with the non-target faces (i.e., higher exemplar density, hence smaller attractor field) were mistakenly identified as "different" when two novel photos of the targets were shown in a same/different matching task. In line with the attractor field model, creating a higher exemplar density surrounding a target face created smaller attractor fields and made it more difficult to tolerate within-person variability.

If the size of an attractor field is associated with the exemplar density of the region where a face is encoded (i.e., distinctiveness), and the range within which within-person variability can be tolerated (i.e., familiarity), then familiar faces should become more distinctive as they become familiar. In other words, the representations of familiar faces would migrate to a region with lower exemplar density that allows them to be recognized across viewpoints, and be surrounded by a larger attractor field. Support for this came from Faerber et al. (2016) who asked subjects to make distinctiveness judgments for familiar and unfamiliar faces, and importantly, their anti-faces. Anti-faces and their corresponding faces are equidistant from the norm, and therefore are located in a region with similar exemplar density. Nevertheless, Faerber et al. (2016) found that familiar faces were rated as more distinctive than their anti-faces, whereas unfamiliar faces were rated as similarly distinctive as their anti-faces. This suggests that familiarity influences how faces are distributed in the face space, where familiar faces likely expand their attractor field as they become familiar, so that more within-person variability can be tolerated.

In another study, Chauhan et al. (2020) provided evidence that familiarity changes the distribution of representations in face space. In their study, they created morph faces by combining either an unfamiliar face with another unfamiliar face (stranger-stranger), an unfamiliar face with a personally familiar face (stranger-friend), an unfamiliar face with the subjects' own faces (stranger-self), a personally familiar face with another personally familiar face (friend-friend), and a personally familiar face with the subjects' own faces (friend-self). These morphs were created along a continuum with 10% increments. Subjects were presented with the morphs along this continuum, followed by the two faces used to create the morph, and were asked which of the two identities was more similar to the morph image. They found that for 50-50 morphs, subjects were more likely to indicate that the morph was more similar to the more unfamiliar face. For example, for stranger-friend morphs, subjects chose the stranger face, and for friend-self morphs, they chose the friend face. The authors argued that viewers become more sensitive to the structural features of familiar faces, hence narrowing the identity-category boundary. In turn, when a 50-50 morph is presented between a familiar and an unfamiliar face, it does not share enough resemblance to the familiar face to be accepted as the familiar face, because the criterion to determine whether an input is similar enough to a familiar face is more conservative for familiar faces. Webster et al. (2004) obtained similar results, finding that observers had narrower category boundaries for own-race and own-gender faces (Webster et al., 2004), indicating greater sensitivity to faces of one's own category. Chauhan et al. explained these results in terms of an expansion of identity-specific subspaces in the face space as a function of familiarity.

This finding seems to contradict the predictions of the attractor field model. Since tolerance to within-person variability is almost perfect for familiar faces (e.g., Jenkins et al., 2011), and familiar faces appear more distinctive than they otherwise would be if they were unfamiliar (Faerber et al., 2016), it is reasonable to assume that familiar faces have a large attractor field. Since a 50-50 morph composed of a face with a larger attractor field (e.g., a distinctive face) and a smaller attractor field (e.g., a typical face) bears more resemblance to the face with the larger attractor field (Tanaka et al., 1998), then a 50-50 morph of a familiar and an unfamiliar face should look more like the familiar face. However, Chauhan et al. (2020) found that a 50-50 morph was judged to be more similar to the unfamiliar face.

One possibility is that Chauhan et al.'s (2020) findings could be solely explained by a distinctiveness effect. Even though a face is *relatively* more distinctive in the face space of the familiar observer than in the face space of the unfamiliar observer (Faerber et al., 2016), this doesn't necessitate that the face is distinctive relative to the average unfamiliar face for observers as a whole, and indeed it may be the case that some unfamiliar faces are more distinctive to an observer than the most distinctive face with which they are familiar. Therefore, the 50-50 morph between a distinctive unfamiliar face and a more typical familiar face may bear more resemblance to the unfamiliar face due to its larger attractor field. However, if their findings were only due to the initial distinctiveness of the face images, it would mean that the distinctiveness of the face images was coincidentally correlated with their level of familiarity. The authors found that between two categories of familiar faces, faces of one's friends and one's self, the 50-50 morph was rated as more similar to the friend's face (i.e., the less familiar face). Therefore, it seems unlikely that their findings can be accounted for exclusively by distinctiveness, but since the initial distinctiveness of the faces is unknown it is possible that differences in the distinctiveness of the parent faces contributes to judgments of 50-50 morphs as well.

Another possibility is that subjects were relying on different reference points when making their judgments. Although the 50-50 morph of two faces is assumed to be equally similar to its constituent *photographs*, it does not have to be equally similar to the *identities* that are used to make the morph face. In Chauhan et al.'s (2020) study, observers were asked to judge which of the two identities was more similar to the morph image. For an unfamiliar face, the only reference point available to the observer is the photograph used in the study. For a familiar face, however, observers have an abstract, rich representation of the person's face beyond the photograph used to make the morph. In turn, when judging which *identity* looks more similar to the morph, observers might be more likely to select the unfamiliar face, because the morph looks more similar to their one and only reference point for the unfamiliar face. If observers are using a different representation as a reference point for the familiar face than the test image, that representation may be located further away from the 50-50 morph than the unfamiliar face, especially if the test image is perceived to be poor likeness.

This would also explain why observers were more likely to select "friend" in friend-self morphs. Previous research has shown that observers are poor at judging the likeness of their own faces, and often choose more favorable versions of themselves as better likeness (e.g., Allen et al., 2009; White et al., 2016; 2017). Therefore, a photograph of the observers taken under controlled conditions as was done in Chauhan et al.'s paper (2020) may not be perceived as a good likeness by the observers themselves, making them more conservative in judging the morph as resembling their face. Perhaps different findings would be obtained if subjects were asked which of the two *photographs* resemble the morph.

It is also possible that the attractor field model does not sufficiently account for familiarity. Although a larger attractor field allows for more inputs to be recognized as that face, as is the case for familiar faces, a larger attractor field also worsens discrimination performance. Tanaka and Corneille (2007) presented subjects with a typical or a distinctive face, followed by either the identical target face image, or a morph face composed of the typical and the distinctive face with varying ratios. Subjects were asked whether the faces were the same or different. They found that subjects were less likely to report "different" when a distinctive face was paired with a morph that included a higher distinctive face ratio compared to a typical face paired with a morph that included a higher typical face ratio. In other words, subjects were more likely to confuse a morph that was *x* units away from a distinctive face with the distinctive face than a morph that was *x* units away from a typical face with the typical face. This is because a face that is *x* units away from a distinctive face is more likely to fall within the boundary of its attractor field, whereas a face that is *x* units away from a typical face is less likely to fall within the boundary of its attractor field (Figure 1.13). Therefore, observers were *less* sensitive to slight changes in the appearance of a face with a large attractor field than a face with a small attractor field. However, Chauhan et al. (2020) found that observers were *more* sensitive to slight changes in the appearance of familiar faces which is contrary to the finding from Tanaka and Corneille (2007).

The assumption that familiar faces have a large attractor field does not hold when it comes to discrimination performance. Indubitably, familiar faces are successfully discriminated from others. In fact, it has been consistently shown that recognition performance for familiar faces is almost always perfect (e.g., Burton et al., 1999; Jenkins et al., 2011). Further, Koca et al. (2023) found that when the exemplar density of faces surrounding a target face was systematically manipulated, in turn manipulating the size of its attractor field, discrimination performance was unaffected. Note that in this study, observers were exposed to within-person variability in all conditions, suggesting that the face space, including the attractor field model, should be evaluated in the context of familiarity.

