

## Formal models of familiarity and memorability in face recognition

THOMAS A. BUSEY,  
*Indiana University, Bloomington, Indiana*

*Note: Figures and Tables are at the end of the article for quick reference.*

The similarity structure of faces has long been recognized as an important mediator of face recognition. Distinctive faces have an enduring quality to them, such that they are rarely confused with other faces. At the same time, we often encounter a situation in which a particular face looks familiar, yet the individual may only bear a resemblance to several acquaintances. The veracity of these introspections has been borne out by empirical evidence, which has served to identify the information used during face recognition. Much of the research has focused on the role of *typicality*, which may be defined in various ways, but is often operationalized as a subjective rating of the difficulty of picking a particular face out of a crowd. Defined as such, typicality embodies the similarity structure of faces, such that typical faces will be similar to lots of other faces, while atypical faces will be very dissimilar and appear distinctive as a result.

This chapter provides an overview of the research on face perception that attempts to discern the role of typicality and the similarity structure of faces in face recognition. The primary discussion will revolve around the 'face-space' representation that was formalized by Valentine (1991a, b) as an extension of previous geometric models from the categorization literature to the area of face recognition. The face-space representation provides the basis for a discussion of the storage and retrieval mechanisms that may account for the effects of typicality described above. To motivate this discussion, extant forced-choice face recognition data is analyzed using a variety of process-oriented models that make predictions for individual faces in the face-space representation. The successes and failures of these models is used to draw conclusions about the nature of the representation of faces in memory and the retrieval processes that work to enable the recognition of faces. This chapter is written in part as a tutorial for those who wish to build quantitative models of face recognition that rely on similarity-based inputs.

The goal of this chapter is to test the 'face-space' representation by proposing an explicit input space in which the similarity relations between faces are separately measured and used to quantify the degree to which different faces influence each other during a recognition experiment. Many of the previous tests of the 'face-space' hypothesis have assumed only that faces are represented as points in a multidimensional space and are normally distributed along the different dimensions. These presumed minimal relations were then used to generate qualitative predictions for tasks such as recognition, classification or categorization. The current approach precisely specifies the 'face-space' representation and uses it to test quantitative models of face recogni-

tion. Although the current emphasis is on recognition, other aspects of the faces may also be computed: For example, a face's location in face space determines factors such as its distinctiveness and similarity to other items (see Busey, 1998; Johnston, Milne, Williams & Hosie, 1997). I will use similarity ratings derived from human participants as the measure of similarity, although other measures based on surface characteristics of the faces are also appropriate, as in principle components analysis (PCA; see O'Toole, Wenger & Townsend, this volume) or connectionist modeling of physical features (see Steyvers & Busey, this volume and Valentin, Abdi, Edelman & Posamentier, this volume for examples).

A variety of studies have demonstrated that distinctive or atypical faces have a characteristic advantage in recognition. Participants discriminate distinctive faces better than very typical faces, such that distinctive targets have high hit rates and low false alarm rates (e.g. Light, Kayra-Stuart, & Hollander, 1979; Vokey & Read, 1992). Typical faces tend to have slightly higher hit rates but produce low discriminability, which results from a very high false alarm rate that more than offsets the higher hit rate. Interestingly, typical faces may engender higher feeling of familiarity regardless of their status as targets or distracters, or in the parlance of an old/new recognition experiment, a higher feeling of 'oldness' (Bartlett, Hurry & Thorley, 1984).

These studies demonstrate that typicality ratings are at least related to those factors that affect recognition. Vokey and Read (1992) addressed the role of typicality in recognition with a principle components analysis of ratings of attractiveness, familiarity, likeability, typicality and memorability. They found that typicality ("how easy is it to pick this face out of a crowd") could be dissociated into two orthogonal components. The first consists of the attractiveness, likeability and familiarity ("how similar is this face to others that you know?") components. The second consists of the memorability rating ("how easy is it to remember this face?"). The typicality ratings loaded equally on the two components. This suggests that two processes affect typicality (and therefore recognition). The first is what Vokey and Read (1992) describe as *context-free* or *structurally-induced* familiarity. In this case, the to-be-identified face engenders a high feeling of familiarity, but there is no indexing of the source of the memory. Such a feeling of familiarity may be erroneously produced by mis-attributing the face to similar faces stored in memory, and thus typical faces are high in this context-free familiarity component. The second process is described as the familiarity *due to prior exposure*. With this process, the identifier matches the target to an item

in memory, or at the very least perceives the target face as more familiar as a result of the prior exposure. Distinctive items are thought to have an advantage that results from encoding and retrieval processes working on the distinctive elements of the face; as a result, distinctive faces tend to gain more familiarity due to prior exposure than typical faces (Bartlett et al, 1984).

The crucial aspect of this framework is that the recognizer is thought not to be able to distinguish between these two forms of familiarity. This results in a situation where typical faces engender high feelings of familiarity, in part through their similarity to other faces. This also leads to confusions, such that a typical distracter will have a high false alarm rate due to erroneous false matches to old items in memory. Distinctive faces have low structurally induced familiarity which will produce very low false alarms when these are used as distracters. However, the distinctiveness provides for easy encoding, making them memorable and giving them high hit rates that more than makes up for the initial low feelings of familiarity due to the context-free component (Bartlett et al, 1984).

O'Toole, Deffenbacher, Valentin and Abdi (1994) extended the work of Vokey and Read (1992) to digitized pictures of faces used as input to a neural network. They trained an associative neural network to recognize Caucasian and Asian faces and found that the memorability component of recognition was due to small, local distinctive features, while the familiarity component of recognition was related to more global aspects of the shape of the face. This reveals what might be a strategic use of information on the part of participants: if a small local feature such as a mole is highly predictive of a face, it will be used by the encoding system to access the context of the study event and provide strong discrimination. In the absence of such features, the recognition system is forced to rely on more generic face information such as shape. In this situation the face is evaluated for its overall familiarity, since the mechanism driven by the memorability component are not engaged by distinctive features.

Uttal, Baruch & Allen (1995a, b) provide evidence that suggests that the information that underlies discrimination may reside mainly in the higher spatial frequencies. This suggests that global shape information (contained in the lower spatial frequencies) may underlie a familiarity mechanism. As a result, what is seen as two mechanisms (familiarity-based and memorability based) may reflect the use of different spatial frequency ranges. Wenger and Townsend (under review) have made similar arguments. In later work, Uttal (this volume) suggests that a multitude of redundant mechanisms are at work, and that the mapping between spatial frequencies and information processing mechanisms may not be all that clean.

In addition to the processes that have been proposed to account for the effects of typicality, several authors have suggested the need for negative evidence. Vokey and Read (1992) found that the memorability component of typicality was correlated with the false alarm rates of typical and atypical faces, which produces the result that distinctive faces have very low false alarm rates. They argue that participants assess the

memorability or the retrieval potential of a particular face, and conclude that if this is high they would have remembered the face if it had indeed been studied. This suggests that subjects evaluate the evidence for a face having been previously presented, and compare that evidence against the likelihood that a face would have been encoded had it actually been studied. This suggests a role for metacognitive processes in the form of an evaluation of subjective memorability on the part of the subjects (Wixted, 1992).

Despite the intuitive appeal of the Vokey and Read (1992) model, there are interpretational problems with the data used to support such a model in face recognition. O'Toole, Bartlett and Abdi (submitted) discuss the difficulties that come from correlating some external rating such as a typicality rating with a dependent measure such as the hit or false alarm rate. For example, a high false alarm rate may result from either low discriminability ( $d'$ ) for typical faces, or a criterion shift in which participants relax their criterion for how much evidence they are willing to accept before calling a face 'old'. O'Toole et al conclude by calling for a model-based approach that makes predictions about which *individual* faces are easy to recognize. Such an approach should consider the similarity structure of the faces, and would have the added advantage of making the assumptions about the use of information explicit. Note that measures such as  $d'$  are important measures of face recognition performance, since the signal detection model presumably separates true sensitivity from any biases that may exist as a result of testing conditions or the location of a particular face in face space. The current recognition data uses a forced-choice paradigm and thus we will not be concerned with criterion shifts as they are usually defined.

The goal of this chapter is to propose and test a model that will account for aspects of the data that may correspond to the familiarity and memorability components described by Vokey and Read (1992). I will describe the foundations of the similarity structure that has been developed in the categorization literature (e.g. Medin & Shaffer, 1978; Nosofsky, 1986) and proposed by Valentine (1991a,b) to account for face recognition. This 'face-space' representation is then used as the input to a face recognition model that uses a sampling rule to account for the data from typical and atypical faces described above. I then test this model on forced-choice recognition data and demonstrate how it can make quantitative predictions. As we shall see, the model will have difficulty accounting for faces that are very similar to studied faces, and I will explore a variety of extensions that might account for these data as well. Finally, I will discuss some future directions for the use of geometric inputs to face recognition models. This discussion will also point out how major model assumptions were derived from existing memory models, in an attempt to adopt a tutorial tone.

## Geometric Models of Faces and Objects

The use of typicality in face recognition research has usually been operationalized as a rating on how easy a face would be to pick out from a crowd. Implicit

in this question is how similar a particular face is to other faces, or how much the face would stand out. One alternative to this approach is to measure the similarity between all pairs of faces in the experiment and compute typicality in terms of the similarity of a particular face to other faces. The role of similarity has been well worked out in the categorization literature, where it has been used in models to make predictions for prototype experiments and test decision rules in categorization experiments. In these experiments, training exemplars are used to construct a prototype stimulus that represents the central tendency of the exemplars. This stimulus is then used at test in a recognition or categorization experiment to assess the existence of a psychological prototype. The prototype is a novel stimulus and should be classified as such, but it is almost always classified as an old stimulus. Although such data are consistent with the existence of a prototype, alternative accounts are also possible. The prototype is by definition similar to the training exemplars, and Nosofsky (1986) has demonstrated that this similarity increases the overall familiarity of the prototype stimulus, and this alone can account for the prototype effect. Thus in many cases there may not be a need to propose a psychological prototyping mechanism.

Similar mechanisms have been proposed for faces. Byatt and Rhodes (1998), Rhodes, Carey and Byatt (1998) tested between a Norm-Based Coding representation, in which a face is compared against a central prototype face, and an exemplar based representation, in which each face is represented as a point in ‘face-space’. These two representations are notoriously difficult to distinguish between, in part because if the exemplars are normally distributed around the center of the space (where the putative prototype would be located) and can extend their influence to nearby locations, then the exemplar-based model acts like a ‘fuzzy’ prototype. For example, as a face gets closer to the center of the space where a prototype would influence it more, it would also get closer to other exemplars that cluster around the center, which would also influence the face more. Often one requires quantitative models to distinguish between these two representations, since qualitatively they produce identical predictions.

Quantitative predictions can be produced by a model that represents the similarity structure of the stimuli as its initial input. The similarity between any two faces can be measured by asking participants to make a similarity rating on a 9 point scale, and repeating this procedure for all pairs of faces. For an experiment with  $r$  faces, this requires  $\frac{r(r-1)}{2}$  ratings on  $\frac{r(r-1)}{2}$  pairs of faces. This provides  $\frac{r(r-1)}{2}$  datapoints, and a more efficient representation can be produced by submitting the similarity ratings to a multi-dimensional scaling algorithm such as ALSCAL. The output consists of an  $n$ -dimensional space (where  $n$  is usually less than 10) that represents each face as a point in this space. The dimensions are not specified by the experimenter, instead they emerge from the MDS procedure according to the dimensions along which faces differ. Gender, age, race, facial fatness, hair color and eye width are all possible dimensions that might emerge. Figure 1 shows an hypothetical ‘face-space’. The location of a face in this space can be used to de-

fine its similarity to other faces, and assuming a normal distribution around the centroid of the space, the most typical face will appear near the center of the space. Distinctive faces will appear at the fringes of this space.

This exemplar-based representation does not make direct predictions for recognition experiments, but it can be used as input to models that work on this representation. This defines the source of information used when recognizing faces: face-space based models assume that the similarity structure of the faces is used as input to some mechanism that will eventually produce an old/new response. This puts an enormous weight on the face-space representation, such that if it is missing some key dimension that is used in recognition, all models based on the representation will be incorrect as well. However, if the face-space representation accurately captures the dimensions that are important for recognition, the model can account for all the hit and false alarm recognition data using a few simple principles that are embodied in mathematical relations with a small number of free parameters. Thus the model can provide a succinct account of face recognition (and perhaps related tasks such as face/non-face classification) by quantifying a few principles into a process-oriented model that describes how information computed from the face-space representation is manipulated to produce a predicted recognition response.

