IN PRESS: JOURNAL OF EXPERIMENTAL PSYCHOLOGY: LEARNING, MEMORY & COGNITION

Accounts of blending, distinctiveness and typicality in the false recognition of faces

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

JENNIFER TUNNICLIFF
University of California at Riverside

The false recognition of distractor faces created from the combination of studied faces has been attributed to mechanisms that create novel traces in memory. However, an alternative account is that the distractors are similar to studied items and therefore seem familiar. In three experiments we examined the false recognition of faces within the context of a larger model of face recognition that accounts for the effects of typicality and distinctiveness. Using morphing image processing techniques, we created a set of distractor faces that were mathematical blends of two 'parent' faces. Subjects studied the parent faces, and the distractors were used to assess the presence of novel memory traces that may have resulted from a psychological binding or blending mechanism that inappropriately combines studied faces. To address the temporal properties of such a mechanism, during study the two parent faces were seen either sequentially or separated by at least 20 other faces. We found very high false alarm rates to the morph distractors, but no effects of the temporal separation manipulation. In a forced choice version, subjects were more likely to choose the distractor over the parent face if the two parents are similar to each other, which demonstrates the strength of the false recognition effect and is consistent with a blending mechanism. Recognition models based on Nosofsky's Generalized Context Model (GCM, Nosofsky, 1986) could account for some but not all aspects of the data. A new model, called SimSample, can account for the effects of typicality and distinctiveness, but still has difficulty accounting for the high false alarm rates to the morphs. A version that includes explicit prototype representations that are created only between similar faces provides a significantly better fit to the morph data. The conclusions are consistent with an account of memory in which novel traces are created in memory as a result of the combination of existing traces; alternative explanations are also explored.

This research concerns the face recognition errors that result when a novel face that has been constructed from studied faces is used as a distractor in old/new recognition experiments. This novel distractor often produces large number of false alarms, and subjects will be extremely confident that this indeed was an old face (e.g. Solso and McCarthy, 1981; Reinitz, Lammers & Cochran, 1992; Reinitz, Morrissey & Demb, 1994;

* Please send correspondence to Thomas A. Busey,Department of Psychology,Indiana University,Bloomington, IN 47405, email: busey@indiana.edu

Kroll, Knight, Metcalfe, Wolf & Tulving, 1996). One interpretation that has been proposed for these effects is that psychological mechanisms somehow combine the features of studied faces into a new representation in memory that matches the distractor, resulting in a false alarm. Empirical and theoretical interest in these types of memory distortions has grown recently, in part because such effects are found under different conditions and with different types of stimulus materials. One of the most active research areas looking at memory illusions uses a technique borrowed from illusory conjunctions in the visual attention literature (e.g. Treisman and Schmidt, 1982). In the applications to memory work, a *conjunction* distractor is formed by combining two pieces from previously-studied faces or words. For comparison, a *feature* distractor is constructed by combining one piece from a studied item and the other piece from a non-studied item. *Novel* distractors consist only of parts from non-studied items. In a typical recognition memory task, subjects are asked to study a list of items, and then perform an old/new recognition test. In most experiments, the probability of saying ’old’ to each stimulus type is given by this ordering: target items > conjunction distractors > feature distractors > novel distractors. The finding that conjunction errors are higher than feature errors is often taken as indication of a specific failure of a binding or conjunction mechanism which has separate properties from those that encode the individual items. For example, Reinitz et al. (1994) conclude "In the absence of encoded relational information, ’memory conjunction errors’ may occur; that is, features from separate stimulus events may be erroneously conjoined to produce a ’memory’ that does not correspond to a previously experienced stimulus." (p. 167). A similar conclusion was reached by Kroll et al. (1996), who state that: "[a] binding process exists, and that its function consists in ’gluing’ together the elements of the incoming information into separately retrievable engrams in the long-term storage....[An] inhibitory component of the binding operation allows the formation of only those conjunctions of elements as long-term engrams that correspond to the temporally organized chunks in primary memory. When the inhibitory component fails, elements corresponding to higher-order units in long-term memory are created heedlessly" (p. 193). Thus, conjunction errors have been explained by a specific mechanism in which parts of studied items are recombined to form novel traces in memory, perhaps as the result of the failure of an inhibitory mechanism that would normally prevent such spurious binding. The resulting novel trace provides a strong match to a conjunction stimulus at test and produces a false alarm.

Converging evidence in support of the creation of novel traces in memory comes from the categorization and recognition memory literature. In these studies, a physical prototype stimulus is constructed by combining studied items, usually through an averaging process. In recognition experiments the prototype often produces a high false alarm rate, even though the prototype is novel (e.g. Franks & Bransford, 1971; Neumann, 1977; Solso & McCarthy, 1981). The magnitude of prototype effects depends upon the similarity of the exemplars to the prototype: the more similar the exemplars are to the prototype, the more the prototype is falsely recognized (e.g. Homa, Goldhardt, Burrel-Homa & Smith, 1993). Metcalfe (1990) proposed a specific model for such prototypes in which studied items might blend together to produce a new trace in memory, and she demonstrated that prototypes were more likely to form between similar exemplars. Knapp and Anderson (1984) also propose a model in which exemplars, if sufficiently similar, combine to produce a prototype in memory. The prototype mechanism is often not specified, although Kroll et al. (1996) suggest that the false recognition of the prototype is the result of binding failure, and therefore they explicitly link prototype effects with the conjunction errors seen in memory illusion work.