Therefore, it is unclear how the face space changes with familiarity. According to the findings of Faerber et al. (2016), a familiar face is located in a region with less exemplar density than an unfamiliar anti-face, suggesting that it is surrounded by a larger attractor field that allows space for multiple inputs to activate the face's representation (Tanaka et al. 1998; Tanaka & Corneille, 2007). Chauhan et al.'s finding is contrary to the predictions of a larger attractor field surrounding familiar faces – a 50-50 morph of a familiar and an unfamiliar face bears stronger resemblance to the unfamiliar face. Their finding, however, can be explained by the initial distinctiveness of the face images that are used. Further, subjects were not asked to make a judgment regarding the constituent *images* of the 50-50 morph, but its constituent *identities*, which likely contaminated the assumption that the 50-50 morph is equidistant from the familiar and the unfamiliar face. Importantly, in both Faerber et al. (2016) and Chauhan et al.'s (2020) studies, only one photo per identity was used, which does not capture the range of exemplars that correspond to the representation of a familiar face. They also used faces that are initially familiar or unfamiliar, so it is not known how the face space changes over time as a face becomes familiar.

The study proposed in this Chapter aims to address this gap by investigating how the face space changes as a face becomes familiar. Observers will be trained to become familiar with a newly encountered face while their perception of the appearance of a 50-50 morph between the target faces and unfamiliar faces will be measured over time. I expect that observers will perceive the 50-50 morph to be skewed towards one of the identities as a function of increasing familiarity. I believe this study will provide insights into how the face space changes over time as a face representation is refined with familiarity.

3.2 Proposed Methods

3.2.1 Participants

Based on a power analysis conducted using G*Power (Faul et al., 2009), to obtain an effect size of η_p^2 = .02 (obtained from Koca & Oriet, 2023) with 80% power for a within-subjects ANOVA with 4 levels, 69 subjects will be required. Undergraduate students will be recruited through the University of Regina Psychology Participant Pool. All subjects will have normal or corrected-to-normal vision. All procedures will be carried out in accordance with the Canadian Tri-Council Policy Statement on the ethical treatment of research participants and will be approved by the University of Regina Research Ethics Board. All participants will sign a consent form prior to their participation, and subjects will receive 1% bonus credit toward a psychology course as compensation. Subjects who are familiar with the faces used in the study will be excluded from the analysis.

3.2.2 Stimuli and Apparatus

16 photos each for six targets, one photo each for 24 unique foil identities, and one photo each for 108 unique distractor identities will be collected from the Internet. Photos belonging to Turkish celebrities will be used to ensure that the targets are unfamiliar in Canada, and that many high-quality ambient images are available for each face. None of the faces will be obstructed (e.g., hands, glasses). Each photo will be cropped so that only the face is shown, and photos will be landmarked using webmorphR (DeBruine, 2022). Four images from each target will be randomly chosen to be used during testing. For all these test images and their corresponding foils, a morph continuum will be created with 10% intervals: 0-100, 10-90, 20-80, 30-70, 40-60, 50-50, 60-40, 30-70, 20-80, 10-90, 100-0 target-foil ratio. The brightness of all the face images will be matched to ensure that the 50-50 morph images do not appear skewed towards one of the identities due to image-specific characteristics. The experiment will be programmed using jsPsych (de Leeuw, 2015). The data will be collected online on pavlovia.org.

3.2.3 Procedure

First, subjects will be shown two unfamiliar face images side by side for five seconds and will be asked to study these faces. After the faces disappear, they will be asked to adjust a morph face so that it bears equal resemblance to the two face images shown. Subjects will be able to adjust back and forth between the faces along the morph continuum using arrow keys such that one full cycle will bring participants back to their starting point. The morph that is presented initially will be randomly selected from the morph continuum. The ratio of the target in the morph that was chosen by the subjects in this initial phase will be recorded as baseline. After subjects report their response, they will be shown four ambient images belonging to one of the identities initially shown to the participants to familiarize them with this identity (i.e., the target face). This face will be interspersed amongst six unique distractor identities to emulate how we learn faces in the real world (Koca et al., 2023; Koca & Oriet, 2023). Subjects will be asked to rate the distinctiveness of these ten faces to encourage them to pay attention to the images, which will not be analyzed for the purposes of the current study.

Following the training phase, subjects will be shown a novel photo of this target identity paired with a novel unfamiliar foil identity for five seconds, and will be asked to adjust a morph face again so that it bears equal resemblance to the two face images shown, exactly as they did the first phase (Figure 3.1). This procedure will repeat a total of three times each for six targets, yielding three training phases and four test phases per target. The trials will be blocked by target identity. The order in which the target identities are assigned to the blocks will be randomized. The order in which the target and the distractor images are presented during the training phase will also be randomized with the constraint that no two consecutive trials will depict the target identity. The order of the target-foil pairs presented during testing will also be randomized. In each test trial, target images will be paired with a novel unfamiliar identity to eliminate any possible identity- or image-specific effects. This will also be done to prevent subjects from incidentally becoming familiar with the foil faces during testing. In other words, this will be done to ensure that the change in the subjects' responses is only influenced by their increasing familiarity with the target. These foil faces will match the target faces demographically (e.g., similar age and gender). Note that, unlike Chauhan et al. (2020), subjects will be asked to adjust the morph so it is equally similar to the two face *images*, and not *identities*. This will be done to ensure that a 50-50 morph would truly be equidistant from the two test stimuli in the face space and the task would not be contaminated by other representations of the target that the subjects have stored from their exposure to different exemplars during the experiment. At the end of the experiment, subjects will be asked whether they were familiar with any of the faces used in the study prior to the experiment. Subjects who reported familiarity with the targets will be removed from the analysis.

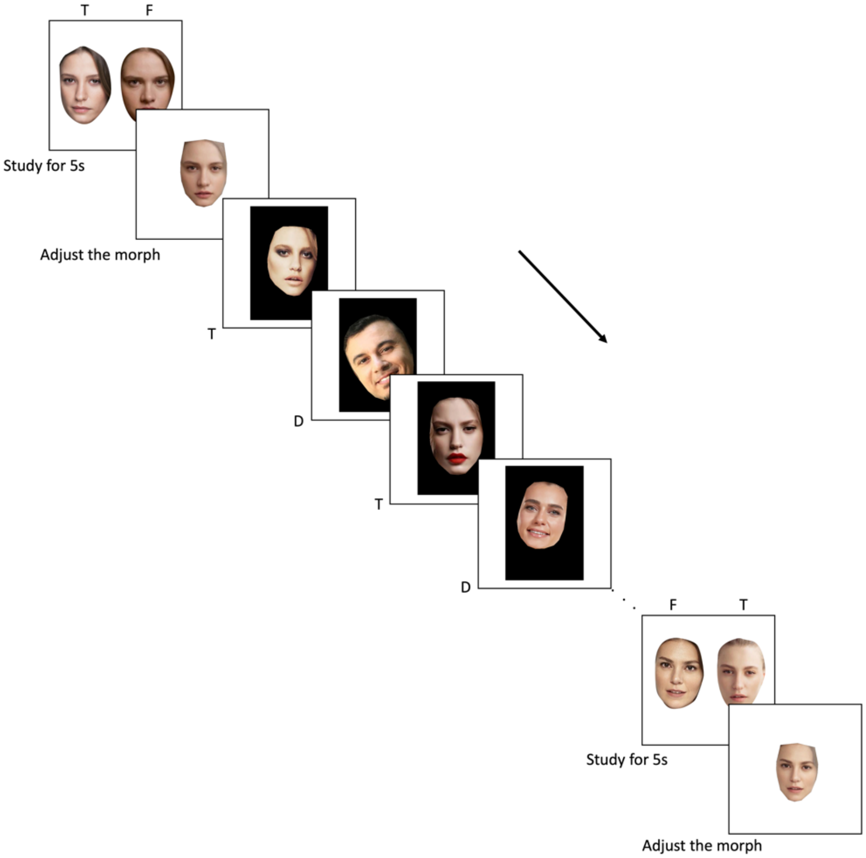

Figure 3.1. An illustration of the procedure for the study. Subjects will first study the faces for 5 seconds, followed by adjusting a morph face of the two faces. After a training session, this procedure will be repeated for a total of 4 times. T = Target, F = Foil, D = Distractor. Prompts and sliders are not included in the illustration.

3.3 Proposed Analyses and Hypotheses

The study will have one within-subjects variable with 4 levels: Time Point (T1-T4). The dependent variable will be the ratio of target in the morph that was chosen to be equally similar to the two faces. A linear model will be fit to the data with Time Point as the predictor of the target ratio using R (R Core Team, 2021). I expect that, in T1, subjects will be accurate at selecting the 50-50 morph as equally resembling the two images since both faces are unfamiliar.