## Face-Space Representations and Models of Recognition

Within the categorization literature, the use of the multi-dimensional scaling approach has been limited to relatively simple stimuli such as color chips, geometric figures, random dot patterns and random polygons. The advantage of such stimuli is that the experience provided by the training portion of the experiment is the only exposure the participant will have for a particular stimulus. In addition, these stimuli are either inherently low-dimensional, or if they are high-dimensional they are constrained to vary along only a few underlying dimensions (e.g. Edelman & Intrator, submitted). However, we have no way to control the prior exposure to faces, except to assume that participants are very experienced with faces and somehow take that into account in the modeling. As a start, we can assume that for novel faces, the similarity relations between the faces provides a representation that captures those dimensions that are relevant for face recognition.

Much of the work with geometric representations provided by MDS applied to similarity ratings assumes a representation such as that shown in Figure 1. Stimuli have values along different dimensions, and a variety of quantities can be computed. Most models assume that the distance  $d_{i,j}$  between any two faces can be computed from the locations in this space,

$$d_{i,j} = \sqrt{\sum_{n=1}^M w_n (x_{i,n} - x_{j,n})^2} \quad \text{Eq. 1}$$

where  $x_{i,n}$  is the coordinate for face  $i$  on dimension  $n$

(out of  $M$  total dimensions) and  $w_n$  is the attentional weight given to dimension  $n$  as described below. This corresponds to the Euclidean distance between faces  $c_i$  and  $c_j$ . Other metrics have been used, including the city-block metric, and this can be generalized via a Minkowski distance metric as

$$d_{i,j} = \left( \sum_{n=1}^M w_n |x_{i,n} - x_{j,n}|^b \right)^{1/b}$$

where  $b$  determines how the information on separate dimensions is combined. In general, for stimuli that tend to be processed holistically or integrally such as faces or colors, the Euclidean distance is appropriate,  $b=2$ . For stimuli that tend to be more separable such as abstract line drawing, the city-block metric is more appropriate ( $b=1$ ), and implies that participants make individual judgments on the separate dimensions and then combine the two decisions rather than compute one overall similarity when comparing two faces (Nosofsky, 1991). In one version of the model fitting I allowed  $b$  to freely vary, and the estimated value was quite close to 2.0.

The similarity,  $\eta_{i,j}$ , between faces  $i$  and  $j$  is defined as

$$\eta_{i,j} = e^{-cd_{i,j}} \quad \text{Eq. 2}$$

where  $c$  is a scaling parameter used to define the relation between distance and similarity (Shepard, 1974; 1987). There is a vast literature in support of this formulation, which Shepard (1987) goes so far as to describe as a universal law of generalization. Nosofsky (1987) demonstrates the ubiquity of this relation in tasks that are related to recognition. This re-computation of similarity enables a mapping of distance to similarity that systematically varies; high  $c$  values produce similarity values that are high only for very short distances and indicate that no item is very similar to any other item. Low  $c$  values imply that all faces bear some similarity to each other, and are difficult to distinguish.

The similarity structure provided by the similarities computed from the MDS distances provides the basic input to models. One such model that has been proposed by Valentine (1991a,b) to account for face recognition is the Identification version of GCM (Nosofsky, 1986, 1987). In this model, distinctive items are more likely to be encoded into memory, which expresses the memorability component described by Vokey and Read (1992). The model uses the similarity values from Eq. 2 to make a prediction for the probability of saying 'old'. For target faces, this value is,

$$P(\text{"old"} | i \text{ presented}) = F \left[ \frac{1}{\sum_{\substack{j \in \text{All Faces} \\ \text{In Memory}}} \eta_{i,j}} \right] \quad \text{Eq 3a}$$

and for distracter faces is,

$$P(\text{"old"} | i \text{ presented}) = F \left[ \frac{\text{Max}_{\substack{j \in \text{All Faces} \\ \text{In Memory}}} [\eta_{i,j}]}{\sum_{\substack{j \in \text{All Faces} \\ \text{In Memory}}} \eta_{i,j}} \right] \quad \text{Eq 3b}$$

where  $F$  is a logistic function,

$$F(x) = \frac{1}{1 + \beta e^{-\theta x}} \quad \text{Eq. 4}$$

with free parameters  $\beta$  and  $\theta$  that map the ratio in Eq 3 into the range of 0 to 1.

The form of the ratio in Eq 3 should provide some intuition for why Valentine (1991a,b) proposed this formal model for face recognition. First, consider the denominator in Eq 3a. When a face is tested in an old/new recognition experiment, the similarity to all other items in memory is computed. Faces that are very atypical tend to lie near the edges of this space, and will therefore will not be similar to many other faces. Thus, the summed similarity from the numerator will be small, making the overall fraction large. Distinctive target faces will therefore have a very high probability of saying old on the basis of the denominator. Typical targets have a larger denominator and thus an overall smaller probability of being called an old face.

While this model predicts the high hit rate to distinctive target items, it may have difficulty accounting for the low false alarm rates to the distinctive distracters. Previously, Vokey and Read (1992) argued that such a situation requires the use of negative evidence, which the Identification version of GCM does not contain. Negative evidence predicts low false alarm rates to distinctive distractors according to the following logic that invokes a notion of subjective memorability (e.g., Gentner & Collins, 1981). Under this theory, participants are aware of the fact that distinctive faces are more memorable than typical faces. During test, when faced with a very distinctive distractor, participants recognize the distinctiveness and assume that if they had studied this particular face it would have been very memorable and therefore they would have remembered it. Thus distinctive distractors can be confidently rejected, in part on the basis of an analysis of the stimulus properties, not because these faces are particularly unfamiliar. Thus subjective memorability may be thought of as a metacognitive process that is somewhat separate from the computation of familiarity based on a comparison of the test item to items in memory.

Although the Identification version of GCM does not contain an explicit assumption of subjective memorability, components of it may reflect this process implicitly. The denominator in Eq 3b will be very small for distinctive distracter which would lead to a higher false alarm rate than for typical distracters. Opposing this tendency is the numerator, which tends to be larger for typical distracters. How the numerator and denominator trade off depends in part on the similarity structure of the face space, and thus quantitative model

predictions are required to evaluate the adequacy of the Identification model.

In summary, the Identification model includes a mechanism that has the properties associated with the memorability component of Vokey and Reed's framework. It may or may not include the familiarity component, which in part may depend on the structure of the face space and the ability of nearby targets to produce false alarms for typical distracters via the numerator of Eq 3b.

### Applications to Forced-Choice Face Recognition

While most models of recognition memory are applied to old/new picture recognition paradigms, the legal setting provides an important forced-choice situation. In a lineup, a witness may often assume that the suspect is present in the lineup, and use a comparison between the faces to name a suspect. Vokey and Reed's breakdown of familiarity into context-free and that provided by previous study raises an interesting possibility for the lineup situation. For example, typical faces tend to induce more context-free familiarity. Distinctive target faces begin with less context-free familiarity but benefit more from study (Bartlett et al, 1984). Consider a situation in which a target and a distracter face are compared in a forced choice task. The target face will have study-induced familiarity in addition to some amount of context-free familiarity. However, if the distracter face is very typical, it may have a large amount of context-free familiarity, causing the participant to choose the distracter over the target. At the very least, such a comparison would be more difficult than if the target and distracters are both distinctive. In an old/new recognition experiment, Solso and McCarthy (1981) demonstrated that prototype faces could attract a large number of false alarms, suggesting that the familiarity induced by the similarity to studied faces could translate into a false recognition. These studies used identikit line drawings that re-combined features from studied faces to produce the prototype, and therefore some of high false alarm rates may be due to misrecognition of individual features rather than the entire face. However, this work does demonstrate that substantial confusions can take place between a test distractor and several studied items, which has a similar effect as the structurally-induced familiarity described above.

The lineup situation is complicated somewhat by assumptions that the witness might make when making an identification (e.g. Wells & Lindsay, 1985). In the present case we will limit ourselves to the case in which exactly one face in a two-alternative forced-choice comparison was presented at study. The data described below was briefly described in Busey and Tunnicliff (submitted). A summary of this experiment is provided below.

#### Experimental Design and Procedures

The stimuli used in this experiment were photos of bald men that ranged in apparent age from mid-twenties to mid 50s (Kayser, 1985). As describe above, the typicality of a particular face is an important mediator of memory performance. In addition to the naturally occurring differences in typicality, we included 16 faces

that were constructed by *morphing* two parent faces. These morphs were included only in the test portion of the experiment, and were used because the morphs tend to be highly typical. At the very least they are similar to the parent faces, and due to the geometry of MDS space, the morphs might be closer to many other faces as well (Busey, 1998). Thus these morphs may provide a stimulus that induces a large amount of context-free familiarity. The parent faces provide the appropriate comparison stimulus, since they are both studied and more distinctive. If the two-process framework described by Vokey and Read is correct, and if the context-free familiarity induced by the typicality of the morphs dominates, then we might find that participants choose the morph over one of the parents in forced choice.

The details of this experiment are provided by Busey and Tunnicliff (submitted), but the essential details are reproduced below. Participants were 119 Indiana University undergraduates who participated in one of 24 groups of up to 5 participants at a time. They received course credit for their participation. The stimuli consisted of 104 pictures of bald men with neutral expressions. Twenty-one of the men had facial hair. Fourteen of the men were black and the rest were Caucasian. Sixty-eight faces were selected for the study portion of the experiment. Thirty-two faces were selected to be parent faces for the morphs. These faces were paired so that 8 pairs had faces that were dissimilar, while 8 pairs had faces that were similar according to a pre-experiment sorting task. This manipulation allows us to evaluate the effect of similarity on the psychological mechanisms that underlie the responses to the morphs. Sixty-eight faces were selected for the study portion of the experiment, including 36 target faces and 32 parent faces. The parent faces were combined to create 16 morph faces as described below. There were 20 distracter faces selected from the faces. The constraints placed by the morphing procedures did not allow us to select faces at random for the parent faces, since faces with facial hair do not morph well. Faces with facial hair tend to be more distinctive, which may influence the forced-choice data. However, we know the structure of MDS 'face space' and therefore will be able to take these differences into account.

Control points were placed on the salient features of each parent face and 50% averages were created using the Morph™ software package (Gryphon Software). At least 150 control points were placed on each parent, and control points were added as required to remove obvious artifacts in the resulting morph. Data was collected by a PowerMac computer using 5 numeric keypads that provided identifiable responses from each keypad. The faces were displayed on a 21" Macintosh grayscale monitor.

There are four types of faces in each experiment. Target and parent faces appear both at study and at test; the only distinction between the two sets of faces is that parent faces tended to be less distinctive because they were all clean shaven. The target faces were a mix of clean-shaven and mustached faces. Morphs and distracters appeared only at test and are therefore distracters. However, the morphs are similar to the parents and as a result we expect higher false alarm rates in

general to the morphs than to other distracter faces.

Participants were asked to view a series of faces in the study phase and remember them for the subsequent recognition test. There were 68 faces in the study phase: 36 target faces (faces not used for morphs but would reappear in test phase) and 32 parent faces (faces previously used to create the morphs). Each face appeared for 1500 ms followed by a two second delay between each face.

At test, participants were given a forced choice recognition test. Participants were required to pick one of two faces presented that was previously studied. Participants either chose between a morph and one of the two parents, or between a target and a distracter. There was a total of 36 faces in the test phase: 16 morph/parent pairs, 20 target/distracter pairs in random order. Although there were 36 targets presented at study, only 20 randomly-chosen targets were tested since we have only 20 distracters.

### Results

The mean probability of choosing a target over a randomly-chosen distracter is .765 (standard error of the mean, SEM=0.015). When comparing the morphs and parents constructed from similar parents, the probability of choosing a similar parent over its associated similar morph is .463 (0.222), which is statistically significantly less than 0.5 ( $t(1448) = 2.12, p < 0.05$ ). Morphs from dissimilar parents show the opposite effect: the probability of choosing a dissimilar parent over the morph is 0.658 (0.161), which is greater than 0.5 ( $t(1448) = 10.6, p < 0.05$ ).