The most compelling evidence for the existence of a specific binding mechanism that creates novel traces in memory comes from experiments designed to produce a specific deficit in its functioning. Two examples that use faces in a memory illusion paradigm illustrate the characteristic results, and are shown in Table 1. The stimuli in these experiments consist of Identikit line drawings of faces, in which individual features from two studied faces are recombined to construct a conjunction distractor face. Reinitz et al. (1994) placed subjects in a divided attention condition in which subjects had to perform a secondary dot-counting task while studying the faces. Other subjects were allowed to devote full attention to faces. The authors hoped that the secondary task would specifically disrupt the binding mechanisms and produce more conjunction errors than those produced by subjects in the full attention condition. The full attention condition does indeed produce fewer conjunction errors. In a related task using the identical stimuli, Kroll et al. (1996) tested subjects from both older adult (control) and patient populations. The patients had right hemisphere damage that the authors suggested would lead to a specific failure in the binding of relational information. The failure to inhibit a runaway binding system would produce more novel traces in memory and therefore more conjunction errors. The data support such a claim: the patients show evidence of more conjunction errors than the controls. Based on these results, the authors argue for the existence of a specific mechanism, perhaps supported by attention or the hippocampus, that underlies the binding of relational information. In the absence of such a mechanism, features are mis-combined to create novel traces in memory, and conjunction errors result.

There are two difficulties with this argument. First, in virtually every example in the literature, subjects produce more false alarms to feature distractors than to novel distractors. If a binding mechanism is the sole mechanism for producing errors (beyond random guessing, which would account for the non-zero novel false alarm rate), we would expect no difference between these two stimuli, because the binding system cannot bind together a feature in memory with a feature not in memory. Second, in the literature, every manipulation that produces an increase in conjunction errors also produces an increase in feature errors. There seems to be no manipulation that selectively affects the conjunction error rate without also altering the feature error rate as well. This happens even in the Reinitz et al. (1994) forced-choice data, so the increase in feature error rates cannot be attributed to criterion shifts. These difficulties raise some doubt about the special status given to the interpretation of conjunction errors.

An equally plausible alternative explanation to the binding failure hypothesis is that the overall familiarity of the distractor may drive the subject to say old. The conjunction stimulus, by virtue of its construction, necessarily bears some similarity to two studied items. A feature stimulus bears a resemblance to only one studied item. As we will describe in detail below, any model that computes a familiarity measure based on the summed similarity between items will predict a higher false alarm rate to conjunction relative to feature distractors. While in early work, Reinitz et al. (1992) discount the role of familiarity, in later work they admit that familiarity may play a role in the recognition processes (Reinitz et al., 1994). What they fail to consider is the possibility that familiarity may underlie *all* of the recognition processes, including the conjunction errors. One may not have to posit an additional mechanisms that create novel traces in memory that cause conjunction errors. Below we describe how a well-established and successful model of categorization and recognition will make predictions that are consistent with the memory conjunction data. This model can, in principle, provide a complete account of the data without assuming any mechanism other than storage of individual exemplars. If this model can account for the memory conjunction data, then parsimony would preclude us from assuming an additional binding or blending mechanism. The importance of testing models has been demonstrated in related domains: in the visual search illusory conjunction literature from which the memory illusion paradigm was adapted, solid evidence for illusory conjunctions has required rigorous hierarchical testing of formal models (e.g. Ashby, Prinzmetal, Ivry, and Maddox, 1996).

The preceding discussion should not be taken as a refutation of a binding mechanism; we wish only to make the point that there are plausible alternative explanations that have not been considered within the context of the memory illusion literature, but that we consider in the present work. In the following section, we describe how a familiarity-based model that relies on a summed-similarity mechanism might account for the memory illusion data. We will show how this model can qualitatively account for the conjunction stimuli in the memory illusion data, and then apply similar models to new face recognition data. In a later section we develop a new model that remedies many of the shortcomings of the familiarity-based model when applied to face recognition. The fundamental logic of our approach is as follows: the hypothesis that novel traces are created in memory as a result of blending or misbinding mechanisms can only receive support after alternative explanations based on exemplar-only models are rejected. Therefore we will test existing and newly-developed exemplar models, and only if these fail to account for the blending effects can we conclude that a blending or binding mechanism is a plausible explanation for the data.

Accounting for Memory Illusions with a Familiarity Model

Nosofsky's Generalized Context Model (GCM, Nosofsky 1986) is perhaps the model best suited to make predictions for the memory conjunction data. GCM assumes that a test item engenders a feeling of familiarity, and items that result in more familiarity are more likely to be labeled as old faces by subjects. Familiarity is a function of the summed similarity between the test face and all other faces in memory, as described below.