In other words, the target ratio of the selected morph at T1 is expected to be around 50%. However, I expect this ratio to change over time as the targets become more familiar.

One possibility is that the target ratio in the morph will increase over time, so subjects will perceive a morph containing a higher ratio of the target face to bear equal resemblance to the two faces. I will call this *the narrowed category boundary hypothesis.* According to Chauhan et al.'s (2020) finding, where a 50-50 morph of a familiar and an unfamiliar face looks more similar to the unfamiliar face, I predict that subjects would perceive the actual 50-50 morph to look more like the unfamiliar face over time. In turn, they would adjust away from the unfamiliar face and choose a morph that has a higher target ratio. If observers are becoming more conservative in perceiving a familiar face as the familiar face so that even slight changes in the appearance of a familiar face would cross the identity boundary, observers would need to add "more" target to the morph mix to be able to perceive the morph to be equally similar to the two faces.

Another possibility is that the target ratio in the morph will decrease over time, so that a morph containing a higher ratio of the unfamiliar face will be perceived as the 50-50 morph. I will call this the *expanding attractor field hypothesis.* If we assume familiar faces are more distinctive (Faerber et al., 2016), and distinctive faces have larger attractor fields (Tanaka et al., 1998), a face's attractor field should become larger as it becomes familiar. Therefore, a 50-50 morph should look more similar to the familiar face over time, causing observers to adjust away from the familiar face to achieve the point of subjective equality.

Both these hypotheses have the assumption that a familiar face moves away from other faces in the face space, expanding the region surrounding them. They differ in terms of whether this expanding region will *push away* a similar-looking exemplar and reject it as the target (i.e., the narrowed category boundary hypothesis), or *pull in* a similar-looking exemplar to improve the likelihood that it would be accepted as the target (i.e., the expanding attractor field hypothesis). These hypotheses, however, do not necessarily answer the question regarding how that region in the face space surrounding a face's representation accommodates or encodes different exemplars of that face. Chauhan et al. (2020) referred to their finding as an expansion of an identity-specific subspace, which is an intriguing approach to how familiarity would be represented in the face space. In Chapter 4, properties of such a subspace will be explored.

4 Exploring Identity-Specific Subspaces in the Face Space

4.1 Introduction

Familiarity yields remarkable benefits to face processing. While observers have great difficulty recognizing novel instances of unfamiliar faces, and perceptually matching two photos of an unfamiliar face, such tasks become trivial for familiar faces, even under poor viewing conditions (e.g., Bruce et al., 1999; 2001; Hancock et al., 2000; Megreya & Burton, 2006; 2007). This led some researchers to argue that familiar and unfamiliar faces are processed in qualitatively different ways such that unfamiliar face processing recruits unsophisticated strategies that rely on image-based cues, while familiar face processing matching relies on abstract representations in memory.

For example, Megreya and Burton (2006) showed that inverted familiar face processing was strongly associated with upright unfamiliar face processing, suggesting that configural information is not as effectively extracted from unfamiliar faces. Further, in card sorting tasks, where subjects are asked to sort multiple ambient unfamiliar face photos in different piles each corresponding to an identity, they make many more piles than the number of identities in the set while rarely placing two different identities in the same pile (e.g., Jenkins et al., 2011; Laurence et al., 2016; but see Andrews et al., 2015). In other words, while unfamiliar observers have great difficulty judging two separate exemplars of the same person as depicting the same person, it is easier to accurately judge that two photos belonging to different people are different people. It seems then that a very interesting problem with face recognition is not only the ability to tell faces apart (i.e., discriminate), but also tell faces "together" (i.e., amalgamate). This is because faces vary in tremendous ways – one person can look drastically different across different viewing conditions (e.g., angle, lighting, age, weight; Figure 1.15). When a face is familiar, this variability is readily tolerated and the face can be recognized. When a face is unfamiliar, the visual system cannot tolerate the variability across different instances of the face, hence it mistakenly categorizes these separate instances as different identities.

Importantly, this variability is unique to each face because being familiar with a face does not transfer that familiarity to a different face. Burton et al. (2016) provided computational evidence that faces vary in idiosyncratic ways – when multiple ambient photos of different celebrities were used in training a PCA model, the model explained the variability in photos of different people in different ways. For example, the same component captured facial expression for one identity but gaze direction in another identity. This means that for a face to be familiar, observers need to be exposed to a sufficient range of variability in that face so that a novel instance is more likely to be successfully recognized. Accordingly, many studies have shown that exposure to high, unsystematic within-person variability captured in sets of ambient images yields superior recognition performance (e.g., Andrews et al., 2015; Burton et al., 2016; Corpuz & Oriet, 2022; Ritchie & Burton, 2017).

Robust and seemingly qualitative differences between processing familiar and unfamiliar faces give rise to the idea that familiar faces are represented as identity-specific subspaces in the face space. This idea is particularly compelling considering that faces vary in idiosyncratic ways.

Therefore, it is unlikely that each face loads differently onto a shared set of dimensions as is often conceptualized in the face space, because the very dimensionality with which each face is encoded is idiosyncratic (Burton et al., 2016). For example, suppose there is a dimension of the face space that encodes age. How would this dimension accommodate different exemplars of a familiar face that vary in age? If we assume that each face is encoded using a common set of dimensions, we need to assume that the age dimension of this space hosts multiple representations of a familiar face in varying age groups. However, recent evidence suggests that this is not the case. Laurence et al. (2022) investigated whether a single representation of a familiar face can incorporate different ages, or whether multiple representations are needed to represent a familiar face across different ages (i.e., younger adult, older adult) by presenting older adults with the faces of older famous adults who had aged alongside them (e.g., Paul McCartney).

In their first experiment, they used a long-lag repetition priming task where subjects were primed with images of famous faces followed by a test phase. The test image could be the same image shown during the priming phase, a different image of the same person from the same age, or an image of the person depicting a different age than the priming phase. During testing, subjects' task was to indicate whether the test image was a famous or a non-famous person, and their reaction times were recorded. They found that the effect of priming was similar across the three test conditions, indicating that the priming effect can transcend different ages of the same face. In their second experiment, an adaptation paradigm was used. Subjects viewed an adaptor image that depicted either an older or a younger photo of a target face, followed by a test image that included the target face morphed with a non-target identity by 50%. The age of the test image was either congruent or incongruent with the age of the adaptor image (e.g., older-older or younger-older, respectively). Subjects were asked to indicate whether the test image better resembled the target or the non-target identity. They found that the magnitude of the aftereffect was similar across the congruent and incongruent conditions, suggesting that identity aftereffects are not contingent on the person's age. Together, these findings suggest that a single familiar face representation can accommodate changes in age.

This is especially fascinating considering that aftereffects are found to be category-contingent, meaning that they do not transfer over different age, race, or species categories (Little et al., 2008). When a face is familiar, however, changes in such categories[8] do not impact the magnitude of adaptation, suggesting that familiar identity representations can encompass a wide range of variability across different viewing conditions. Support for this conclusion came from Hole (2011), who found that identity-specific aftereffects transfer over a range of viewing conditions for familiar faces like inversion, changes in viewpoint, and even vertically stretching the face three times the normal height. Transference of aftereffects within different images of a familiar identity despite drastic changes in low-level image properties (Hole, 2011), and despite distinct changes in categories like age (Laurence et al., 2022) suggest that familiar faces may be represented as subspaces in the face space.

Further evidence for identity-specific subspaces come from computer vision, where PCA models that utilize identity-specific subspace methods outperform the traditional Eigenface methods (Aishwarya & Marcus, 2010; Shan et al., 2003). Put simply, in the traditional Eigenface methods, the model is trained with multiple standardized images each belonging to a different individual whereas in the subspace method multiple models are created per individual, each trained with multiple images of that individual. The model's recognition performance is often

[8] Certainly, a familiar identity is unlikely to switch between particular categories like race, unless perhaps in rare cases like Michael Jackson.

measured by projecting a novel image in the face space and calculating its distance from previously trained exemplars, which is referred to as reconstruction error. Evidence suggests that reconstruction error of a novel image of a previously trained identity is reduced for identity-specific subspace models (Aishwarya & Marcus, 2010; Shan et al., 2003).