These results provide tentative evidence in support of the framework suggested by Vokey and Read (1992). Target faces tend to be very distinctive, because many had facial hair. This distinctiveness may have resulted in a large amount of familiarity due to the prior exposure. Similar morphs are similar not only to the two parents, but also to many other faces (Busey, 1998). As a result, they may have engendered a large amount of context-free familiarity and therefore have been chosen over the parent face in the forced-choice comparison. This suggests that the morphing process provides a reasonable technique for producing novel stimuli in face space in order to control and manipulate the degree of structurally-induced familiarity.

Although these results are consistent with the familiarity and memorability view of Vokey and Read (1992), there are several aspects of this framework that are troubling. First, it is not clear that context-free familiarity is a distinct construct that is separable from familiarity due to prior exposure. Clearly the face at least must be recognized *as* a face, which must require some form of active search through memory. There may also be an additional search through memory that corresponds to the familiarity due to the prior exposure. As a result of this overlap between the two processes, a single-process model might be able to account for both the good discriminate of the target faces and the errors made by participants to the similar morphs. One possible starting point is the Identification version of the GCM model developed by Nosofsky (1986; 1987) and suggested by Valentine and Ferrara (1991) as a good model for face recognition. Below I describe how this

model can be extended to forced-choice data and demonstrate the adequacy of this model.

### Measuring Face Space: Similarity Ratings

Before a model can make quantitative predictions for individual faces, the similarity structure of the faces must be measured. The procedures used to gather similarity ratings and produce the multidimensional scaling output are described in Busey (1998), but are briefly sketched below. A set of 104 faces requires multiple similarity ratings on all (100\*99/2) pairs of faces. This required 373 Indiana University undergraduates making ratings on 177 randomly-chosen pairs of faces, on a scale from 1 (most similar) to 9 (least similar). These similarity ratings were submitted to the ALSCAL multidimensional scaling algorithm, which produced a 6 dimensional solution. Because the program could only handle 100 stimuli, 4 target faces were selectively deleted from the set. The dimensions of the solution were all interpretable, and included dimensions such as age, race, facial pudginess, and facial hair.

### Accounting for Forced-Choice Data

One possible extension of the Identification model to forced choice data would be to consider the model's predicted familiarity for both faces, and whichever face produces the higher familiarity is the face that is selected. While intuitively plausible, this model cannot be correct, because without noise or some other process it would always predict that the target would be chosen over the distracter with probability 1.0 (assuming that the distracter did not have more context-free familiarity than the target). In the data, the targets were chosen over distracters about 77% of the time, while dissimilar parents (which tend to be less distinctive) were chosen over the dissimilar morphs 66% of the time. Any model must account for these gradations in choosing rates that appear to depend upon typicality or the similarity structure of the faces.

In an old/new recognition task, the participant typically makes either an 'old' or a 'new' response that is presumably based on some internal value that reflects the test face's familiarity or match to items in memory. In some sense this is a categorization task of each test face into either the old or new category. Most categorization tasks also include two categories, in which members of category *a* are distinguished from the members of category *b* according to some criteria. What distinguishes recognition from most categorization tasks is that the population of new items in recognition is generally unknown. A forced-choice task is much closer to the categorization task, since the target face is directly compared with a known quantity, the distracter face. In categorization, the probability that test item *i* is classified as a member of category *a* is,

$$P("a" | i) = \frac{G[A]}{G[A] + G[B]} \quad \text{Eq. 5}$$

where *A* is the evidence that item *i* belongs to category *a*, *B* is the evidence for category *b*, and *G* represents a monotonic transform. Often, *G* takes the form of an exponential,

$$P("A" | i) = \frac{e^{\zeta A}}{e^{\zeta A} + e^{\zeta B}} \quad \text{Eq. 6}$$

where  $\zeta$  represents the extent to which small differences between the evidence for faces  $A$  and  $B$  are magnified into a large likelihood of choosing face  $A$ . For example, for small  $\zeta$ , virtually all choosing probabilities will be close to 0.5, since the exponents will all be close to 0. However, for large  $\zeta$ , this emphasizes the impact that  $A$  and  $B$  can have, such that if  $A$  dominates  $B$  only slightly, the participant will be very likely to say "A". Thus  $\zeta$  can be thought of as a confidence parameter that indicates how much the evidence of  $A$  over  $B$  influences the resulting choosing rate. It also may reflect to some degree the noisiness of the comparison process, since if  $A$  and  $B$  are similar and the system is noisy, the participants would choose  $B$  on some proportion of the trials. The model would mimic this behavior by reducing the probability of choosing  $A$  by having a fairly small  $\zeta$  parameter. An alternative interpretation is offered by Nosofsky and Palmeri (1997), in which the  $\zeta$  parameter represents the amount of evidence that must be accumulated by a random walk before it reaches threshold.

Eq. 6 can be used to adapt the Identification version of GCM for the forced-choice recognition paradigm if we assume that the participant computes familiarity of faces  $a$  and  $b$  via Eq 3 or some other process, which provides the values  $A$  and  $B$  for Eq 6. In situations where  $A$  and  $B$  are about equal (that is, both faces  $a$  and  $b$  seem equally familiar), the probability of choosing face  $a$  will be close to 0.5. However, as one face tends to dominate, Eq 6 will get closer to 1.0.

The forced-choice data were fit as follows. The data from target faces consists of the probability of choosing a given target face over one of the distracters. Over the course of the experiment, each target face was tested against 24 randomly-chosen distracters, and the modeling must reflect this. This was accomplished by computing the familiarity of each target and distracter face (as expressed as the probability of saying old) via Eqs 1-4, and then computing the probability of choosing the target face over a given distracter via Eq 6. These probabilities were then averaged over all such comparisons involving that particular target face. This process was repeated for the parent, morph and distracter faces, although morphs were always compared only with their parent faces and vice versa. Thus the forced-choice predictions reflect the degree to which one face seems more familiar than the other, as defined by Eq 3 for the Identification model.

One issue that has been raised in the categorization literature is the idea that participants may selectively attend to one dimension over another. For example, in recognition, age may be a particularly salient dimension, while other dimensions such as the color of the facial hair may be less so. The MDS procedures normalize the dimensions, and to compensate we add 6 weight value that allow the dimensions to have differential effects on the computation of distance (and therefore similarity) via Eq 1. These are constrained to sum to 1.0, and so this adds 5 free parameters to the model.

This model has 9 free parameters; the similarity gradient parameter  $c$ , the 5 weight parameters, and  $\beta$  and  $\theta$  that map the ratio in Eq 3 into a familiarity (probability of saying old), and  $\zeta$  that determines how the evidence for face  $a$  is compared with the evidence for face  $b$ . The best-fitting parameter values are given in Table 2.

Figure 2 shows the fit of the GCM-Identification model, with the probability of choosing a face from the data on the abscissa and the theory's predictions on the ordinate. One measure of the model's goodness of fit is the root-mean-squared error (RMSE), which was 0.120. Overall the fit is not bad; in general the points fall on the diagonal. However, there are systematic deviations for some of the target faces, the similar parents and the distracters. Most telling is the failure to account for the fact that participants tend to choose the similar morphs over the similar parents; the model places the similar morphs (upright open triangles) to the left of the similar parents (upright filled triangles), where the reverse should be true.

Despite these failings, overall the model is accounting for the distinctiveness effects seen in the faces. The target and distracter faces tend to be more distinctive than the dissimilar parents and dissimilar morphs. We see in Figure 2 that the targets have a higher choosing rate than the dissimilar parents, which demonstrates that the model can account for the effects of distinctiveness. Thus, the Identification model might be associated with the Memorability component described by Vokey and Read (1992). What it apparently lacks is some mechanism to account for the similar morphs, which are very typical distracters. This might either require a better model formulation or a separate familiarity mechanism to include context-free familiarity.

## The SimSample Model

Before adopting a second process to account for something like context-free familiarity, consider an alternative model that might account for both the effects of highly typical and highly distinctive faces. This model involves sampling from memory according to the similarity of the target face to items in memory, and thus I term it SimSample. This model has previously accounted for old/new recognition data (Busey & Tunnicliff, submitted) and might provide a better account of the forced-choice data as well.

Before describing the model, I'd like to describe how it was derived, as motivation not only for its assumptions but also to provide a tutorial on the modeling process. Additional information on the process of model building can be found in Shiffrin and Nobel (1997). Memory models that rely on relations between stored vectors of features have been remarkably successful at accounting for a variety of recognition, cued- and free-recall tasks using works. Examples include CHARM (Metcalfe, 1990). Minerva 2 (Hintzman, 1986), and SAM (Gillund and Shiffrin, 1984). Although SAM is technically not a vector model, it shares some of the characteristics of the other models. In these models, a test item is used to probe memory in such a way that a trace may be sampled and potentially

recovered. Similarity is often represented abstractly rather than in terms of an MDS space. However, the model structures are similar.

In developing SimSample, we observed that in the free- and cued-recall literature, the memory models were exhibiting behavior that might correspond to those in our data. In particular, the models had a tendency to be better at recovering distinctive items such as low frequency words. In addition, if a cue was similar to lots of studied words, the cue might produce an incorrect item (vs no item at all in the case of a distinctive cue in an error is made). This seemed to be analogous to our high false alarm rate for morphs in old/new recognition. The models all differ in their assumptions, but in general the recall mechanisms involve some sort of sampling process. In the case of SAM, one item is sampled, but may not be recovered. This sampling process has a tendency to favor distinctive items (if something is recalled at all) because they are in an isolated region. Items in a dense region have a tendency to recall incorrect item information due to confusions with similar items.

These model mechanisms were developed primarily to handle recall effects; most models had separate mechanisms for recognition. However, the two mechanisms often produce highly correlated outputs, because the item that is the most likely to be recovered in recall also contributes the most to the familiarity mechanism that is assumed to underlie recognition. In addition, we felt that there may be a role for recall in recognition, as suggested by Yonelinas, Dobbins, Szymanski, Dhaliwal and King (1996). As a result, we felt comfortable using a recall-based mechanism for our recognition data.

To develop the SimSample model we assume that similarity is constructed from the MDS face space according to Eqs 1-2. We then assume that for each test face in the forced-choice comparison, the test face is used to probe memory, and exactly one face is sampled from memory. Not all items are equally likely to be sampled, however. The probability that the observer samples face  $k$  in memory given face  $i$  was presented at test is,

$$P(\text{sample } k/i \text{ presented}) = \frac{\eta_{i,k}}{\sum_{\substack{j \subset \text{All Faces} \\ \text{In Memory}}} \eta_{i,j}} \text{ Eq 7}$$

which is simply the Luce Choice Rule. This function has two nice properties: first, it constrains the probability of sampling *something* to 1.0. That is, the sum of Eq 7 for all items  $k$  is 1.0. Second, it has the property that the similarity between  $i$  and  $k$  is relative to the summed similarity of the test item ( $i$ ) to all other faces. Thus this makes the similarity between  $i$  and  $k$  sensitive to the density of  $i$ . Typical faces and distinctive faces will have different denominators and this will affect how the similarity between  $i$  and  $k$  is evaluated. This is a critical aspect of the model that will be explored below. The Luce Choice sampling rule is adapted from the Search of Associative Memory (SAM)

model proposed by Gillund and Shiffrin (1984), although SAM uses strengths rather than similarities to compute the Luce Choice ratio. To characterize the relation between SimSample and SAM (and thereby provide support for the assumptions underlying SimSample) below I describe the nature of the sampling and familiarity process.

When the SAM model is used to predict free- or cued-recall data (usually words), context or a test cue is used as a probe to sample items from memory. One item is always sampled from memory, but it may or may not be recovered. If the item information is recovered, the participant reports the contents (i.e. the studied word). This process of sampling and retrieval continues until either all words are recovered or some other stopping rule is applied. A separate mechanism has been proposed for recognition, in which the information gained from sampling all items in memory is combined to produce an overall familiarity value for a test item. This familiarity mechanism is more akin to Nosofsky's GCM, although similarity is defined not in terms of distance in MDS space but instead by the feature overlap of two items. Although the sampling component of SAM has been associated with recall, in more recent work Shiffrin, Huber and Marinelli (1995) have suggested that there may be a recall component in recognition. Under this assumption, participants would sample once from memory, and if the test item is recovered, participants would assess the recovered information and respond old or new, rather than assessing an item's overall familiarity. If no recovery is made, participants instead respond on the basis of the familiarity computation which assesses the test item's similarity to all items in memory. In principle these are separate mechanisms, but in practice they produce highly correlated results since the familiarity process tends to be dominated by one or two traces in memory, and those are the ones that primarily affect the sampling and recovery process. Thus it is reasonable to assume that a sampling process could be at work in face recognition. As we will see, the sampling process of SimSample implicitly contains elements of both a recall and a familiarity-based recognition system.