To derive specific predictions for the probability of calling each face 'old' in the test session of a recognition experiment, the similarity relations between all pairs of faces must be represented. Traditionally this is accomplished by asking for similarity ratings between pairs of faces and then submitting the results to multi-dimensional scaling algorithms. These algorithms produce a psychological space in which individual stimuli are represented as points in a multidimensional space. The dimensions of this space represent stimulus dimensions along which faces vary, such as age or race, and various metrics such as the distance between two points can be computed. Within GCM, the distance between any two faces i and j, $d_{i,j}$, in a space with M dimensions is computed as,

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

where $x_{i,n}$ is the coordinate for face i on dimension n and w_n is an attentional weight given to dimension n[1]. The similarity, $\eta_{i,j}$,

[1] 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 Minkoski distance metric, as: $d_{i,j} = \left(\sum_{n=1}^{6} w_n \left| x_{i,n} - x_{j,n} \right|^b \right)^{1/b}$. In the city-block metric, b is set to 1.0. In one version of the model

between faces i and j is defined as

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

where *c* is a scaling parameter used to define the relation between distance and similarity (Shepard, 1974). A large *c* implies that only very close faces are deemed similar, while a small *c* value implies that subjects consider most items to be similar to each other. For MDS spaces scaled between -2 and 2, typical values of *c* range from 2-4.

A prediction of familiarity for test item k is simply the summed similarity between item k and all other faces in memory,

$$familiarity_k = \sum_{\substack{j \subset studied \\ faces}} \eta_{k,j} \qquad \text{Eq. 3}$$

which can be converted to a prediction of the probability of saying old using some monotonic transform F of familiarity,

$$P("old" \mid k \quad presented) = F[\ familiarity_k\ ] \qquad \text{Eq. 4}$$

where F in this case is a logistic function,

$$P("old" \mid k \quad presented) = \frac{1}{1+\beta e^{-\theta \quad familiarity_k}} \qquad \text{Eq. 5}$$

with two free parameters β and θ.

GCM Simulations

One of the difficulties with applying GCM to the data in Table 1 is that the similarity relations between the studied faces were not measured. It is still possible to generate predictions by making assumptions about the distributions of the similarity relations among the features, and in this section we describe how this was done. In the original experiments, each face was constructed from 4 separate features (hair, mouth, eyes and nose). A conjunction stimulus was created by taking the hair and mouth from one face and combining these with the eyes and nose of a second face. A similar process was used to create the feature stimuli, except one set of features comes from an unstudied face. We simulated this procedure by assuming that given instances of a particular feature (e.g. eyes) would fall along a normally distributed unidimensional scale that would reflect some measure of similarity between the actual features. Thus to construct a representation of a studied face, we chose values for each of the 4 features by sampling from a normal distribution with a mean of zero and unit variance. This was repeated for two additional study faces and one face that would be designated as an unstudied face. To create the conjunction stimulus, the values representing two features were copied from the first studied face, while the second study face supplied the values corresponding to the second set of two features. In a similar manner, the feature values from the third study face and the unstudied face were combined to produce the feature stimulus.

The values corresponding to each feature can be thought of as the values along a particular dimension. Thus each face can be represented in a 4-dimensional space, and distances between faces can be computed via Eq 1. Similarity follows directly via Eq 2, and the familiarity values are computed by summing the similarity between each test stimulus and the three studied faces. The similarity between the unstudied face and the three studied faces was used to produce predictions for the novel stimulus. The familiarity values were then transformed via Eq 5 to produce the predicted probability of saying old to each stimulus type, as shown in Table 1 for the Kroll et al (1996) data[2]. In order to eliminate possible problems that might crop up with a particular choice of feature values, the procedure described above for construct-

fitting we allows b to freely vary, and the estimated value was quite close to 2.0.

[2] The data from Reinitz et al (1994) involve a forced-choice procedure that requires more parameters and model complexity to fit, which doesn't provide strong constraints on a model like GCM.

ing representations of the various stimulus types was repeated 5000 times, which in essence replicates the above simulation to produce 5000 datasets and attempts to fit all the results with a single set of model parameters. The Solver function in Microsoft Excel was used to minimize the least square differences between the predicted and actual data, using independent parameter estimates for control and deficit subjects.

The predictions from GCM are shown in Table 1, and demonstrate two effects. First, GCM correctly predicts higher feature errors than novel errors, which is a result that a binding mechanism explanation cannot account for. It does so because a feature stimulus is similar to one face in memory, while a novel stimulus is not. Second, GCM correctly predicts that as conjunction errors increase in the deficit condition, feature errors will increase as well. This is another effect that the binding mechanism account cannot explain. GCM achieves the increase in both conjunction and feature errors by assuming that the sensitivity parameter c gets smaller for subjects in the 'deficit' condition, which implies that these subjects are having difficulty distinguishing between the faces. That is, to subjects with brain damage, all faces look more similar, which causes an increase in the rate at which both feature and conjunction stimuli look like target faces. This results in an increase in the false alarm rates for both types of stimuli, although because conjunction stimuli share features with two target items and are therefore similar to more faces in memory, conjunction stimuli will continue to have higher false alarm rates than feature stimuli. Thus GCM correctly predicts that both feature and conjunction errors will increase in the 'deficit' condition, and conjunction errors should continue to be higher than feature errors.