The identity-specific subspace account can also explain how refining a familiar face's representation can facilitate *both* amalgamation and discrimination. In the attractor field model (Tanaka et al., 1998), a face with a larger attractor field improves the likelihood that more exemplars can be recognized as that face. There is evidence in the literature suggesting that the size of an attractor field is associated with tolerance to within-person variability (Laurence et al., 2016), and there is also evidence suggesting that the region surrounding a familiar face's representation in the face space is larger (as indexed by distinctiveness) than an unfamiliar anti-face located in a region with comparable exemplar density (Faerber et al., 2016). This indicates that familiar faces should have a larger attractor field, allowing for multiple exemplars to be accepted as that face (i.e., amalgamation).

However, a larger attractor field also increases the likelihood of false alarms. Tanaka and Corneille (2007) found that discriminating between a distinctive target (i.e., a face with a large attractor field) and a similar-looking distractor was more difficult than discriminating between a typical target (i.e., a face with a small attractor field) and an equally similar-looking distractor. However, they only used unfamiliar faces in their study. A more recent study by Chauhan et al. (2020) found that observers were less likely to judge a morph composed of a familiar and unfamiliar face to resemble the familiar face, suggesting that when a face is familiar, observers are *more* conservative in accepting a similar-looking exemplar as the familiar face. Further, Koca et al. (2023) found that manipulating the size of the attractor field surrounding a face's representation in a familiarization paradigm did not influence discrimination performance in a matching task.

While assuming that the size of the attractor field surrounding a face’s representation increases with familiarity can explain the process of amalgamation, it does not necessarily explain discrimination performance for familiar faces. Perhaps a limitation of most of the face space models (e.g., the attractor field model, Tanaka et al., 1998; the Voronoi model, Lewis & Johnston; 1999) is the assumption that *all* faces occupy the same layer of the face space to the same extent. Because familiar and unfamiliar faces are represented very differently, it is difficult to make unifying predictions with a model of face recognition that possesses the same set of functions for processing both familiar and unfamiliar faces. However, this can be reconciled if it is assumed that the same faces are represented in different *layers* of face space. One layer of the face space would represent all faces and would be responsible for discriminating faces from each other. Another layer of the face space represents familiar faces as distinct categories with their own set of idiosyncratic dimensions, and would be responsible for amalgamation. Such a mechanism was proposed by Benson and Perrett (1991).

Qualitative differences between processing familiar and unfamiliar faces, the phenomenon that within-person variability is idiosyncratic, evidence from computational models trained for face recognition, and current limitations within the models of the face space all raise the possibility that there are identity-specific face spaces for familiar faces that can successfully allow observers to recognize novel instances of that face, while achieving perfect performance in discriminating the face from similar others. The goal of the study proposed in this chapter is to take first steps into exploring the properties of an identity-specific subspace.

An intriguing question is whether norm-based coding is a property of identity-specific subspaces as hypothesized for the population face space. Previous research showed that averaging faces can be an underlying mechanism of face learning (Burton et al., 2005; Koca & Oriet, 2023; Kramer et al., 2015). This suggests that an identity-specific norm is represented for familiar faces, which is a compelling account since different exemplars of a face vary greatly and the process of averaging would wash out this within-person variability allowing identity-invariant aspects of a face to be represented in memory. Recently, however, Davis et al. (2024) argued that ensemble coding (i.e., set averaging) may not be a route to face learning. In their study, subjects were presented with an array of four images depicting the same person. The array contained either four different ambient images (the different-images condition), the average of these four presented four times (the average-only condition), or a single image presented four times (the same-image condition). Recognition was tested using a novel image of the person. The authors found that presenting subjects with the average set yielded no recognition advantage. In fact, recognition performance was similar in the same-image and the average-only condition, both worse than the different-images condition. Since presenting subjects with an average did not generalize over recognition of a novel image, the authors reasoned that extracting an average representation is not sufficient and that exposure to within-person variability is imperative to learning a new face.

However, presenting observers with an average image may not be an appropriate operationalization of ensemble coding. To an unfamiliar eye, an average image is just another exemplar of that face, which would explain why Davis et al. (2024) found comparable recognition performance between the average-only and the same-image condition. If ensemble coding is assumed to be an underlying mechanism of face learning, it is important to consider not only the average representation that is created as an end product, but also the *process* of extracting this average representation. Given that ensemble coding serves to reduce redundancies and noise in the perceptual domain (perhaps even the social domain, see Whitney & Yamanashi-Leib, 2018), arguably, ensemble coding is only meaningful when the ensemble has some sort of underlying variability. If there is no variability in the set, then any one item is as good a representation as the average. Therefore, I argue that exposure to variability, in the case of faces, within-person variability, is an imperative aspect of the ensemble coding process and should be considered in investigations of ensemble coding as an underlying mechanism of face learning.

Nevertheless, as Davis et al. (2024) argued, it is unlikely that *only* an average is represented for familiar faces. The idiosyncratic nature of within-person variability suggests that not only the central tendency, but also the distribution of exemplars is critical for face processing (Burton et al., 2016). Importantly, many identity-invariant aspects of a face that one can conceptualize, like metric distances between features, are subject to within-person variability. If representing varying exemplars is important for recognizing a face, then what is the purpose of representing the average in addition to exemplars? Since an average representation is refined with increasing familiarity (Koca & Oriet, 2023), I propose that that this identity-specific norm serves as a reference point to allow within-person variability to be encoded.

In this study, I aim to investigate whether different exemplars of a face are encoded relative to a norm by using an adaptation paradigm similar to Leopold et al.'s study (2001). In their experiment, Leopold et al. created anti-faces by exaggerating an average face in the opposite direction of an identity's vector (Figure 1.11). They found that adapting to an anti-face, such as "anti-Adam", made the average face appear more like Adam, decreased the identification threshold for recognizing Adam, and increased the identification threshold for recognizing a non-

Adam face. This suggests that adaptation to an anti-face shifts perception in a *systematic* way - that is, along the identity trajectory of the anti-face - providing evidence for norm-based coding of the face space. If norm-based coding is also a property of the hypothesized identity-specific subspaces, then adapting to an anti-exemplar of a familiar face should produce similar results for different exemplars of a familiar face.

4.2 Proposed Methods

4.2.1 Participants

Ten undergraduate students will be recruited through the University of Regina Psychology Participant Pool. All subjects will have normal or corrected-to-normal vision. All procedures will be carried out in accordance with the Canadian Tri-Council Policy Statement on the ethical treatment of research participants and will be approved by the University of Regina Research Ethics Board. All participants will sign a consent form prior to their participation, and subjects will receive 1% bonus credit toward a psychology course as compensation. Subjects who are unfamiliar with the faces used in the study will be excluded from the analysis.

4.2.2 Stimuli and Apparatus

The stimuli construction will follow Leopold et al.'s (2001) approach, but instead of using images of different people, different types of stimuli will be created by using images of a single identity. To do so, eight images each for four famous celebrities will be collected from the internet. Using webmorphR (DeBruine, 2022), all images will be automatically delineated (i.e., points that correspond to facial features will be marked) and will be manually corrected when necessary. After delineation, four images of each target will be randomly selected to create an average face, and the remaining four images will be used as test stimuli. The average face for each target will be transformed away from each test stimulus by 80% (-0.8 strength). This is done by creating a vector between the test stimuli and the average, and then transforming the average image along that vector in the opposite direction of the test stimuli. This will create four anti-exemplars to be used as adapting stimuli. Lastly, the average face and each exemplar will be morphed in different ratios to create a set of "weaker" versions of each exemplar that gradually look more similar to the average face to measure identification thresholds.

The experiment will be programmed using PsychoPy3 (Peirce, 2007) and data will be collected in-person. Data will be collected using monitors with a 1920x1080 resolution. Subjects will be asked to sit 54 cm away from the screen so that 1 cm on the screen will correspond to 1 degree of visual angle.

4.2.3 Procedure

The procedure will closely follow Leopold et al.'s (2001) study, with the difference that instead of creating a norm-based face space using images of different identities, a norm-based face subspace will be created using different images of the same person. One subspace will be created for each target. The entire experiment will be blocked by target identity with a rest period in between each block, and the order of blocks will be randomized for each participant. The sections below will explain what happens in a given block where only one target identity is seen.