In order to make predictions for old/new recognition within the SimSample model, we assume exactly one face is sampled from memory. This is different from SAM, which allows multiple sample and recovery attempts. We assume that the sampled face is compared with the test face, and if they are similar enough the observer concludes that they have a match and says "old". This involves a criterion such that if face  $k$  is sampled when face  $i$  is used to probe memory,

$$\text{Say "old" if } \eta_{i,k} > \text{criterion} \quad \text{Eq 8}$$

where the similarity criterion is a free parameter. If the similarity between the sampled item and the test face is less than the criterion, the model predicts that the observer will say "new". More formally, we can compute the probability of saying old to item  $i$  as the probability of sampling all items that are similar enough such that if sampled, the observer would say old. Define function  $\theta(\eta_{i,k})$  such that

$$\Theta(\eta_{i,k}) = \begin{cases} 1 & \text{if } \eta_{i,k} > \text{criterion} \\ 0 & \text{if } \eta_{i,k} < \text{criterion} \end{cases} \quad \text{Eq 9}$$

which is simply the probability that the observer will say old to item  $i$  given item  $k$  is sampled. The probability that the observer says old when viewing face  $i$  at test is,

$$P(\text{"old"} | i \text{ presented}) = \sum_{\substack{k \text{ faces in} \\ \text{memory}}} P(\text{sample } k | i \text{ presented}) \Theta(\eta_{i,k}) \quad \text{Eq 10}$$

where the first term inside the summation comes from Eq 7 and the second from Eq 9.

For a variety of reasons it is reasonable to assume that the similarity criterion in Eqs 8 and 9 is not fixed, but has normally distributed variability due to some internal noise or differences across participants. In this case, we can redefine Eq 9 according to a cumulative gaussian function with mean set to the criterion and standard deviation set to a free parameter  $\text{critSD}$ ,

$$\Theta(\eta_{i,k}) = \int_{-\infty}^{\eta_{i,k}} \frac{e^{-(x-\text{criterion})^2 / 2\text{critSD}^2}}{\sqrt{2\pi}\text{critSD}^2} dx \quad \text{Eq 11}$$

which implies that if  $\eta_{i,k}$  equals the criterion, the probability that the observer says old when face  $k$  is sampled is 0.5. No modification of Eq 10 is necessary to accommodate this change to  $\Theta(\eta_{i,k})$ .

As stated above, the sampling and criterion assumptions embodied by Eqs 7, 10 and 11 are related to the sampling and testing processes of the SAM model, although in SAM the model is allowed to sample multiple times, whereas the SimSample model is only allowed to sample once. Various multiple-sampling versions of SimSample were attempted, with little success.

The SimSample is extended to account for the forced-choice data by assuming that the subject computes the probability of each item having been previously presented (i.e. the probability of saying 'old' to each item) and then uses these probabilities via Eq 6 to predict the choosing rate for the target stimulus. The model has 9 free parameters (which is the same number as the Identification model): 1 generalization gradient parameter  $c$ , 5 attention weights, the response criterion and the standard deviation of the response criterion and  $\zeta$ , which controls the comparison behavior between the two faces.

### Accounting for Distinctiveness

At a minimum, the SimSample model must account for the finding that participants are very good at recognizing distinctive targets and rejecting distinctive distracters. The upper-left panel of Figure 3 demonstrates how the SimSample model accounts for the high hit rates to distinctive targets. A distinctive target is not similar to many other items in memory, making the denominator in Eq 7 small. When sampling its own

item in memory, the numerator in Eq. 7 is 1.0, and for all other faces the numerator is much less than 1.0. This implies that distinctive faces are very likely to sample their own image in memory, and of course when they do,  $i = k$ , and  $\eta_{i,k} = 1.0$ , which exceeds the similarity criterion in Eq 8. Less distinctive targets are less likely to sample themselves in memory, since even though the numerator is still 1.0 in Eq 7, the denominator is larger for more typical faces. When a moderately typical test face samples other faces in memory, they may be far enough away such that  $\eta_{i,k} < \text{criterion}$  and the observer will incorrectly say 'new'. Thus the SimSample model correctly predicts that more distinctive target faces will be more likely to be chosen over a distracter than less distinctive targets.

The SimSample model can also account for the fact that distinctive distracters are easily rejected by observers (and are not often chosen in forced choice), as demonstrated by the upper-right panel of Figure 3. As with target faces, a distinctive distracter will sample some face in memory. However, it cannot sample itself because it wasn't placed into memory at test. If there are no faces near enough to fall inside the criterion in MDS space, the observer will never make a false alarm. The noise added to the criterion insures that all distracters have above-zero false alarm rates, but the model will produce very few false alarms. More typical distracters will have a greater chance of being near a target face that is inside the criterion, which if sampled will produce a false alarm or be erroneously chosen in a forced-choice paradigm. Thus the SimSample model can account for the low false alarm rates to distinctive distracters without assuming negative evidence as suggested by Brown, Lewis and Monk (1977).

### Accounting for Familiarity

In addition to accounting for the effects of distinctiveness, the SimSample model can in principle account for the fact that very typical faces engender a high feeling of familiarity. The bottom panels of Figure 3 demonstrate how SimSample can in principle account for the high false alarm rates to the morphs created from similar parents, as well as the relatively high hit rates to typical parents. When a morph is used to probe memory, it cannot sample itself because it was not presented at study. However, it does have the opportunity to sample nearby items in memory and will produce a false alarm if the sampled item is inside the criterion. In the case of the morphs created from similar parents, there are likely to be at least two studied faces (the two parents) that are similar enough to fall inside the criterion. In addition, the morphs tend to be among the most typical of faces, since the morphing procedures tend to place the morphs near the middle of MDS face space (Busey, 1998). Thus the SimSample model correctly predicts higher false alarm rates to the morphs than to more distinctive distracters.

This same explanatory principle can account for the fact that very typical parents have higher hit rates than moderately typical parents, as seen in Figure 3. Typical parents are likely to be similar to lots of other faces in memory, and even though such a face is not very likely to sample its own trace in memory, it is very likely to sample a nearby face. Typical parents

have lots of other faces nearby, and if one of these is sampled the observer will say old. When this happens, the observer is making a correct response but doing so for the wrong reason. Less typical parents have fewer opportunities to sample nearby faces that would generate an old rating, and therefore cannot take advantage of incorrect samplings that result in a correct decision.

The fit of the SimSample model is shown in Figure 4, and the best-fitting parameters are provided in Table 1. The RMSE is 0.112. The fit is an improvement on the fit of the Identification model despite the fact that it has the same number of free parameters. The RMSE is reduced, and the systematic deviations for the distracters and targets are no longer present. However, the model still has difficulty with the similar morphs and parents: the similar morphs are still systematically to the left of the similar parents. Thus either the model cannot account for very typical faces, or there is some other mechanism such as noise, clustering or blending that is going on that might account for these faces.

### *Extensions to the Exemplar-Based Model*

One possible explanation for the high choosing rate for the similar morphs is that there is enough noise in the recognition system such that the morph is confused with one or both of the parent faces. This seems somewhat unlikely given the fact that one of the parent faces is shown with the morph, and participants know that one and only one studied face is shown at test. Nevertheless, noise may play a role in the false recognition of the morphs, since the central location of a prototype may make it more immune to noise than the exemplars. One mechanism to introduce noise into the locations of the faces in MDS space is to assume a gaussian similarity gradient rather than an exponential gradient,

$$\eta_{i,j} = e^{-c(d_{i,j})^2} \quad \text{Eq. 12}$$

which tends to make the sharp drop of the similarity gradient 'fuzzy'. This provided a very slight improvement in the RMSE, reducing it to 0.108. However, it could not predict that participants would choose the similar morphs over the similar parents.

A second mechanism which might save the exemplar-based version of the SimSample model is to assume some sort of clustering mechanism that might work to bring similar faces even closer together. This clustering goes against other effects of density as described by Krumhansl's Distance-Density hypothesis (Krumhansl, 1978), where experience with a dense region tends to make the items in that region appear *less* similar, not more. A clustering algorithm was appended to the SimSample model according to the following logic. Studied items were placed into memory at their locations in MDS space. If they were close enough to other items (as determined by a free parameter), all items inside this pre-defined region were systematically moved closer to each other by an amount proportional to their distance and a free parameter. This mechanism reduced the RMSE only slightly, to 0.107, and did not predict that the participants would choose the similar

morphs over the similar parents. Thus it appears as if a clustering mechanism cannot help the SimSample fit.

A third mechanism that might help the morphs is the assumption of a global prototype that exists in addition to the individual exemplars. Such a model assumption is similar to the norm-based coding model proposed by Byatt and Rhodes (in press), Rhodes, Carey and Byatt (in press) and Valentine and Bruce (1986). In this model, some form of blending or abstraction mechanism is at work that creates a new trace in memory that represents a global prototype of all bald men. We do not know the location or strength of this prototype, but we can estimate its values on each dimension and its strength as free parameters. Thus to the SimSample model I added 7 more parameters: 6 parameters dealing with the location of the prototype on the 6 dimensions, and 1 free parameter that determined the strength of the prototype when computing its contribution to the probability of saying old via the SimSample process. This model only reduced the RMSE to 0.111, which is not a significant reduction in error given the addition of 7 free parameters (see Table 1 for F-values). In addition, the model did not reverse the discrepancy between the similar morphs and parents. Thus a global prototype does not seem to be a plausible extension to the SimSample model.

It is interesting to note that the global prototype could have completely dominated the individual exemplars by choosing a very large prototype weight. This would have been equivalent to the Norm-Based Coding model, which assumes that faces are coded relative to a global prototype rather than as individual exemplars. The failure of this model casts doubt on the norm-based coding model, and demonstrates that quantitative predictions are necessary to distinguish a prototype model from an exemplar model.

### *Prototype Extensions to the Exemplar-Based Model*

Given the failure exemplar-based or global prototype mechanism to account for the finding that participants choose the similar morphs over the similar parents, an alternative is to propose individual prototypes that form between parent faces and correspond to the morph locations. Similar extensions have been proposed in categorization work (e.g. Homa, Goldhardt, Burrue-Homa, & Smith, 1993). As with the previous prototype theories, these prototypes would act like faint traces in memory at the locations of the morphs in MDS space and could in principle help the similar morphs be chosen over the similar parents. The morphs were included in the original similarity rating experiment and so we know the locations of the morphs in MDS space. The morphing operation introduces artifacts into the resulting blended face (Busey, in press) moving it away from the midpoint in MDS space between the two parent faces. However, by including the morphs into the scaling solution, we presumably have eliminated these biases by informing the model of the morph's true location.

One possible prototyping mechanism would blend nearby faces, creating a prototype trace in memory that would be treated from the perspective of a model as a

faint version of a real trace. The strength of the prototype affects both the likelihood that a prototype is sampled, as well as the probability of saying old if it is sampled. In general, when sampling items from memory, the probability that face  $k$  is sampled (where  $k$  can now be either a parent, a target or a morph) is,

$$P(\text{sample } k/i \text{ presented}) = \frac{\eta_{i,k}}{\sum_{j \in \text{All Faces In Memory}} \eta_{i,j} + \sum_{j \in \text{prototypes}} pw \eta_{i,j}} \quad \text{Eq 13}$$

for faces actually studied, and,

$$P(\text{sample } k/i \text{ presented}) = \frac{pw \eta_{i,k}}{\sum_{j \in \text{All Faces In Memory}} \eta_{i,j} + \sum_{j \in \text{prototypes}} pw \eta_{i,j}} \quad \text{Eq 14}$$

for the morphs. In both Eq 13 and Eq 14,  $pw$  is the prototype weight that determines the strength of the prototype in the sampling process. Once a face has been sampled (and now a prototype may be the sampled face), the probability that the observer says old is related to the similarity between the test face and the sampled face as in Eq 8. This is modified such that if the morph is sampled, the similarity used to compute the probability of saying old via Eq 9 is reduced by the prototype weight. This is in keeping with the idea that the prototype trace is fainter than a real face's trace, and this influences both the sampling and decision processes<sup>1</sup>. This assumption implies that we compute  $\Theta(pw \eta_{i,k})$  when a prototype is sampled, rather than  $\Theta(\eta_{i,k})$  as in Eq. 9. To compute the overall probability of saying old to item  $i$ , we compute,

$$P(\text{"old"}|i \text{ presented}) = \sum_{k \in \text{faces in memory}} P(\text{sample } k/i \text{ presented}) \Theta(\eta_{i,k}) + \sum_{k \in \text{prototypes}} P(\text{sample } k/i \text{ presented}) \Theta(pw \eta_{i,k}) \quad \text{Eq 15}$$

which simply extends Eq 9 to include the possibility of sampling a prototype, and if one is indeed sampled, the probability of saying old. Note that this addition of prototypes to the SimSample model is somewhat arbitrary, since it assumes that prototypes are only created

between two parents and not between any other pairs of faces. However, since we are only probing the locations between two parents with the morphs, this seems like a reasonable assumption.