While GCM can account for the qualitative aspects of the data, the model performs less well at a quantitative level. For both control and deficit subjects, the conjunction error rate is much larger than GCM can predict, while the feature error rate is lower that GCM can predict. One possible explanation for these deviations is that our method of simulating the feature values rather than measuring similarity relations between the faces may have missed an important aspect of the relation between studied faces and the conjunction faces. A stronger test of GCM requires a complete multidimensional scaling solution to use as input to the model, rather than the technique simulated here. However, we also cannot reject the possibility that the binding hypothesis proposed to account for the conjunction errors may indeed be partially correct. Thus, in addition to a familiarity-based mechanism like GCM, subjects may be mis-combining features from studied faces to create novel exemplars in memory. The preceding analysis demonstrates that GCM is a plausible alternative to the binding explanation, but that the extant data is not sufficient to provide a complete test of exemplar-based models. In the following section we describe a memory paradigm that is designed to collect quantitative data that can be used to evaluate exemplar models and therefore assess the need to assume a blending or binding mechanism.

A paradigm to study the mechanisms of false recognition of faces

The preceding analysis of the memory illusion data demonstrates that before a claim can be made for a mechanism that creates new exemplars in memory, quantitative fits of exemplar-based models must be done to disconfirm a familiarity-based explanation of the conjunction error data. The goal of the present work is to provide a rigorous test of exemplar-based models using stimuli in which the similarity relations are known and can be used to evaluate an exemplar-based model. If we are able to demonstrate the inadequacy of exemplar-based accounts, this finding would suggest the need for an additional mechanism that creates novel traces in memory, such as the binding mechanism proposed by Reinitz et al. (1992, 1994) and Kroll et al. (1996). However, it would be necessary to provide a specific representation of such a mechanism within an exemplar model and demonstrate that it provides a much better account of the conjunction errors than that provided by a pure exemplar-based model. We will refer to such mechanisms in generic terms as blending or binding mechanisms, although later we propose the characteristics of such a mechanism.

In addition to addressing the false alarms found with the conjunction stimuli, we have a second set of goals relating to the general applicability of familiarity-based models to complex stimuli such as faces. As Reinitz et al (1994) acknowledge, the Identikit faces used in memory illusion experiments may differ in important ways from realistic faces, and in our own work we have noticed how attributes of the texture of faces provides an extremely potent cue for memory. In an effort to generalize the memory conjunction paradigm and exemplar-based models to naturalistic stimuli, we have chosen to use grayscale photographs of faces as our stimuli. This required that we change the way in which we constructed the conjunction stimuli, in part because it is difficult to combine the features of real faces in an artifact-free fashion. Rather than conjoining the features from studied faces, we averaged the features using morphing techniques. This may actually provide stronger evidence for blending mechanisms. Schooler & Tanaka (1991) distinguish between composite stimuli, which are similar to the conjunction stimuli of the memory illusion literature in that they are composed of pieces of studied stimuli, and compromise stimuli, which are blends or averages of the features of two studied stimuli. They claim that evidence of a blending or binding mechanism is more compelling if demonstrated with a compromise stimulus that averages the features rather than conjoining them.

The morphing process is a two-step process in which control points are placed on important landmarks on each of two faces (e.g. left eye, tip of nose), and the locations of these landmarks are averaged to create a set of average control points. The photographs of the two faces are then digitally warped so that the features of both faces align with the locations of the average control points. With the features in alignment, the two faces are now averaged together on a pixel-by-pixel basis to produce the average face. Figure 1 shows examples of morphs constructed from similar and dissimilar parents, which look quite realistic despite the fact that neither are real faces.

We constructed 16 morphs from pairs of parent faces chosen from a large set of photographs of bald men. The finding that feature errors are greater than errors to novel distractors, as well as the finding that feature errors increase as conjunction errors increase, suggests that similarity is an important variable in false recognition and memory illusions. To evaluate this proposition, we chose our parent faces such that they differed in their similarity. Using a sorting technique (Goldstone 1994; see method section for details), we obtained a rough estimate of the similarity between all possible pairs of parent faces. Using this data we selected 8 pairs of parent faces that were very similar to each other and 8 pairs that were dissimilar. This pre-experiment was designed to provide only a rough estimate of the parent similarities; the exact similarity relations can be determined from the complete set of similarity ratings which were collected as part of the experiment.

In addition to addressing the effects of similarity in the possible blending or binding mechanisms using the above manipulation, we also instituted a temporal separation manipulation during the study sequence, which we hoped would reveal the role of temporal contiguity or context in any binding or blending process. This manipulation was motivated in part by an intriguing finding in the verbal task of the Kroll et al. (1996) study. They manipulated the number of intervening items between the presentation of the two parent words used to construct a conjunction word, such that the two parent words appeared sequentially or separated by 5 words. For patients with left hemisphere damage, they found more conjunction errors when the parent words had been sequentially presented than if they had been presented with 5 words in between. This may suggest a role of the left hemisphere in binding, or in this case, the inhibition of spurious binding. If the inhibitory mechanism is damaged, the binding system is allowed to mis-combine features from different study items and produce an elevated conjunction error rate. As intriguing as this effect of temporal separation is, the effect was found only for the left hemisphere patients, not those with right hemisphere damage, nor with the two sets of control subjects. In addition, this difference does not replicate when pictorial stimuli are used instead.