4.2.3.1 Practice. Following Leopold et al. (2001)'s procedure, subjects will first complete a practice phase to familiarize them with the test images. During the training phase, subjects will be shown four different images (i.e., exemplars) of a target face, and they will be asked to press labeled buttons to indicate which particular image they saw. For example, if the exemplar is an image of Keanu Reeves in the Matrix, they will be asked to press "a", if it is an image of Keanu Reeves in John Wick, they will be asked to press "s". This procedure will be carried out for 80 randomized trials, and subjects will receive feedback. Subjects who score below 80% in the training phase will not be eligible to participate in the experiment.

4.2.3.2 Baseline Phase. After the practice phase, baseline identification thresholds of the exemplar images will be measured using the method of constant stimuli. To do so, subjects will be presented with the exemplars and their weaker versions (created by morphing the exemplars with the average of the target in varying ratios) and will be asked to identify the exemplar as they did in the training phase. To measure the baseline for chance performance, the average target faces will also be included in this phase. Subjects will be informed that they will see some "weaker" versions, that is, a face that may not look like any of the exemplars, to which they should randomly respond. This will be done over 150 randomized trials, and subjects will not receive feedback for this phase.

4.2.3.3 Adaptation Phase. In the adaptation phase, subjects will view anti-exemplars presented on the screen for 5 seconds. Subjects will be able to move their gaze on the screen since it has been shown that after-effects for faces can transcend low-level properties like retinal size (Rhodes & Jeffery, 2006; Valentine et al., 2016). After viewing the adapting stimulus, subjects will be shown a test stimulus. The test stimulus can be the average face of the person, the exemplar image along the same vector of the anti-exemplar (matching-exemplar), or an exemplar image of the same person with a different vector direction (mismatching exemplar). As was done in the training and the baseline identification phases, weaker versions of these exemplars will also be used as test stimuli to measure identification thresholds. Subjects will be asked to identify these images using the same set of keys they used during the previous two phases. This will be done over about 1000 randomized trials.

4.3 Proposed Analyses and Hypotheses

Responses will be recorded as correct for trials where an exemplar was correctly identified, and when the average test image (0.0 strength) is identified as the corresponding exemplar of the adapted anti-exemplar. Accuracy will be calculated for each subject within each exemplar strength level (manipulated by morphing the exemplars with the target's average at varying levels). Following Rhodes and Jeffery (2006) and Leopold et al. (2001), using R a four-parameter logistic function will be fitted to the data separately for the baseline phase, the matching-exemplar trials (adaptation phase), and the mismatching-exemplar trials (adaptation phase). Identification threshold will be determined as the identity strength that is correctly identified 50% of the time and will be separately calculated for the baseline phase, the matching-exemplar trials (adaptation phase), and the mismatching-exemplar trials (adaptation phase).

I expect that adapting to an anti-exemplar will shift the appearance of test stimuli in systematic ways. If identity-specific subspaces are architectured in a norm-based fashion, adapting to an anti-exemplar should selectively increase sensitivity to the matching-exemplar image since they are along the same vector. This would be observed as decreased identification thresholds for the matching-exemplar images compared to baseline. Similarly, when the target's average image is shown post-adaptation, the appearance of the average image would be biased towards the matching-exemplar, suggesting that adapting to an anti-exemplar would increase sensitivity to the matching-exemplar image.

An interesting question is regarding what would happen to the identification threshold for the mismatching-exemplar images. One possibility is that the identification threshold will increase for mismatching-exemplars, suggesting that increased sensitivity to the matching-exemplar will make it more difficult to identify mismatching-exemplars. This is similar to what was found in Leopold et al. (2001)'s and Rhodes and Jeffery (2006)'s studies using mismatching-*identities*. The authors found that while adapting to anti-Adam made it easier to identify Adam (matching-identity), it made it more difficult to identify Jim (mismatching-identity). This is because adapting to anti-Adam increased sensitivity to Adam's facial features which are not likely present in Jim.

A mismatching-*exemplar*, however, is still the same identity (i.e., it is a mismatching exemplar for Adam) and hence has the same identity-specific characteristics as the matching-exemplar, so it is likely that the features to which the observers have increased sensitivity after adaptation to an anti-exemplar will be present in the mismatching-exemplar as well. Previous research showed that adaptation effects transfer within different views of the same identity (Carbon & Ditye, 2012; Hole, 2011; Laurence et al., 2022). Therefore, adaptation to an anti-exemplar can decrease the identification threshold for the mismatching-exemplars. Nevertheless, I predict that even in that case, the magnitude of identification threshold decrease will be larger for the matching-exemplar than for the mismatching exemplar.

It is difficult to predict what would be observed if the architecture of the identity-specific subspace is exemplar-based. Since exemplar-based accounts of the face space assume that a face representation is activated based on its similarity (Euclidean distance) to the test face, the adaptation effects would be stronger for test faces that are similar to the anti-exemplar, irrespective of the vector trajectory. Therefore, identification thresholds would increase for exemplars that are perceptually similar to the anti-exemplar. This account will not be explicitly explored in this study and is a potential avenue for future research.

5 Anticipated Significance of the Proposed Research Program

Face processing is a difficult problem for the visual system to solve. While different individuals share common patterns (i.e., two eyes above a nose above a mouth), different views of one individual have great variability from one instance to the next (e.g., age, make-up, weight, lighting, angle, and so on). Understanding how faces are represented in memory can contribute to our understanding of how the human mind can organize and process information in a way that yields successful recognition of a familiar face despite the idiosyncratic nature of face variability while achieving successful differentiation of different individuals. The goal of this research program is to reconcile these two aspects of face processing, providing a comprehensive explanation of how the visual system can achieve the spectacular feat of recognizing thousands of faces.

Faces are an important and ubiquitous aspect of our daily lives. Our day-to-day social interactions rely heavily on our ability to successfully recognize a face. People may feel insulted when we fail to recognize their face, or it may create awkward social situations if we falsely recognize a stranger as someone familiar. Further, recognizing a face not only helps us retrieve the person's name, but also access semantic information about the person (e.g., occupation, recent life events) that helps guide social interactions (Bruce & Young, 1986).

Beyond social interactions, many important decisions in our society also rely on successfully recognizing faces. For example, our justice system can use eye-witness testimonies to decide whether to convict somebody of a crime. We use photo identification to decide whether someone truthfully represents themselves while entering a country, or to decide whether we should sell them alcohol. We develop computer algorithms to recognize faces for security reasons - most of us even have smartphones that unlock based on face recognition. We think that capturing a face on camera recording a crime means that someone, somewhere will recognize the perpetrator if only the image is circulated widely enough, ignoring whether the image captured is a good likeness of the individual. Understanding how faces become familiar can help develop more effective systems to improve eye-witness testimonies, photo identifications, and security.

References

Addams, R. (1834). LI. An account of a peculiar optical phenomenon seen after having looked at a moving body. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *5*(29), 373-374. https://doi.org/10.1080/14786443408648481

Aishwarya, P., & Marcus, K. (2010). Face recognition using multiple eigenface subspaces. *Journal of Engineering and Technology Research*, *2*(8), 139-143.