The prototype strength ( $pw$ ) for morph face  $i$  was assumed to be a function of the distance between the two parent faces, under the assumption that blending is more likely to occur between two similar faces than between two dissimilar faces. Thus,

$$pw_i = (\eta_{i,p_1} + \eta_{i,p_2}) \rho \quad \text{Eq 16}$$

which gives the prototype model one additional free parameter,  $\rho$ , which determines the relation between distance and prototype strength. The results of this model fit are shown in Figure 5. Not only does this model provide a significant decrease in the RMSE, it places the similar morphs to the right of the similar parents, which previous models failed to do. Thus this model can account for the finding that similar morphs are chosen over their parents in forced-choice.

The prototype create method described in the previous section may seem arbitrary. Why should prototypes only be created between randomly chosen parent faces? To address this issue, we extended the previous prototype model to include a mechanism by which prototypes are created between all possible pairs of faces (as defined by the average in MDS space of the two faces), with two additional assumptions. First, the strength of the prototype in memory is proportional to the similarity of the two faces that are being used to create the prototype. This is consistent with Eq 16 above. Second, there was a threshold implemented such that the prototype creation occurred only for faces that were a minimum distance apart. This proved necessary because the full model that allowed all possible prototypes was computationally intractable. This model performed significantly better than the SimSample model, with a RMSE of 0.1066. In addition, this model accurately predicts that the similar morphs would be chosen slightly more often than the similar parents. This supports the proportional prototype assumption that underlies our prototype extension.

I would like to conclude this section on prototype extensions to the SimSample model with a few comments about prototype mechanisms. The existence of prototypes (at least as identifiable by testing prototype models) is the subject of furious debate within the categorization literature. Prototype models (or more properly, mixed modes that includes both prototypes and exemplars) are often mimicked by pure exemplar models, and from my study of the literature there is no firm evidence one way or the other (although others might disagree). The danger in concluding that prototypes exist just because a mixed model fits better than a pure exemplar model is that the prototypes may be making up for some deficit in the pure exemplar version of the model. There are other possible explanations for the tendency to choose the morphs over the parents in the forced-choice paradigm: for example, morphs may seem younger or more attractive due to the smoothing effects of the morphing process. Unless these effects are re-

<sup>1</sup>A version of the model in which the prototype weight influenced only the sampling process, not the decision process, was attempted, although the fit was markedly worse. In support of this assumption to have the strength of the item affect both sampling and decision processes, I suggest that subjects have an intuition about the strength of the match between the test face and the sampled face. This is reflected in other work, in which we have found a strong correlation between confidence and accuracy for target faces, which suggests that subjects can monitor the output of their sampling processes and use the strength of the output to make confidence judgments as well as old/new or forced-choice responses.

flected as dimensions in the MDS representation they will not be included in the modeling process. In addition, there may be context effects that occur when a morph is compared with its parents, which are described below. In general, I view the apparent need for a prototype extension of SimSample to indicate just how powerful the morph effect is, and how difficult it is for existing exemplar-based models to account for it. Thus at the very least the prototype extensions to SimSample quantify the range of the effects of the data and sketch out the types of data patterns that an exemplar will have to account for if it is to do it without a prototype extension. The improvement of the SimSample model over existing categorization models based on GCM should not be lost in the somewhat unrelated debate over the existence of prototypes.

### Context Dependencies in Face Spaces

The apparent need for prototype extensions in the fits of the SimSample model to the forced-choice recognition data suggests that, at the very least, the morphs are very similar to items stored in memory. The case of the SimSample model, this required a prototype extension, although other models may be developed that can account for these based on stored exemplars alone. An alternative solution suggests that there might be something special about the parent-morph relationship that makes the morphs confused with the parents more than one would predict on the basis of an exemplar-based model. One possibility is context-specific effects that make the morph appear more similar to one of its parents than one would otherwise expect on the basis of other considerations. Consider an example that illustrates this point. Suppose one were to morph an African-American face with a Caucasian face. The resulting morph would have a middle-gray skin tone in a black-and-white photo. When compared with the dark parent, however, the morph might look darker, and when compared with the light parent the morph might appear lighter. That is, when making similarity ratings the participant tends to ignore the differences across the two faces and look at only the similarities. Goldstone and Medin have demonstrated several effects in which feature representations of ambiguous items are borrowed from less ambiguous items (Medin, Goldstone & Gentner, 1993; Goldstone, Medin & Halberstadt, 1997). These sorts of assimilation effects may contribute to the memory data as well, since the participant may use similar matching or comparison procedures when accessing memory traces. The first step, then, is to identify context effects and then investigate whether they affect the memory data. To anticipate our findings, we find evidence for context effects with the morphs but find that they cannot account for the morph data from the forced-choice experiment.

When evaluating possible context effects, the first step is to determine an appropriate metric by which to compare parents to morphs. If context effects come into play, we would find that the morphs and parents are rated as more similar to each other than we would otherwise expect. Figure 6 illustrates an MDS space with lines drawn in to represent the size of the raw similarity ratings. In general the raw similarities correspond to the derived MDS distances, but the raw similarities be-

tween morphs and parents are much shorter than the morph-parent distances in MDS space. This is drawn to reflect the hypothetical context effects.

We can look for evidence for these effects in our data by comparing the raw similarity ratings against the derived distances computed from MDS space. In MDS space, the two parents and the morph have locations in MDS space as determined by the ordinal relations between each face and *all* other faces. Thus the location of each morph (as well as all other faces) is constrained by 99 numbers, only two of which represent parent/morph comparisons. In Figure 6, the MDS algorithm may try to adjust the location of the morph in order to account for the very close similarity reported between the morph and its two parents, perhaps by moving the morph upward. However, the position of the morph is also constrained by the ratings to the other 97 faces, and thus the MDS algorithm cannot account for the context effects that are present in the similarity ratings. Thus the MDS may be thought of as a representation that does not allow context effects, and therefore represents an appropriate measuring stick to compare the raw similarity ratings against. We expect context effects to primarily affect a morph/parent comparison, and thus we will single this out for examination.

In a stimulus set with 100 faces, there are  $(100 \cdot 99)/2 = 4950$  possible paired comparisons. Between each pair of faces we obtain a mean similarity rating from the data, and we can also compute the distance in MDS between each pair of faces. Figure 7 shows the scatterplot that compares the MDS distance against the rated similarity ratings. In general the correlation is quite high, and reflects the overall good fit of the MDS solution. When the stimuli are broken down into separate comparisons, however, we find that there are systematic and interesting deviations. The morph-parent pairs are singled out as large squares in the plots, and are systematically shifted below the rest of the data. This implies that when morphs are compared to their parents, they are systematically rated *more similar* than one would expect on the basis of the rest of the comparisons. This is exactly what one would predict if the assimilation hypothesis is correct, and demonstrates evidence for context-dependent effects in the morph-parent comparisons. The right panel of Figure 7 removes the parent and other faces, and just shows pairs that include at least one morph (as well as the pairs of parents used to construct the morphs). We find downward-shifted points only for pairs in which a morph is compared with one of its parents, not when that morph is compared to other parents or even other morphs. So these context dependencies do not result from the fact that morphs might be strange in general (due perhaps to artifacts in the morphing process), and appear only when the morph is compared with its parents.

### Context Effects in Recognition

The MDS solution was obtained using a set of data in which the morph-parent similarity ratings were set to missing data values, which allows the MDS to find a solution without taking these similarity values into consideration. The MDS fit did not significantly change, nor did the analysis in Figure 7 change significantly. This is what one would expect, given that for a

morph, the similarity rating to its parents are only two of 99 constraints that determine its location in MDS space, and so eliminating two ratings doesn't greatly affect the overall solution.

One question that is raised by evidence for context effects is whether these effects might also account for the tendency to choose the morphs over the similar parents in the forced-choice recognition data described in previous sections. If the morphs are drawn into the parents by a tendency to accentuate similarities, this might also make the morphs more likely to be chosen over the parents. These effects cannot be accounted for by the MDS, but can be implemented in a model in which the raw similarity ratings are used instead of the MDS distances as input to the SimSample model. The first stage of computation when fitting SimSample is to compute the similarities of each face to all other faces from the MDS coordinates, using Eqs 1 and 2. To fit a version in which we rely not on MDS distances but instead on the raw similarity ratings, we computed similarity between faces  $i$  and  $j$  by the following formula, which replaces Eqs 1 and 2,

$$\eta_{i,j} = e^{-as_{i,j}+b} \quad \text{Eq. 17}$$

where  $s_{i,j}$  is the raw similarity rating between faces  $i$  and  $j$ , and  $a$  and  $b$  are free parameters that linearly scale the raw similarity ratings. Note that while the original similarity ratings are on a scale of 1 (most similar) to 9 (least similar), these scores were converted to z-scores by subtracting the mean and dividing by the standard deviation of all scores for each participant. The  $b$  parameter is required because some scores are negative as a result of the z-score transformation.

Because we are using raw similarity ratings and not MDS coordinates, this model fit does not include attentional weights. Thus this fit of SimSample has 5 free parameters:  $a$  and  $b$  that map the raw similarity values into computed similarity and act like the generalization gradient parameter  $c$  in GCM, parameters that represent the response criterion and the standard deviation of the response criterion and  $\zeta$ , which controls the comparison behavior between the two faces.

The fit of this version of SimSample was not an improvement over the other versions that rely on the MDS space as input. The RMSE was 0.1171, and more importantly, the model could not account for the finding that morphs were chosen over their parents more than half of the time. Thus the context effects evident in the similarity rating data appear not to be able to account for the very high choosing rates of the morphs that occur despite the fact that the morphs were not studied.

### Face-Space Representations and the Other-Race Effect

The face-space representation used to model the present data and described in Busey (1998) allow the investigation of one of the major applications of face-space modeling. In the cross-race effect, observers who have limited contact with faces of other races are asked to identify faces of their own race and of another race.

In a remarkably consistent effect, memory for same-race faces was superior to memory for other-race faces (see Bothwell, Brigham & Malpass, 1989 for a review). African American and Caucasian subjects both demonstrate a bias for their own race in 79% of the experiments in the literature (Bothwell, et al, 1989). Interestingly, other-race faces tend to be classified faster than same-race faces in a race classification task. Valentine and Endo (1992) proposed an exemplar-based explanation of both of these effects in which other-race faces are in a separate part of face space and are more densely distributed. Figure 8 shows a hypothetical space. Identification is thought to be a function of the density surrounding an exemplar in face space, and exemplars in denser regions will become difficult to distinguish. This explains the fact that distinctive faces are more memorable in general (Valentine, 1991), and suggests that same-race faces, because they are more distributed will support better recognition. Classification, on the other hand, is accomplished by summing the similarity to all members in a category, which will provide a benefit for faces in a denser distribution. Thus the Figure 8 face space, in conjunction with separate decision rules for identification and classification, can account for the cross-racial effects.

Chiroro and Valentine (1995) addressed the contact hypothesis and found support for the exemplar-based account of the cross-racial identification effect. However, recently Levin (1996) argued that the face-space model was insufficient to account for classification and visual search tasks. He suggested that a dimension not included in the face-space representation, that of a quickly-coded race feature, was at work in the search and classification tasks, but not in the identification tasks. Thus, at the very least, face-space models must allow for the possibility of the use of different information for different tasks.