Despite the fact that the temporal separation manipulation in the Kroll et al. (1996) data produced differences only in one group of patients with one set of materials, we thought it worth while to manipulate temporal separation as an exploratory variable. This would address, for example, whether blending might be more likely to occur for two faces studied together. Alternatively, two highly similar faces presented sequentially might allow subjects to focus on subtle details that distinguish one face from another. To systematically vary the context of the two parent faces, at study we either presented the two parent faces sequentially or separated by at least 20 other faces. The effects of this separation manipulation will show up as differences in the false alarm rates to the associated morph stimuli. For example, if blending is more likely to occur between sequentially studied faces, the morph associated with parent faces presented sequentially will have a higher false alarm rate than when the parents are separated by other faces. We should stress that this manipulation is somewhat exploratory, and finding no effect of temporal separation does not rule out a blending mechanism that blends faces over larger intervals of time than our experiment. However, finding an effect of temporal separation would suggest the properties of a blending or mis-binding mechanism.

We report the results of 3 experiments that address the role of similarity, distinctiveness and temporal separation in the false recognition of faces. In all three experiments, we use pairs of parent faces to create morph blends that will be used as distractors at test. We present the parent faces during the study session, and vary the temporal separation so that either one parent face follows the other at study, or the two parents are separated by at least 20 other faces. The predictions for our distractor stimuli are fairly straightforward. We expect that the morphs will generate more false alarms than other distractors, and that the false alarm rate to morphs constructed from similar parents will be greater than the rate associated with morphs from dissimilar parents. A variety of similarity and sampling models will make these predictions, and thus the models (including those that include a blending or binding mechanism) must be tested at the quantitative level. However, the demonstration that similar morphs have higher false alarm rates than dissimilar morphs would indicate a significant contribution of similarity, which has been ignored in the memory conjunction literature, and provide further support for a summed-similarity explanation rather than a binding error explanation for the conjunction errors.

A third prediction concerns the comparison between the false alarm rates for the morphs (especially the similar morphs) and the hit rates to the associated parents that were used to construct the morph. Related work with faces by Solso & McCarthy (1981) found that subjects were more likely to say old to the conjunction stimulus than to original faces. It is important to point out, however, that such prototype effects are often accounted for by pure exemplar models (Shin & Nosofsky, 1992; Homa, Sterling & Trepel, 1981; Medin & Schaffer, 1978). These models account for the prototype effects because the prototype stimuli, having been used to construct the studied exemplars, are similar to many other items in memory. As a result, they has a high summed-similarity value and therefore a high familiarity value.

Despite the success of exemplar-based models, in some cases exemplar-based models have failed to account for prototype effects. Of particular interest is a finding by Homa et al. (1993), who found a false alarm rate to a prototype stimulus that was below the hit rate of the parent stimuli, but still rejected a pure-exemplar model and had to resort to a version that included prototypes. Because of its ability to distinguish between models, the relation between the false alarm rate to the morphs and the hit rate to the parents is of particular interest when addressing the adequacy of exemplar-based models. In some cases the relation will allow discrimination between models at a qualitative level, although in other cases specific quantitative fits are required to rule out any particular class of models.

The logic of our research is as follows. We generate data from a face recognition experiment that includes distractor faces created by averaging studied faces. We also record similarity relations between the faces, which is used to construct an input space to similarity-based models. Such models represent a face as a point (or exemplar) in multi-dimensional space, and compute values such as similarity based on the distances between the points. If exemplar-based models can account for the false alarms produced in response to the morph distractors, there would be no reason to propose a blending or binding mechanism. However, if such models cannot account for the high false alarm rates, and if we can propose an additional blending mechanism that does, then we would have evidence in favor of a blending mechanism that creates novel exemplars in memory by combining existing exemplars.

In Experiment 1 we compare the hit rates of the parents to the false alarm rates to the associated morphs. In Experiment 2 we extend these findings to a forced-choice paradigm, which at test enables a direct comparison between the morph and one of the two parents. In Experiments 3a and 3b we obtain estimates of the morph false alarm rates and the parent hit rates from separate sets of subjects, which prevents knowledge of the existence of morphs from affecting the hit rates of the parents. To interpret the role of a blending mechanism that may produce the high false alarm rates to the morph distractor faces, we model the individual hit and false alarm rates of each face in the study using extensions of summed-similarity models. The results of Experiment 1 are modeled using Nosofsky's GCM (Nosofsky, 1986) and a variant known as Identification, as well as a new model that attempts to account for both familiarity and distinctiveness using a sampling rule adapted from the SAM model (Gillund & Shiffrin, 1984). This model is used to determine whether the blending mechanisms can be accounted for by an exemplar-based model that uses a summed similarity rule, or whether active blending or binding mechanisms must be assumed that create novel traces in memory. To anticipate our results, we find that no exemplar-based model could provide a complete account of the data, and the best fits came from a version of a model that included a prototype mechanism. This is consistent with the blending or binding hypotheses described in the literature.