Allen, H., Brady, N., & Tredoux, C. (2009). Perception of 'best likeness' to highly familiar faces of self and friend. *Perception*, *38*(12), 1821-1830. https://doi.org/10.1068/p6424

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology*, *68*(10), 2041-2050. https://doi.org/10.1080/17470218.2014.1003949

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, *12*(3), 219-228. https://doi.org/10.3758/BF03197669

Benson, P. J., & Perrett, D. I. (1991). Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology*, *3*(1), 105-135. https://doi.org/10.1080/09541449108406222

Blakemore, C., & Sutton, P. (1969). Size adaptation: A new aftereffect. *Science*, *166*(3902), 245-247. https://doi.org/10.1126/science.166.3902.245

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2021). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition*, *208*, 104341. https://doi.org/10.1016/j.cognition.2020.104341

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin*, *15*(1), 19-25. https://doi.org/10.1177/0146167289151002

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305-327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruce, V., Burton, M. A., & Dench, N. (1994). What's distinctive about a distinctive face?. *The Quarterly Journal of Experimental Psychology*, *47*(1), 119-141. https://doi.org/10.1080/713755777

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339. https://doi.org/10.1037/1076-898X.5.4.339

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. https://doi.org/10.1037/1076-898X.7.3.207

Burton, A. M., Jenkins, R., Hancock, P. J., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*(3), 256-284. https://doi.org/10.1016/j.cogpsych.2005.06.003

Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223. https://doi.org/10.1111/cogs.12231

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243-248. https://doi.org/10.1111/1467-9280.00144

Burton, M. A. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485. http://dx.doi.org/10.1080/17470218.2013.800125

Byatt, G., & Rhodes, G. (1998). Recognition of own-race and other-race caricatures: Implications for models of face recognition. *Vision Research*, *38*(15-16), 2455-2468. https://doi.org/10.1016/S0042-6989(97)00469-0

Byatt, G., & Rhodes, G. (2004). Identification of own-race and other-race faces: Implications for the representation of race in face space. *Psychonomic Bulletin & Review*, *11*(4), 735-741. https://doi.org/10.3758/BF03196628

Carbon, C. C., & Ditye, T. (2012). Face adaptation effects show strong and long-lasting transfer from lab to more ecological contexts. *Frontiers in Psychology*, *3*, 3. https://doi.org/10.3389/fpsyg.2012.00003

Carey, S. (1992). Becoming a face expert. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 95-103. https://doi.org/10.1098/rstb.1992.0012

Carroll, L. (1946). Through the looking glass and what Alice found there. New York: Random House.

Chauhan, V., Kotlewska, I., Tang, S., & Gobbini, M. I. (2020). How familiarity warps representation in the face space. *Journal of Vision*, *20*(7), 18-18. https://doi.org/10.1167/jov.20.7.18

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *48*(4), 879-894. https://doi.org/10.1080/14640749508401421

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition*, *11*(7), 857-869. https://doi.org/10.1080/13506280444000021

Cohen, M. E., & Carr, W. J. (1975). Facial recognition and the von Restorff effect. *Bulletin of the Psychonomic Society*, *6*(4), 383-384. https://doi.org/10.3758/BF03333209

Corpuz, R. L., & Oriet, C. (2022). Within-person variability contributes to more durable learning of faces. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *76*(4), 270–282. https://doi.org/10.1037/cep0000282

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Cross, J. F., Cross, J., & Daly, J. (1971). Sex, race, age, and beauty as factors in recognition of faces. *Perception & Psychophysics*, *10*(6), 393-396. https://doi.org/10.3758/BF03210319

Davis, E. E., Matthews, C. M., & Mondloch, C. J. (2024). Ensemble coding of facial identity is robust, but may not contribute to face learning. *Cognition*, *243*, 105668. https://doi.org/10.1016/j.cognition.2023.105668

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1-12. https://doi.org/10.3758/s13428-014-0458-y

DeBruine, L. (2022). _webmorphR: Reproducible Stimuli_. https://debruine.github.io/webmorphR/, https://github.com/debruine/webmorphR.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115*(2), 107–117. https://doi.org/10.1037/0096-3445.115.2.107

Ellis, H. D., Deregowski, J. B., & Shepherd, J. W. (1975). Descriptions of white and black faces by white and black subjects. *International Journal of Psychology*, *10*(2), 119-123. https://doi.org/10.1080/00207597508247325

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. https://doi.org/10.1068/p080431

Faerber, S. J., Kaufmann, J. M., Leder, H., Martin, E. M., & Schweinberger, S. R. (2016). The role of familiarity for representations in norm-based face space. *PloS one*, *11*(5), e0155380. https://doi.org/10.1371/journal.pone.0155380

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect?. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 628. https://doi.org/10.1037/0096-1523.21.3.628

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is" special" about face perception?. *Psychological Review*, *105*(3), 482. https://doi.org/10.1037/0033-295X.105.3.482

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception*, *29*(2), 159-170. https://doi.org/10.1068/p3012

French, J. (2023). smerc: Statistical methods for regional counts. *R package version 1.8.2*, <https://CRAN.R-project.org/package=smerc>.

Ge, L., Zhang, H., Wang, Z., Quinn, P. C., Pascalis, O., Kelly, D., Slater, A., Tian, J., & Lee, K. (2009). Two faces of the other-race effect: Recognition and categorisation of Caucasian and Chinese faces. *Perception*, *38*(8), 1199-1210. https://doi.org/10.1068/p6136

Gibling, F., & Bennett, P. (1994). Artistic enhancement in the production of Photo-FIT likenesses: An examination of its effectiveness in leading to suspect identification. *Psychology, Crime and Law*, *1*(1), 93-100. https://doi.org/10.1080/10683169408411939

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of Experimental Psychology*, *20*(5), 453. https://doi.org/10.1037/h0059826

Going, M., & Read, J. D. (1974). Effects of uniqueness, sex of subject, and sex of photograph on facial recognition. *Perceptual and Motor Skills*, *39*(1), 109-110. https://doi.org/10.2466/pms.1974.39.1.109

Goldstein, A. G. (1979). Race-related variation of facial features: Anthropometric data I. *Bulletin of the Psychonomic Society*, *13*(3), 187-190. https://doi.org/10.3758/BF03335055

Goldstein, A. G., & Chance, J. E. (1980). Memory for faces and schema theory. *The Journal of Psychology*, *105*(1), 47-59. https://doi.org/10.3758/BF03329829

Goldstein, A. G., & Chance, J. E. (1985). Effects of training on Japanese face recognition: Reduction of the other-race effect. *Bulletin of the Psychonomic Society*, *23*(3), 211-214. https://doi.org/10.3758/BF03329829

Hagen, M. A., & Perkins, D. (1983). A refutation of the hypothesis of the superfidelity of caricatures relative to photographs. *Perception*, *12*(1), 55-61. https://doi.org/10.1068/p120055

Hancock, P. J., & Little, A. C. (2011). Adaptation may cause some of the face caricature effect. *Perception*, *40*(3), 317-322. https://doi.org/10.1068/p6865

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330-337. https://doi.org/10.1016/S1364-6613(00)01519-9

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception*, *31*(10), 1221-1240. https://doi.org/10.1068/p3252

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 93-103. https://doi.org/10.1002/wcs.1203

Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology*, *5*(1), 1-10. https://doi.org/10.20982/tqmp.05.1.p001

Jenkins, R., Dowsett, A. J., & Burton, A. M. (2018). How many faces do people know?. *Proceedings of the Royal Society B*, *285*(1888), 20181319. https://doi.org/10.1098/rspb.2018.1319

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323. https://doi.org/10.1016/j.cognition.2011.08.001

Johnston, R. A., Milne, A. B., Williams, C., & Hosie, J. (1997). Do distinctive faces come from outer space? An investigation of the status of a multidimensional face-space. *Visual Cognition*, *4*(1), 59-67. https://doi.org/10.1080/713756748

Kaufmann, J. M., & Schweinberger, S. R. (2008). Distortions in the brain? ERP effects of caricaturing familiar and unfamiliar faces. *Brain Research*, *1228*, 177-188. https://doi.org/10.1016/j.brainres.2008.06.092

Koca, Y., & Oriet, C. (2023). From pictures to the people in them: Averaging within-person variability leads to face familiarization. *Psychological Science*, *34*(2), 252-264. https://doi.org/10.1177/09567976221131520

Koca, Y., Corpuz, R. L., Oriet, C. (2023). Crowding the face-space: The attractor field hypothesis and within-person variability. [Manuscript submitted for publication]

Kramer, R. S., Manesi, Z., Towler, A., Reynolds, M. G., & Burton, A. M. (2018). Familiarity and within-person facial variability: The importance of the internal and external features. *Perception*, *47*(1), 3-15. https://doi.org/10.1177/0301006617725242

Kramer, R. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision*, *15*(4), 1-1. https://doi.org/10.1167/15.4.1

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling* (No. 11). Sage.