Surprisingly, the representation proposed in Figure 8 that has been used as an explanation for the cross-racial identification effects has never been measured (at least to our knowledge). There are a variety of ways to measure face-space, the simplest of which is to ask observers for similarity ratings between all possible pairs of faces and then submit the results to multi-dimensional scaling algorithms. The face-space initially described by Busey (1998) included twelve African American faces, and the 278 observers in that study were overwhelmingly white with somewhat limited exposure to African American faces. In addition, the scaling solution revealed that Race was the second dimension (after Age) that comes out of the scaling solution, thus demonstrating that race was a salient feature to the subjects.

The Age and Race dimensions are shown in Figure 9, along with the pictures of selected faces. As is apparent in Figure 9, African American faces are located in a separate part of the space and appear to be more densely clustered than Caucasian faces. To test this, we computed the average distance for each face to all other faces of the same race. Because this is an average, it is not influenced by the number of faces of that race. We find a strong effect of density, which is entirely consistent with the model proposed by Valentine and Endo

(1992). The average distance between Caucasian faces is 3.08 (SEM = 0.037) and between African-American faces is 2.14 (0.056). This difference is significant ( $p < 0.001$ ). This is not a computational artifact resulting from the fact that we had only 12 African American faces; when the computation is restricted to a randomly-chosen set of 12 Caucasian faces the mean is quite similar (3.12). This provides the first clear empirical evidence that the density-based explanation proposed to account for cross-racial identification and classification also appears in the similarity-based face space representation.

There are several interpretational issues with the previous dataset that should be discussed. First, because the study did not set out to investigate other-race issues, we did not attempt to recruit African American observers. This would have allowed us to demonstrate whether Caucasian faces are more densely distributed for these observers. In addition, the number of African American and Caucasian faces should have been equated, although for our present purposes we wished to keep the percentages about equal to that in the general population to avoid attentional effects. Equal numbers of Caucasian and African American faces would have reduced problems associated with the fact that African American faces appeared relatively rarely during the similarity ratings experiment, which may have made observers treat these faces differently. Of course, this is exactly what may happen in the real world when an observer encounters a relatively rarely-seen other-race face.

Despite these difficulties, the preceding analysis demonstrates that the density-based explanation proposed by Valentine and Endo (1992) is essentially correct, and validates the face-space approach. This opens the way for quantitative models based on empirical face space representations, in which identification and classification performance can be predicted for each face in the set.

### Applications of Face-Space Modeling

The successes of the current quantitative modeling provide direction for future work. Much of the face-recognition literature revolves around three central themes: how are faces represented in memory (i.e. as exemplars or relative to a central prototype), what mechanisms determine how faces are retrieved from memory (familiarity and memorability (Vokey & Read 1992) or just memorability (Valentine, 1991a,b)), and how the structure of the face-space can affect the storage and recognition process. This third theme has been discussed in a number of domains, including applications to the cross-racial identification data, in which members of another race are represented in a separate cluster in MDS space, where the individual exemplars are grouped together more tightly, making individual identifications more difficult (Chiroro & Valentine, 1995).

The models applied to the forced-choice recognition data in this chapter allow a number of conclusions about these themes. First, we find much more support for an exemplar-based representation than a central-prototype representation, although in some instances it appears that prototypes are necessary to account for

very typical faces. It is important to point out that an exemplar-based model that had a different formulation of typicality or a different sampling mechanism might account for the similar morphs without assuming prototypes. The current modeling merely demonstrates the failures of existing exemplar-based models. The fact that we find a poor fit from the global prototype model suggests that the norm-based coding model is not a reasonable model to account for recognition data, although it may be useful for other types of comparisons where a rating is made on a face relative to some standard (i.e. he is attractive for his age). In general, discriminating between exemplar-based and prototype-based models is difficult without a representation of the scaling space as input to a quantitative model.

The SimSample model demonstrates how a sampling process in conjunction with a similarity-based decision mechanism could incorporate mechanisms that account for both a familiarity-based recognition mechanism and a recall-based mechanism. The sampling process that depends upon the similarity structure of the faces in memory tends to favor distinctive items, which corresponds to the memorability component of Vokey and Read (1992). The tendency of typical items to lie near one another and be mis-sampled during the sampling process will increase both the hit and false alarm rate for typical items, which previously was associated with a separate familiarity-based mechanism. Evidence in favor of the SimSample model comes from the previous work done on the various components of the model. There is strong evidence in favor of the exponential similarity gradient, as well as a great deal of work on the Luce Choice Rule. The sampling process comes directly from the SAM model (Gillund and Shiffrin, 1984). Thus the assumptions that underlie SimSample have a history of successfully accounting for recognition and recall data with words.

While the SimSample model could account for both the effects of distinctiveness and typicality, it failed to account for the effects of very typical faces, as demonstrated by its inability to account for the behavior of the similar morphs. Fixing this problem might require one of several possibilities. One is the prototype fix described above. The second possibility is an alternative mechanism that provides a better account of the most typical faces, perhaps by adopting a different sampling or response mechanism. Third, there might be a second familiarity-based mechanism at work, in addition to the current model. This possibility was investigated by adding Nosofsky's Generalized Context Model (GCM, Nosofsky, 1986) to the SimSample process, which effectively increased the probability of choosing very typical faces. The addition of this explicit familiarity component to the SimSample model did not improve the fit, nor did it place the similar morphs above the similar parents.

A final possibility is that there is something strange about the morphs that tend to make them seem familiar. This quantity would have to be outside the domain of the dimensions recovered by the similarity ratings experiment. For example, morphs appear smoother and younger than their parent faces, that this may have attracted responses in the forced-choice task.

This raises the larger issue of the limits of the geometric model. Levin (1996) points out that in cross-racial classification, the dimensional information used is different from that used to make cross-racial identifications. One might resort to shifting attention along different dimensions, as the modeling in this chapter has adopted, but this will work only if the recovered dimensions from the MDS solution correspond to the dimensions that are used in recognition or classification. Alternatively, a subspace model might be adopted, that defines in advance which dimensions are relevant for a particular task.

The geometric space representing similarity relations between faces can be obtained by measures other than similarity ratings. For example, the reaction time to call two faces 'different' in a same/different may be used as input to the MDS algorithm, under the assumption that similar faces require more time to note differences. Another approach is to construct an input space on the basis of physical features, as described by

Steyvers and Busey (this volume). These input spaces may highlight different aspects of faces, such that a similarity rating task may highlight similarities (and perhaps rely on lower spatial frequencies) while a reaction time task may highlight differences (see Uttal, this volume).

The limitations of geometric spaces should not be seen as disconfirmation of what I believe is a very promising approach. The MDS approach allows predictions for individual stimuli, which in turn provides evidence for the role of the similarity structure of faces. This structure can then be used to ask questions about the retrieval mechanisms that enable recognition. There are important links that can be made between a large literature involving geometric models and an equally large literature involving global memory models. Faces seem to be an elegant and important stimulus that can be used to bridge both literatures.

## References

- Bartlett, J., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, *12*, 219-228.
- Bothwell, R., Brigham, J., & Malpass, R. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin*, *15*, 19-25.
- Brown, J., Lewis, V.J., & Monk, A.F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology*, *29*, 461-473.
- Busey, T. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science*, *9*, 476-483.
- Busey, T. A & Tunnicliff, J.(submitted). Accounts of blending, typicality and distinctiveness in face recognition. Submitted to *Journal of Experimental Psychology: Learning Memory and Cognition*.
- Byatt, G. & Rhodes, G. (1998). Recognition of own-race and other-race caricatures: Implications for models of face recognition. *Vision Research*, *38*, 2455-2468.
- 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*, *48A*, 879-894.
- Edelman, S. & Intrator, N. (submitted). Learning as extraction of low-dimensional representations.
- Gentner, D., & Collins, A. (1981). Studies of inference from lack of knowledge. *Memory & Cognition*, *9*, 434-443.
- Gillund, G. & Shiffrin, R. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *92*, 1-38.
- Goldstone, R., Medin, D. & Halberstadt, J. (1997). Similarity in context. *Memory and Cognition*, *25*, 237-255.
- Hintzman, D. (1986). "Schema Abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411-428.
- Homa, D., Goldhardt, B., Burrue-Homa, L., & Smith, J.C. (1993). Influence of manipulated category knowledge on prototype classification and recognition. *Memory & Cognition*, *21*(4), 529-538.
- Jacoby, J. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513-541.
- Jacoby, L.L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *3*, 306-340.
- Johnston, R., Milne, A., 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*, 59-67.
- Kayser, A. (1985). *Heads*. New York: Abbeville Press.
- Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, *85*, 445-463.
- Levin, D. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 1364-1382.
- Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning & Memory*, *5*, 212-228.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*, 252-271.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.
- Metcalf, J. (1990). Composite Holographic Associative Recall Model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General*, *119*, 145-160.
- Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254-278.
- Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.
- Nosofsky, R.M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87-108.
- Nosofsky, R.M. & Palmeri, T. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.
- O'Toole, A.J., Bartlett, J.C., & Abdi, H. (in press). A signal detection model applied to the stimulus: Understanding covariances in face recognition experiments in the context of face sampling distributions. In press, *Visual Cognition*.
- O'Toole, A.J., Deffenbacher, K.A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, *22*, 208-224.
- O'Toole, A., Wenger, M., & Townsend, J. (1999). Quantitative models of perceiving and remembering faces: Precedents and possibilities. *This volume*.
- 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*, 2307-2321.
- Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*, 373-421.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

- Shiffrin, R., Huber, D., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 267-287.
- Shiffrin, R. & Nobel, P. (1997). The art of model development and testing. *Behavior Research Methods, Instruments, and Computers*, *29*, 6-14.
- Steyvers, M. & Busey, T. (1999). Predicting similarity ratings to faces with physical descriptions. This volume.
- Solso, R. L., & McCarthy, J. E. (1981). Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology*, *72*, 499-503.
- Uttal, W., Baruch, T., & Allen, L. (1995a). The effect of combinations of image degradations in a discrimination task. *Perception & Psychophysics*, *57*, 668-681.
- Uttal, W., Baruch, T., & Allen, L. (1995b). Combining image degradations in a recognition task. *Perception & Psychophysics*, *57*, 682-691.
- Uttal, W. (1999). This volume.
- Valentin, D., Abdi, H., Edelman, B., & Posamentier, M. (1999). 2D or not 2D? That is the question: What can we learn from computational models operating on 2D representations of faces? This volume.
- Valentine, T. (1999). Face-space models of face recognition. This volume.
- Valentine, T. (1991a). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43A*, 161-204.
- Valentine, T. (1991b). Representation and process in face recognition. In Watt, R. (Ed.), *Vision and visual dysfunction*. Vol. 14: Pattern recognition in man and machine series editor, J. Cronley-Dillan). London: Macmillan.
- Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognizing and classifying faces. *Perception*, *15*, 525-535.
- Valentine, T. & Endo, M. (1992). Towards an explanatory model of face processing: the effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology*, *44A*, 671-703.
- Valentine, T. & Ferrara, A. (1991). Typicality in categorization, recognition and identification: Evidence from face recognition. *British Journal of Psychology*, *82*, 87-102.
- Vokey, J. & Read, J (1992). Familiarity, memorability and the effect of typicality on the recognition of faces. *Memory & Cognition*, *20* 291-302.
- Wells, G.L., & Lindsay, R.C.L. (1985). Methodological notes on the accuracy-confidence relation in eyewitness identifications. *Journal of Applied Psychology*, *70*, 413-419.
- Wenger, M., & Townsend, J. (submitted). Spatial frequencies in short-term memory for faces: A test of three frequency-dependent hypotheses. *Submitted to Memory and Cognition*.
- Wixted, J. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 681-690.
- Yonelinas, A., Dobbins, I., Szymanski, M., Dhaliwal, H., & King, L. (1996). Signal-detection, threshold and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition*, *5*, 418-441.