Experiment 1:

The goal of Experiment 1 is to measure the hit and false alarms for the parent and morph faces, as well as for the target and distractor faces in which these stimuli were embedded. Although we are specifically interested in the data from morphs and parents, we will also model data from the target and distractors faces. Thus each face in this experiment may be thought of as a condition in and of itself. Subjects saw only one face from a parent/morph triplet at test, and always studied both parents. We measured whether the false alarm rates to the morph distractors from conditions in which the parent faces were studied together or separately; any differences can be used to address the properties of a blending mechanism. In addition, the effects of the similarity of the parent faces on the morph false alarm rates can be determined. Finding a higher false alarm rate for the similar morphs relative to the dissimilar morphs would suggest a significant contribution of similarity in the memory conjunction data, which previous explanations could not account for. In addition, the quantitative relation between similarity and the false alarm rate will prove useful during model testing.

Subjects

Subjects were 180 undergraduates attending Indiana University fulfilling part of a class requirement. There were 45 groups of subjects run with an average of 5 subjects participating at any given time.

Stimuli

The stimuli were 104 pictures of bald men (Kayser, 1985). The pictures were all taken under similar lighting conditions and all men had similar 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, including 36 target faces and 32 parent faces. The parent faces were combined to create 16 morph faces as described below. From the faces remaining out of the original 104, 20 distractor faces were selected. 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. Since we could not select the faces for the various conditions at random, the specification of the features of the target and distractors is important. Of the 36 target faces, 20 had facial hair and there were 8 African-American faces. Five of the African-American faces had facial hair. Of the 20 distractor faces, 3 had facial hair, 3 were African-American and none were both. None of the parent faces had facial hair, and there were 3 African-American faces. This created 16 morphs, 3 of which were blends of a white and an African-American face. Faces with facial hair tend to be more distinctive, which may influence the target hit rate relative to the parent hit rate, for example. However, our critical comparison is between the parents and the morphs, making these differences less critical.

The morph stimuli were created by choosing 8 pairs of faces that were dissimilar and 8 pairs of faces that were similar. Similarity of the parent faces was determined in a pilot study, in which a member of the lab manipulated all parent faces on a computer screen to place similar faces near each other and dissimilar faces far apart according to a procedure described by Goldstone (1994). When repeated for 30 sortings, this produces an output in which the relative similarity between all possible pairs of faces is represented. From these values, pairs of faces were chosen that were either very similar or very dissimilar. 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.

The faces were digitized and displayed on a 21" Macintosh grayscale monitor using luminance control and gamma correction provided by a Video Attenuator and the VideoToolbox software library (Pelli & Zhang, 1991). The background luminance was set to 5 cd/m^2 . The contrast of naturalistic images is difficult to define; here we simply scaled the grayscale values in the images to cover the range between 5 cd/m^2 for black to 80 cd/m^2 for white.

Data was collected by a PowerMac computer using 5 numeric keypads that provided identifiable responses from each keypad. Up to 5 participants completed the experiment at the same time.

Design and Procedure

The study session procedures were identical in all three experiments. There are four types of faces in each experiment. Target and parent faces appear both at study and at test; the only difference between the two sets of faces is that parent faces tended to be less distinctive then target faces because they were all clean shaven, and that the parents were used to construct the morphs. The target faces were a mix of clean-shaven and mustached faces. Morphs and distractors appeared only at test and are therefore distractors. However, the morphs are similar to the parents and we therefore expect higher false alarm rates in general to the morphs than to other distractor faces.

The two levels of parent similarity (similar and dissimilar) are combined factorially with two types of parent separation manipulation (sequential or separate). In the sequential condition the two parent faces used to construct a morph distractor were shown one after another in the study sequence. In the separate condition, the two parent faces were separated by at least 20 other faces. The presentation order of the study faces was randomized for each group of subjects. A particular parent pair was either in the similar or dissimilar condition (because the morphs were computed prior to the experiment). However, the parents were completely counterbalanced, such that a particular parent pair could either be in the sequential or the separate condition for each group of subjects. In addition, the ordering of the parents was also counterbalanced, to eliminate problems that might develop if one parent was always seen first.

Subjects 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.

The ordering of faces in the study phase was important for the temporal separation manipulation. The faces were presented in the following order at study: 8 target faces, 16 parent faces mixed with 4 target faces, 12 targets, 16 parent faces mixed with 4 target faces, and 8 targets faces. Within these constraints the faces were randomized within the study list. The parent faces were selected to appear at a designated position within the parent blocks, as explained below. The blocks reserved for parent faces in the study list ensured that any primacy or recency effects would not affect the parent faces.

The parent faces were divided into two equal groups of 8 parent pairs per group. The sequential parent pair appeared back to back in one of the parent face blocks. In the separate condition, one parent face appeared in the first block of parents, while the second appeared in the second block. These parent faces were randomized within each parent block, with the exception of the sequential parents which appeared one after another. The order of presentation of the two parent faces was varied randomly.

The test phase immediately followed the study session and the subjects were asked to view a series of 72 faces. During the test period for each group of subjects, one face out of each morph/parent triplet was randomly chosen to be presented. This selection was counterbalanced so that across all 24 groups each face in a triplet was shown an equal number of times at test, but a group of subjects saw only one face from a given parent/morph triplet. Thus it takes three groups of subjects to get hit (and false alarm rates) to all three members of each parent/morph triplet. A group of subjects saw 8 morphs, 8 parents, 36 targets and 20 distractors at test, and the presentation order of faces during this phase was randomized. Subjects made old/new judgments to each face using numeric keypads. The subjects were told to respond old if they thought that the test face appeared in the study list, and new otherwise.