Laurence, S., Baker, K. A., Proietti, V. M., & Mondloch, C. J. (2022). What happens to our representation of identity as familiar faces age? Evidence from priming and identity aftereffects. *British Journal of Psychology*, *113*(3), 677-695. https://doi.org/10.1111/bjop.12560

Laurence, S., Zhou, X., & Mondloch, C. J. (2016). The flip side of the other-race coin: They all look different to me. *British Journal of Psychology*, *107*(2), 374-388. https://doi.org/10.1111/bjop.12147

Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science*, *11*(5), 379-385. https://doi.org/10.1111/1467-9280.00274

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), 89-94. https://doi.org/10.1038/82947

Leopold, D. A., Rhodes, G., Müller, K. M., & Jeffery, L. (2005). The dynamics of visual adaptation to faces. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1566), 897-904. https://doi.org/10.1098/rspb.2004.3022

Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1364–1382. https://doi.org/10.1037/0278-7393.22.6.1364

Levin, D. T. (2000). Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, *129*(4), 559. https://doi.org/10.1037/0096-3445.129.4.559

Lewis, M. B., & Ellis, H. D. (2000). Satiation in name and face recognition. *Memory & Cognition*, *28*, 783-788. https://doi.org/10.3758/BF03198413

Lewis, M. B., & Hills, P. J. (2018). Perceived race affects configural processing but not holistic processing in the composite-face task. *Frontiers in Psychology*, *9*, 396719. https://doi.org/10.3389/fpsyg.2018.01456

Lewis, M. B., & Johnston, R. A. (1998). Understanding caricatures of faces. *The Quarterly Journal of Experimental Psychology Section A*, *51*(2), 321-346. https://www.doi.org/10.1080/713755758

Lewis, M. B., & Johnston, R. A. (1999). A unified account of the effects of caricaturing faces. *Visual Cognition*, *6*(1), 1-42. https://doi.org/10.1080/713756800

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(3), 212. https://doi.org/10.1037/0278-7393.5.3.212

Mauro, R., & Kubovy, M. (1992). Caricature and face recognition. *Memory & Cognition*, *20*(4), 433-440. https://doi.org/10.3758/BF03210927

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*, 865-876. https://doi.org/10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*, 1175-1184. https://doi.org/10.3758/BF03193954

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology*, *64*(8), 1473-1483. https://doi.org/10.1080/17470218.2011.575228

Mileva, M., Kramer, R. S., & Burton, A. M. (2019). Social evaluation of faces across gender and familiarity. *Perception*, *48*(6), 471-486. https://doi.org/10.1177/03010066198489

Ng, W. J., & Lindsay, R. C. (1994). Cross-race facial recognition: Failure of the contact hypothesis. *Journal of Cross-Cultural Psychology*, *25*(2), 217-232. https://doi.org/10.1177/0022022194252004

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39. https://doi.org/10.1037/0096-3445.115.1.39

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, *162*(1-2), 8-13. https://doi.org/10.1016/j.jneumeth.2006.11.017

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rhodes, G. (1997). *Superportraits: Caricatures and recognition*. Psychology Press.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, *46*(18), 2977-2987. https://doi.org/10.1016/j.visres.2006.03.002

Rhodes, G., & Tremewan, T. (1994). Understanding face recognition: Caricauture effects, inversion, and the homogeneity problem. *Visual Cognition*, *1*(2-3), 275-311. https://doi.org/10.1080/13506289408402303

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive psychology*, *19*(4), 473-497. https://doi.org/10.1016/0010-0285(87)90016-8

Rhodes, G., Byatt, G., Tremewan, T., & Kennedy, A. (1996). Facial distinctiveness and the power of caricatures. *Perception*, *26*(2), 207-223. https://doi.org/10.1068/p260207

Rhodes, G., Carey, S., Byatt, G., & Proffitt, F. (1998). Coding spatial variations in faces and simple shapes: a test of two models. *Vision Research*, *38*(15-16), 2307-2321. https://doi.org/10.1016/S0042-6989(97)00470-7

Rhodes, G., Locke, V., Ewing, L., & Evangelista, E. (2009). Race coding and the other-race effect in face recognition. *Perception*, *38*(2), 232-241. https://doi.org/10.1068/p6110

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology*, *70*(5), 897–905. https://doi.org/10.1080/17470218.2015.1136656

Ritchie, K. L., Kramer, R. S., & Burton, A. M. (2018). What makes a face photo a ‘good likeness’?. *Cognition*, *170*, 1-8. https://doi.org/10.1016/j.cognition.2017.09.001

Ross, D. A., Deroche, M., & Palmeri, T. J. (2014). Not just the norm: Exemplar-based models also predict face aftereffects. *Psychonomic Bulletin & Review*, *21*, 47-70. https://doi.org/10.3758/s13423-013-0449-5

Rossion, B. (2018). Humans are visual experts at unfamiliar face recognition. *Trends in Cognitive Sciences*, *22*(6), 471-472. https://doi.org/10.1016/j.tics.2018.03.002

Rusch T., de Leeuw J., Chen L., Mair P. (2023). _smacofx: Flexible multidimensional scaling and 'smacof' extensions_. R package version 0.6-6, <https://CRAN.R-project.org/package=smacofx>.

Sandford, A., & Burton, A. M. (2014). Tolerance for distorted faces: Challenges to a configural processing account of familiar face recognition. *Cognition*, *132*(3), 262-268. https://doi.org/10.1016/j.cognition.2014.04.005

Shan, S., Gao, W., Chen, X., Cao, B., & Zeng, W. (2001). A face-unlock screen saver by using face verification based on identity-specific subspaces. In *Advances in Multimedia Information Processing—PCM 2001: Second IEEE Pacific Rim Conference on Multimedia Beijing, China, October 24–26, 2001 Proceedings 2* (pp. 1096-1101). Springer Berlin Heidelberg.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219-246.

Shepherd, J. W., & Deregowski, J. B. (1981). Races and faces—a comparison of the responses of Africans and Europeans to faces of the same and different races. *British Journal of Social Psychology*, *20*(2), 125-133. https://doi.org/10.1111/j.2044-8309.1981.tb00485.x

Shepherd, J. W., Deregowski, J. B., & Ellis, H. D. (1974). A cross-cultural study of recognition memory for faces. *International Journal of Psychology*, *9*(3), 205-212. https://doi.org/10.1080/00207597408247104

Shepherd, J. W., Gibling, F., & Ellis, H. D. (1991). The effects of distinctiveness, presentation time and delay on face recognition. *European Journal of Cognitive Psychology*, *3*(1), 137-145. https://doi.org/10.1080/09541449108406223

Strobach, T., & Carbon, C. C. (2013). Face adaptation effects: reviewing the impact of adapting information, time, and transfer. *Frontiers in Psychology*, *4*, 318. https://doi.org/10.3389/fpsyg.2013.00318

Takane, Y., Young, F. W., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, *42*, 7-67.

Tanaka, J. W., & Corneille, O. (2007). Typicality effects in face and object perception: Further evidence for the attractor field model. *Perception & Psychophysics*, *69*, 619-627. https://doi.org/10.3758/BF03193919

Tanaka, J., Giles, M., Kremen, S., & Simon, V. (1998). Mapping attractor fields in face space: the atypicality bias in face recognition. *Cognition*, *68*(3), 199-220. https://doi.org/10.1016/S0010-0277(98)00048-1

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*(1), 15.

Tversky, B., & Baratz, D. (1985). Memory for faces: Are caricatures better than photographs?. *Memory & Cognition*, *13*(1), 45-49. https://doi.org/10.3758/BF03198442

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, *79*(4), 471-491. https://doi.org/10.1111/j.2044-8295.1988.tb02747.x

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. https://doi.org/10.1080/14640749108400966

Valentine, T. (2001). 3 Face–space models of face recognition. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 83–113). Lawrence Erlbaum Associates Publishers.