## Tables

	# p	c	W1	W2	W3	W4	W5	W6	$\theta$	$\beta$	$\zeta$	RMSE	F-Value	Crit. F	
GCMID	9	2.301	0.13	0.07	0.02	0.45	0.21	0.12	48.88	10000	1.603	0.120	---		
SimSample	# p	c	crit	W1	W2	W3	W4	W5	W6	C.Var	$\zeta$				
	9	3.478	0.01	0.08	0.10	0.13	0.59	0.08	0.07	0.192	6.900	0.112	---		
SimSample-Gaussian Noise	# p	c	crit	W1	W2	W3	W4	W5	W6	C.Var	$\zeta$				
	9	2.061	0.59	0.04	0.03	0.11	0.19	0.04	0.02	0.206	6.333	0.108	---		
SimSample-Clustering	# p	c	crit	W1	W2	W3	W4	W5	W6	C.Var	$\zeta$				
	11	4.000	0.12	0.08	0.08	0.35	0.33	0.04	0.05	0.000	2.122	0.107	5.06 *	3.10	
SimSample-Global Prototype	# p	c	crit	W1	W2	W3	W4	W5	W6	C.Var	$\zeta$				
	17	4.037	0.26	0.14	0.05	0.15	0.33	0.07	0.00	0.000	2.170	0.111	1.21 (NS)	2.12	
		gpd1	gpd2	gpd3	gpd4	gpd5	gpd6	GPWeight							
		-0.46	-0.93	0.92	0.27	-0.48	-1.26	0.32							
SimSample- Proportional Prototypes	# p	c	crit	W1	W2	W3	W4	W5	W6	C.Var	$\zeta$				
	10	4.346	0.05	0.15	0.17	0.00	0.38	0.25	0.65	0.548	4.358	0.097	31.10 *	3.95	
SimSample- Raw Similarity Ratings	# p	a	b	crit							C.Var	$\zeta$			
	5	4.0	31.1	.0005							0.040	3.18	0.1171		

Table 1. Parameter values for all fits. The obtained F-values compare the model with the original SimSample model. # p represents the number of parameters, RMSE is the root-mean-squared error that has been corrected by subtracting the number of parameters (p) from the number of datapoints (n) in the denominator:

$$\sqrt{n - p}$$

Condition	Forced-Choice Data	SimSample Model Fit	SimSample + Gaussian Similarities	SimSample + Clustering	SimSample + Global Prototype	SimSample + Proportional Prototypes Fit
Targets	0.754	0.750	0.751	0.744	0.749	0.767
Distracters	0.251	0.273	0.273	0.276	0.274	0.264
Similar Morphs	<b>0.539</b>	<b>0.457</b>	<b>0.489</b>	<b>0.472</b>	<b>0.469</b>	<b>0.533</b>
Similar Parents	<b>0.470</b>	<b>0.542</b>	<b>0.508</b>	<b>0.529</b>	<b>0.529</b>	<b>0.468</b>
Dissimilar Morphs	0.345	0.361	0.364	0.342	0.349	0.353
Dissimilar Parents	0.655	0.641	0.637	0.659	0.651	0.650

Table 2. Mean probability of choosing data for the 6 experimental conditions, along with the fits for the various models. Only the Proportional Prototypes model can account for the reversal between Similar Morphs and Parents (Bold numbers).

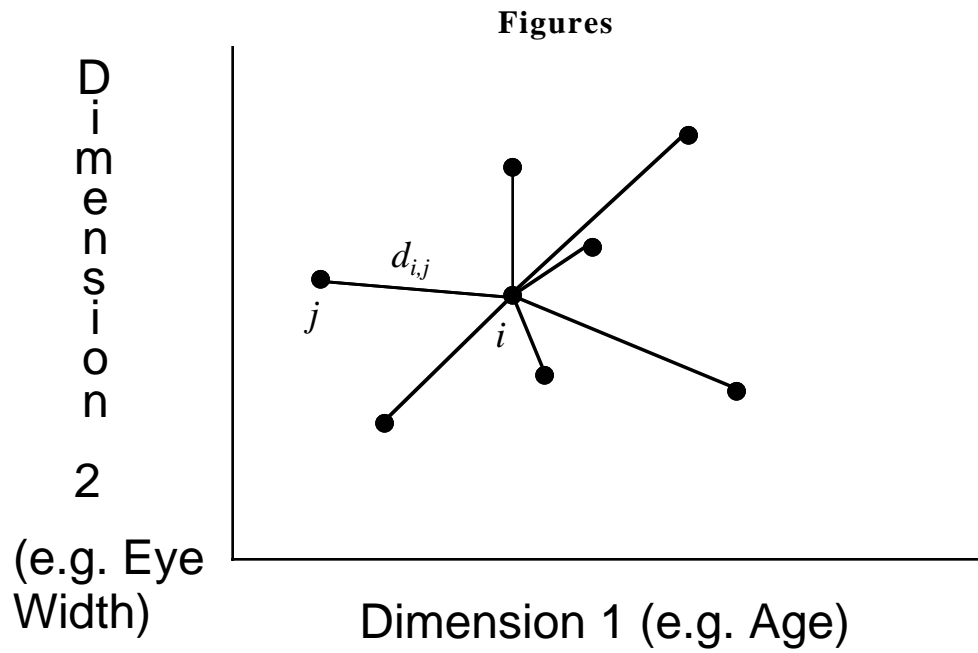


Figure 1. Hypothetical 'Face-Space' derived from MDS procedures applied to similarity ratings on all pairs of faces. Each face is represented as a point in this space, and values such as distance can be computed directly from this representation.

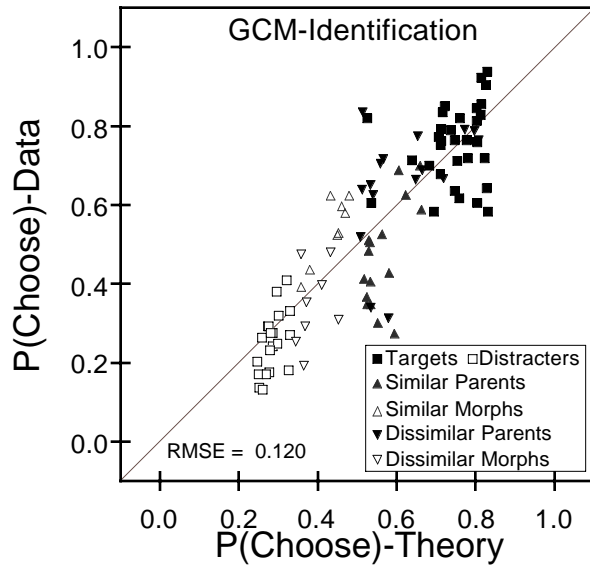


Figure 2. Fit of GCM-Identification model.

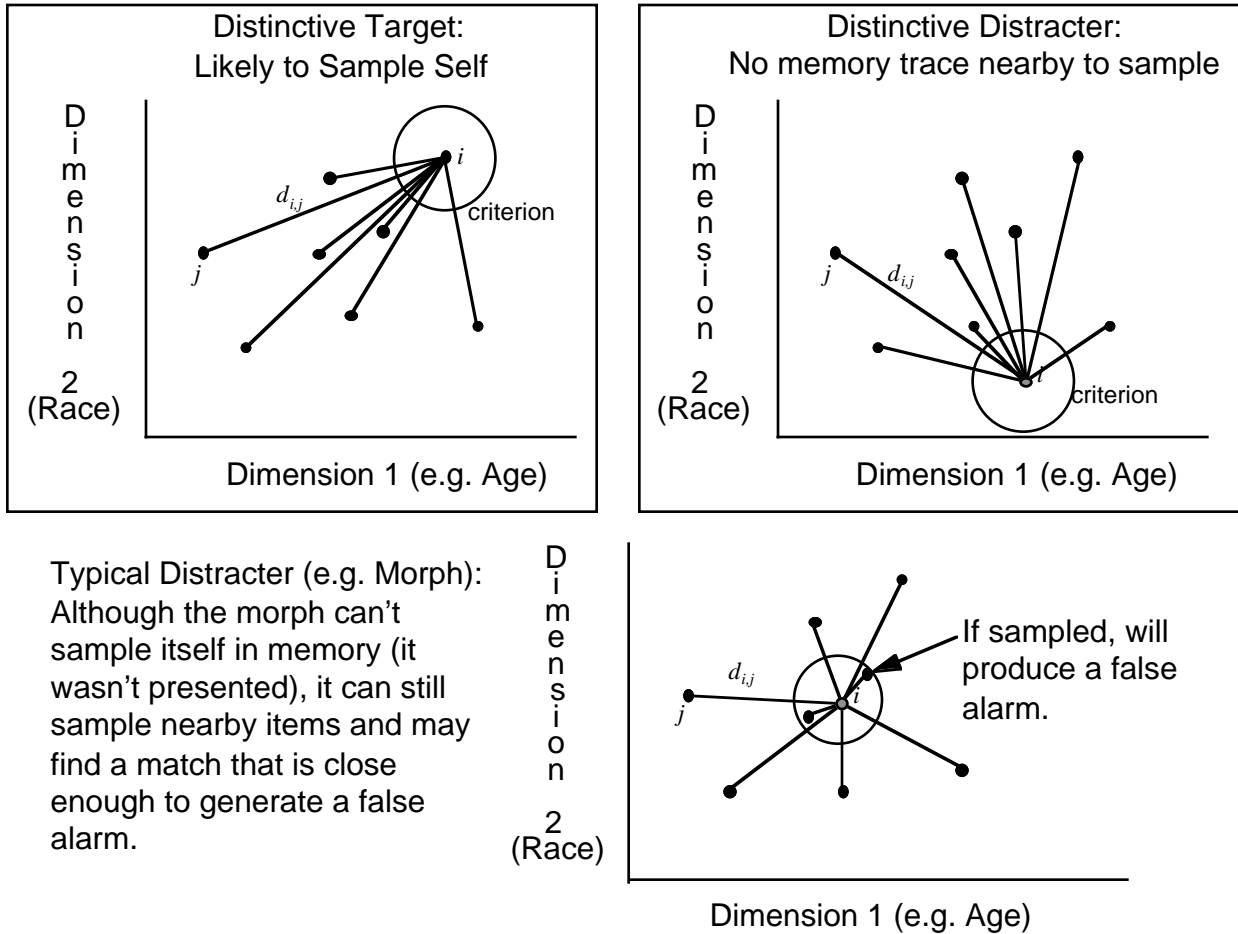


Figure 3. Predictions of the SimSample model to Distinctive Targets, Distinctive Distracters, and Typical Distracters (morphs). Upper Left: A distinctive target is very likely to sample itself and thus has a high hit rate. Upper Right: A distinctive distracter cannot sample itself and may not have any nearby faces that could produce a false alarm if sampled. Bottom Panel: A very typical distracter may produce a false alarm if a nearby item is sampled by mistake and is within the criterion for responding old. Typical target faces will have a high hit rate if either the face samples itself or samples a nearby target that lies inside the criterion. For forced-choice data, the more likely an observer is to say old to a face, the more likely they will choose it in a forced-choice task, all other factors being equal.

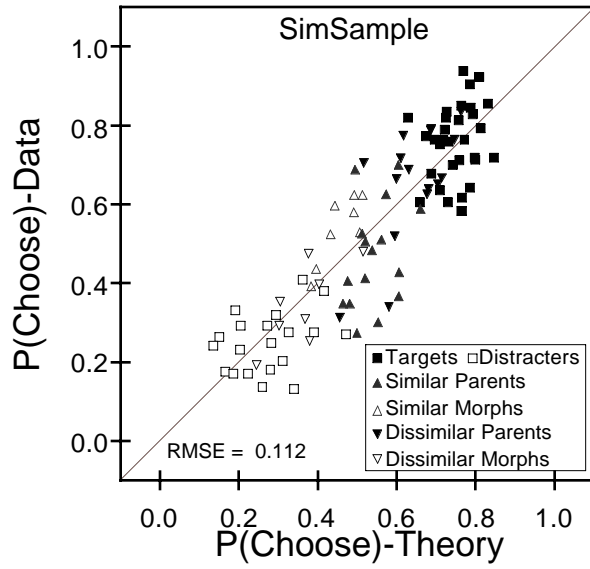


Figure 4. Fit of SimSample model.

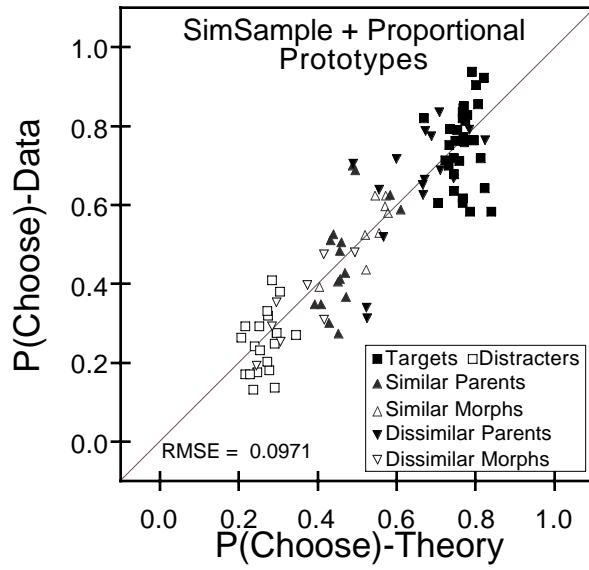


Figure 5. Fit of Proportional Prototypes version of the SimSample model. This model places the Similar Morphs to the right of the Similar Parents, which previous models could not do.