There was a practice session at the beginning of all four experiments to ensure the subjects' comprehension. This session consisted of three study faces and six test faces, none of which were used in the study or test phases of the experiments.

Results

The most important comparison in the Experiment 1 data compares the hit rate of the parents to the false alarm rates of the morphs. These data are shown in Table 2, which demonstrate that subjects are just as likely to say 'old' to a similar morph as to an associated similar parent. A repeated-measures ANOVA comparison reveals that the morph false alarm rate is marginally greater than the parent hit rate (F (1,179) = 3.382, p=0.07)[3] for similar morph/parent triplets. Note that while the morph false alarm rate is only marginally higher than the parent hit rate for similar morphs, the morph false alarm rate effect is much larger than that observed in the memory illusion literature, where under similar study conditions the conjunction error rate is typically 1/3 the size of the old stimulus hit rate. The similarity of the parent faces has a strong influence on this effect, actually reversing the direction of the two conditions: For dissimilar parents and morphs, an observer is more likely to say 'old' to dissimilar parents than to the associated morphs (F (1,179) = 53.72, p<0.05).

The average hit rates for the targets was .613 (SEM= 0.008) and the average false alarm rates to the distractors is 0.277 (.011). This hit rate is higher than the hit rate to the similar parents, although since only the target faces had facial hair, this may have made them more distinctive and therefore more memorable than the similar parent faces. There was a significant difference between the similarity conditions for both the parents (F(1, 179)=14.7; p<.05) and morphs (F(1,179)=31.6; p<.05). Similar parents were more difficult to recognize than dissimilar parents, and subjects made more false alarms to similar morphs than dissimilar morphs. It is perhaps not surprising that subjects were more accurate with the dissimilar parents than with the similar parents. When choosing faces that are very dissimilar, we may have chosen faces that were very distinctive, since this distinctiveness allows faces to be very dissimilar (typical faces tend to be more similar to each other). Vokey & Read (1992) and Bartlett, Hurry & Thorley (1984) both demonstrate that distinctive faces are easily identified as old faces, perhaps because distinctive faces are more likely to engage recollective processes.

There was no effect for the sequential/separate condition for parents (F(1,179)=2.14, p> 0.05) or morphs (F(1,179)=.014, p> 0.05). There was also no interaction between the similarity condition and separation for parents or morphs.

Discussion

The Experiment 1 data demonstrate three major effects. First, the false alarm rate for both similar and dissimilar morphs is higher that the false alarm rates for other distractors. Second, the false alarm rates for the similar morphs is higher than that for dissimilar morphs. Third, and most interesting, the false alarm rates for the similar morphs is marginally higher than the hit rates for the associated similar parents. This effect will prove challenging for exemplar-based familiarity models, as described in a subsequent section on modeling.

We find no effects at all of the temporal separation manipulation at study on the false

[3] In this article we use an α level of 0.05 for all significance tests. However, for effects that are trend level we report exact p-values.

alarm rates for either type of morph distractor. This suggests that if a binding or blending mechanism underlies the false recognition to the morphs, it does not have a temporal component. There are other aspects of the data that suggest that temporal separation does not play a large role in face recognition. For example, we plotted the hit rate of the target faces (excluding parents and morphs) as a function of their serial position. Traditional serial position curves show increases at the beginning and end of the study list, which are known as the primacy and recency effects. The serial position curve for the targets is quite flat, with no appreciable upturns either at the start or end of the study list. Thus temporal context may not be as important for perceptual stimuli such as faces as it is for verbal stimuli such as words.

Experiment 1 demonstrates large false alarm rates for the morphs that are consistent with the conjunction errors reported in the memory illusion literature, and also demonstrates a significant contribution of similarity to these effects. The next step is to determine whether a familiarity-based model such as GCM could account for the morph false alarm rates on the basis of summed similarity. However, before doing so we would like to briefly report the results of two related experiments that establish the strength and reliability of the effects reported in Experiment 1. We will then use the evidence and model fits from all three experiments to draw conclusions about the adequacy of exemplar-based models and the evidence in favor of blending mechanisms.

Experiment 2

The goal of Experiment 2 is to provide a direct comparison between parents and morphs in a forced-choice paradigm. The study conditions are identical to Experiment 1, but at test, a morph is always paired side-by-side with one of its parents. This design is motivated from several potential concerns with an old/new recognition paradigm. First, the morphs, by virtue of their construction, tend to lie in a very dense region of MDS space (Busey, 1998). This may result in criterion shifts on the part of subjects that may inflate morph false alarm rates above the hit rates of the parents. Under a traditional signal-detection model, subjects in a forced-choice trial choose the stimulus with the highest amount of evidence. There is no criterion to shift in forced choice. A second potential difficulty with Experiment 1 is that the relatively short, 1500 ms presentation time may have introduced enough perceptual noise that a morph may have seemed perceptually indistinguishable from one of the parents. As a result, subjects may have made a false alarm to the morph thinking it was one of the parents. A forced choice paradigm directly addresses this concern by telling subjects (truthfully) that one and only one of the faces on each test trial appeared during the study session, and they should pick that face. Since one parent is paired with a morph at test, subjects presumably cannot mistake the morph for the parent when making a response.