Valentine, T., & Bruce, V. (1986). Recognizing familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology / Revue Canadienne de Psychologie, 40*(3), 300–305. https://doi.org/10.1037/h0080101

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception, 15*(5), 525–535. https://doi.org/10.1068/p150525

Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *The Quarterly Journal of Experimental Psychology*, *44*(4), 671-703. https://doi.org/10.1080/14640749208401305

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, *69*(10), 1996-2019. https://doi.org/10.1080/17470218.2014.990392

Vokey, J. R., & Read, J. D. (1988). Typicality, familiarity and the recognition of male and female faces. *Canadian Journal of Psychology / Revue canadienne de psychologie, 42*(4), 489–495. https://doi.org/10.1037/h0084202

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, *20*(3), 291-302. https://doi.org/10.3758/BF03199666

Webster M. A. MacLeod D. I. A. (2011). Visual adaptation and face perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1702–1725. https://doi.org/10.1098/rstb.2010.0360

Webster, M. A., & Maclin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review*, *6*(4), 647-653. https://doi.org/10.3758/BF03212974

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*(6982), 557-561. https://doi.org/10.1038/nature02420

White, D., Burton, A. L., & Kemp, R. I. (2016). Not looking yourself: The cost of self-selecting photographs for identity verification. *British Journal of Psychology*, *107*(2), 359-373. https://doi.org/10.1111/bjop.12141

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual review of psychology*, *69*, 105-129. https://doi.org/10.1146/annurev-psych-010416-044232

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141. https://doi.org/10.1037/h0027474

Young, A. W., & Burton, A. M. (2021). Insights from computational models of face recognition: A reply to Blauch, Behrmann and Plaut. *Cognition*, *208*, 104422. https://doi.org/10.1016/j.cognition.2020.104422

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *42*(11), 1166-1178. https://doi.org/10.1068/p160747n

Appendix

Table 1

Overview of literature discussed

Paper	Main points	Limitations	Implications for the proposed research
Valentine (1991)	Proposing the face space theory as a unifying account for inversion, distinctiveness, caricature, and other-race effects.	Assumes similar properties for representing familiar and unfamiliar faces.	Provides a basis for understanding how a population of faces are represented in memory by conceptualizing a psychological space where faces are encoded based on their similarity to other faces in the face space.
Light et al., (1979)	Inter-item similarity ratings correspond to facial distinctiveness, which predicts recognition performance.	Inter-item similarity ratings were obtained only using a single image per person, which does not account for within-person variability. Stimuli were only unfamiliar faces.	Describes inter-item similarity as a basis for how faces are represented in the face space. This assumption is central to the study I propose in Chapter 2 by collecting inter-item similarity ratings using multiple ambient photos for each identity.
Vokey and Read (1992)	Distinctiveness has two orthogonal components: memorability and familiarity. A distinctive face is high in memorability and low in familiarity.	Distinctiveness was assumed to be an invariant property of a face. Familiarity and memorability judgments were collected using one image per face, which does not account for within-person variability.	Assuming distinctiveness to be composed of memorability and familiarity is most relevant considering unfamiliar faces. Familiar faces are both memorable and familiar, which raises the question of how distinctiveness interacts with familiarity.
Lee et al. (2000)	Distinctiveness judgments of caricatures, anti-caricatures, and veridical images correspond to their theorized locations in the	Judgments were collected using one image per person, which does not account for within-person variability. Stimuli were only familiar faces.	The multi-dimensional scaling (MDS) approach will be adopted from this study to model a theoretical face space using inter-item similarity ratings using multiple ambient photos for each identity to understand how within-person

Paper	Main points	Limitations	Implications for the proposed research
	face space with caricatures in less dense regions.		variability is represented in the face space for familiar and unfamiliar faces.
Lewis and Johnston (1999)	The Voronoi-based model was proposed as an extension of the exemplar-based model of the face space where faces are represented as distinct identity regions where multiple inputs can activate a face's representation.	The size of a region corresponding to an identity is directly determined by distinctiveness, which does not account for familiarity. Since the Voronoi-based space is entirely filled with identity regions, the model cannot give an "unfamiliar" response.	The Voronoi-based model of the face space is a step toward conceptualizing the face space in a way that can accommodate variability across different viewpoints.
Leopold et al. (2001)	Adapting to an anti-face (e.g., anti-Adam) makes it easier to identify the target face (e.g., Adam), and makes an average face more likely to appear as the target face (e.g., Adam), suggesting that the architecture of the face space is norm-based.	Only unfamiliar faces were used, and only one image per identity was used, not accounting for familiarity and within-person variability.	Adaptation paradigms using anti-faces can provide insights into how the face space is constructed. As described in Chapter 4, the methodology of this study will be followed to understand how identity-specific sub-spaces are represented.

Paper	Main points	Limitations	Implications for the proposed research
Tanaka et al. (1998)	Proposed the attractor field model of the face space. Like the Voronoi-based model, faces are represented as identity regions that can accept multiple inputs as the face, accounting for changes in viewpoints.	The size of an attractor field is determined by distinctiveness, which does not necessarily explain the role of familiarity.	The attractor filed model can account for within-person variability. Understanding how the size of an attractor field changes as a face becomes familiar would provide insights into how familiarity is represented in the face space. I propose a study based on this in Chapter 3.
Laurence et al. (2016)	An important factor of the other-race effect (ORE) is that tolerance to within-person variability is less for other-race faces, which is predicted by the attractor field model.	Faces that were used in this study were unfamiliar to the participants, so it is not clear how the attractor field model can account for tolerance to within-person variability for familiar faces.	This study provides evidence that the attractor field model can make predictions about how within-person variability can be represented in the face space.
Burton et al. (2016)	Found that a computer model trained to recognize faces make more errors recognizing a non-target face after being familiarized with a target face, providing evidence for idiosyncratic within-person variability.	Does not provide behavioral data.	This paper proposes the existence of identity-specific sub-spaces – since within-person variability is idiosyncratic, each face must be encoded with idiosyncratic dimensions. In the proposed research, I explore the properties of such identity-specific sub-spaces.

Paper	Main points	Limitations	Implications for the proposed research
Burton and Vokey (1998)	Provided a mathematical account that truly typical faces are actually rare in the multidimensional face space, and that most faces are distinctive in at least one dimension.	Does not provide behavioral data.	The assumption that most faces are distinctive in at least one dimension raises the possibility that the perceived distinctiveness of a face can be subject to within-person variability, and that distinctiveness can be a photo-centred property as much as it is a person-centred property. The latter was the assumption in previous studies without explicit consideration of the former. I explore this in Chapter 2.
Faerber et al. (2016)	Familiar faces were rated as more distinctive as their unfamiliar anti-faces, despite being in regions with similar exemplar density in the face space.	Ratings were collected using a single photo per identity, which does not account for within-person variability.	I aim to gain further insight into perceived distinctiveness of familiar faces by collecting inter-item similarity ratings on a stimulus set that incorporates within-person variability, described in Chapter 2. Further, this research suggests that familiar faces migrate to more isolated regions in the face space due to their increased perceived distinctiveness. I propose to test this prediction in Chapter 3.
Chauhan et al. (2020)	Found that identity boundaries are narrower for familiar faces, as a 50-50 morph between familiar and an unfamiliar face was perceived to be more like the unfamiliar face despite being equidistant in the face space.	Subjects were asked which of the two *identities* was more similar to the morph, which means that subjects may have relied on different quality representations for the familiar and unfamiliar faces in the study.	The narrowed identity boundary account contradicts the predictions made by the attractor field model, which further raises the question of how a face’s representation changes in the face space as it becomes familiar.

Paper	Main points	Limitations	Implications for the proposed research
Laurence et al. (2022)	The magnitude of adaptation is the same for the same identity across different ages. This suggests that a single face representation hosts different ages.	A limitation identified in the context of the current research program is that it is unclear whether similar findings would be obtained for faces that change in dimensions other than age.	The assumption that there is a single representation for each familiar face that hosts different ages provides further support for the existence identity-specific sub-spaces, as it suggests that an identity is encoded with its own set of dimensions. If age was only a common dimension in the face space, there would be multiple representations for one face across different ages.
Koca and Oriet (2023)	An average representation is refined in memory as a face becomes familiar, suggesting that averaging is an underlying mechanism for face familiarization.	It is unclear how exemplars or within-person variability is represented for familiar faces. If exemplars are represented alongside the average, it is unclear why an average would be represented in the first place.	The average may serve as "the norm" for the identity-specific sub-spaces, allowing for within-person variability to be encoded as exemplars in reference to this norm. This norm-based model of the identity-specific sub-spaces are explored in Chapter 4.

Note. The papers are listed roughly in the order they are presented throughout the proposal. I selected papers that have direct implications for the studies that are proposed.