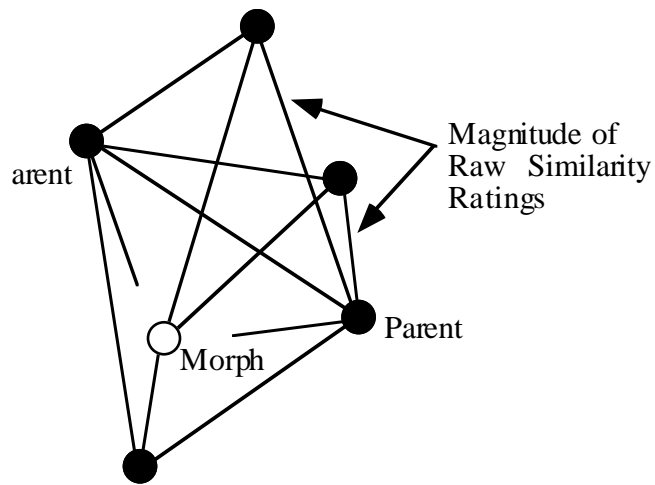


Figure 6. Hypothetical MDS space in which the magnitude of the similarity ratings are shown as black lines. Context effects are represented as lines that are deliberately shorter than the distance between the morph and its parents. The MDS program might try to move the morph upward to account for these raw similarity ratings that are poorly fit, but the other constraints provided by the similarity ratings to the other 97 faces (3 are shown) prevents the morph from shifting.

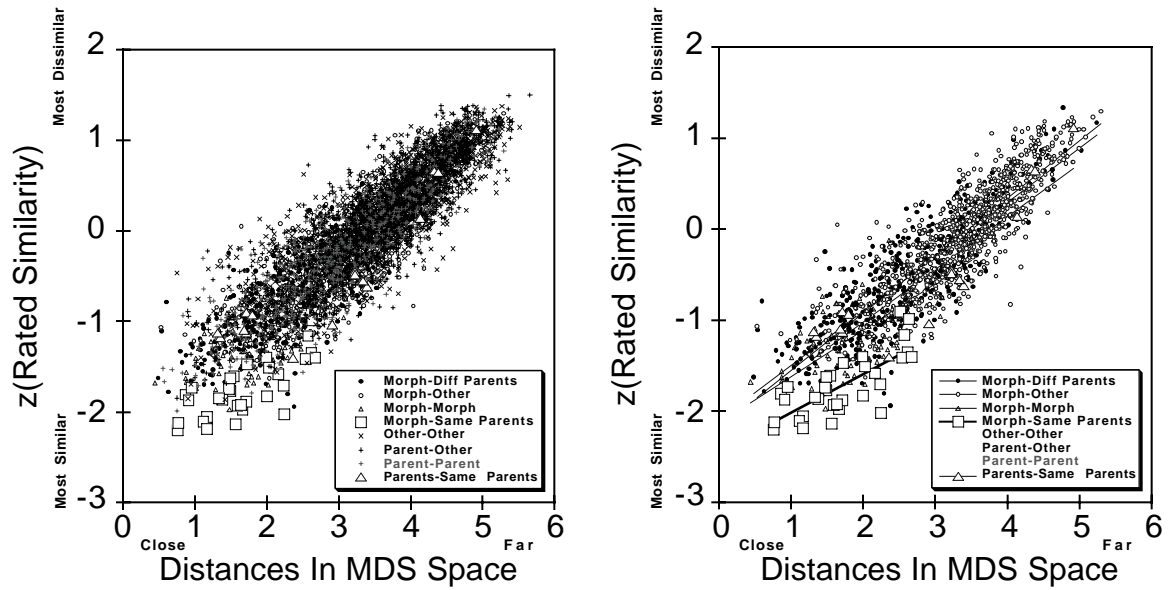


Figure 7. Raw similarity values (converted to z-scores) compared with the computed MDS distance for all pairs of faces. In general the fit is quite good, but the morph/parent pairs (large squares) are systematically shifted below the rest of the points. This is consistent with context-specific effects (see text).

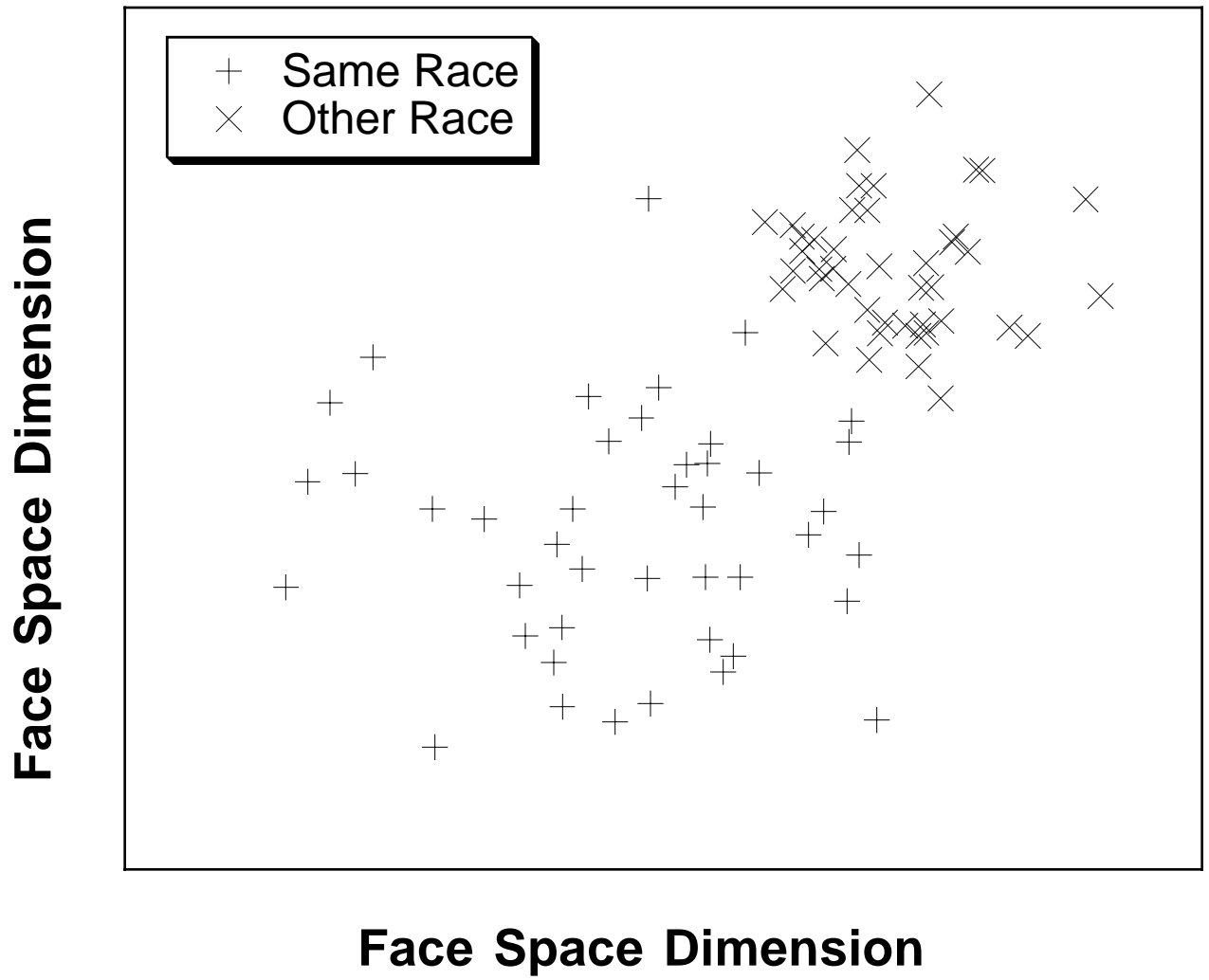


Figure 8. Hypothetical face-space proposed by Valentine and Endo (1992) to account for other-race effects.

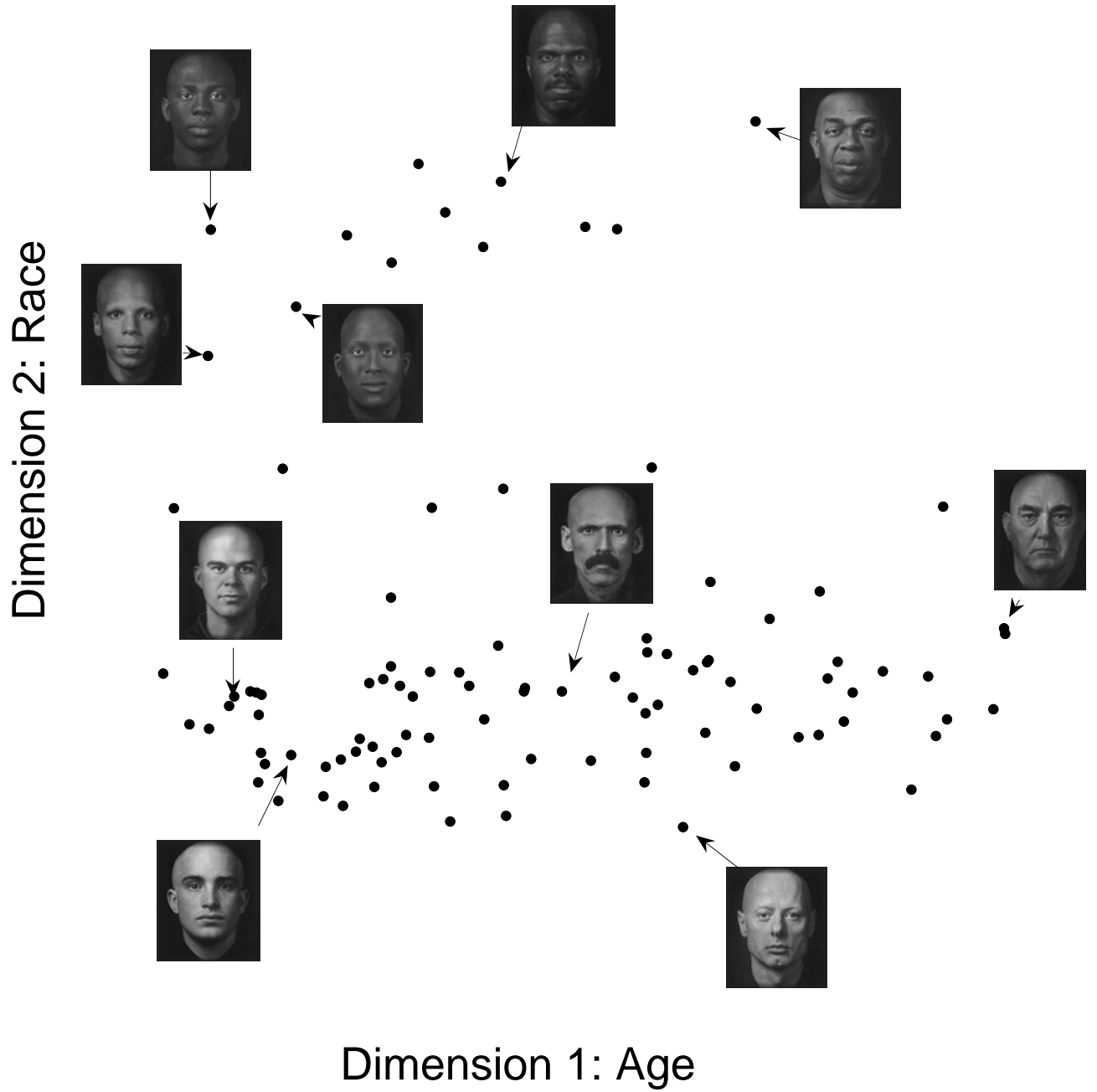


Figure 9. Empirical face space obtained from a scaling solution derived from similarity ratings.