The experimental conditions were identical to those of Experiment 1, with the exception that at test subjects made a series of 36 forced-choice responses. These responses were always between a target and a randomly-chosen distractor, or between a morph and one of the two parents.

Subjects

181 subjects participated in 25 different groups.

Procedure

Subjects viewed the same combination of faces as in Experiment 1: thirty-two parents and thirty-eight targets in the same presentation conditions as Experiment 1. At test, subjects were given a forced choice recognition test. Subjects were required to pick one of two faces presented that was previously studied. Subjects either chose between a morph and one of the two parents, or between a target and a distractor. There was a total of 36 trials in the test phase: 16 morph/parent pairs and 20 target/distractor pairs in random order. Although there were 36 targets presented at study, only 20 randomly-chosen targets were tested since we have only 20 distractors.

Results and Discussion

The results for Experiment 2 are clear-cut, and are shown in Table 3 as the probability of choosing a parent over the morph for the four conditions. For the similar morphs, subjects were more likely to select the morph over the parent. Averaged over separation condition, the probability of choosing a similar parent over its associated similar morph is .461, which is statistically significantly less than 0.5 (t(180) = 2.68, p < 0.05). This is a somewhat small effect, but it is a qualitative effect that provides strong discriminablility between models, since some exemplar-based models will have trouble accounting for this effect. In addition, it rules out models that suggest that the morph is being confused with one of the parents, since these models could only predict that the morphs would be chosen equally often as their parents. As with Experiment 1, this effect reverses for dissimilar morphs: the probability of choosing a dissimilar parent over the morph is 0.658, which is greater than 0.5 (t(180) = 11.5, p<0.05).

The probability of choosing a target over a distractor was .765. There was a significant difference between the similarity conditions for parents (F(1, 180)=65.4, p<0.05) and for morphs (F(1, 180)=121.8, p<0.05), with is consistent with Experiment 1. There was no difference between studying parents sequentially or separately, nor was there an interaction.

The Experiment 2 data replicate the major findings reported in Experiment 1, which is the tendency for subjects to favor the similar morphs over their associated parents. The reverse is true for morphs constructed from dissimilar parents, and any model that proposes a binding or blending account of these false alarm rates will have to account for the role of similarity.

Experiment 2 represents an extreme case in which parents and morphs are allowed to interact, in that when subjects see a morph, they are also shown its parent. Experiment 1 also mixed parents and morphs at test, although subjects only saw one face of each parent/morph triplet. To eliminate problems associated with mixing parents and morphs at test, in Experiment 3 we gathered independent hit and false alarm rates to the parents and morphs in a between-subjects design. This experiment, which we refer to as Experiments 3a and 3b, also provides for a separate evaluation of the temporal separation effects and a comparison, albeit across-experimental, of the morph false alarm rates and parent hit rates. Experiment 3a also provides a partial replication of Experiment 1 and therefore can be used to test models in a subsequent section.

Experiments 3a and 3b

Experiments 3a and 3b are partial replications of Experiment 1, with the exception that no parent faces are shown at test in Experiment 3a, while in Experiment 3b no morphs are shown at test.

Method

Subjects

There were 119 subjects in Experiment 3a and 112 subjects in Experience 3b. Subjects were tested in 24 groups for each experiment.

Procedure

The study procedures were identical to Experiments 1 and 2. At test, subjects saw all 36 target faces, 20 distractor faces and either 16 morphs (Experiment 3a) or 16 parent faces (Experiment 3b). In Experiment 3b, one parent was randomly selected out of each parent pair. Subjects made old/new recognition judgments using the numeric keypads.

Results and Discussion

For Experiment 3a, the hit rate was .637 (SEM= .015) for target faces and the false alarm rate for the distractors was .281 (.013).

For Experiment 3b, these numbers changed only slightly. The hit rate for target faces was .636 (.010) and for distractors the average false alarm rate was .301(.021). The hit rates are nearly identical, and the false alarm rates do not differ by an unpaired t-test (t(229) = 1.17, n.s.). This indicates that although this was a between-subjects design, the subjects did not differ in their placement of criterion, and therefore the parent hit rates and morph false alarm rates are directly comparable.

For the morphs and parents, the results mirror those of Experiment 1, which replicates the finding that the similar morph false alarm rate is higher than the similar parent hit rate. However, the differences in Experiment 3 are larger than in Experiment 1. The results are shown in Table 4. Averaging across separation condition, for Experiment 3a the similar morph false alarm rate was 0.702 (0.025), while for Experiment 3b the similar parent hit rate was 0.551 (0.027). The morph false alarm rate is significantly higher than the parent hit rate (t(229)=5.81, p<0.05). As in Experiment 1, the reverse held for the dissimilar morphs and parents, which were significantly different but in the opposite direction (t(229) = 6.05, p<0.05). As with previous experiments, there was no effect of temporal separation for either similar or dissimilar morphs in Experiment 3a (F(1,118)< 1.0). There was a significant effect of temporal separation for the dissimilar parents in Experiment 3b, with dissimilar parents more likely to be identified as old if they were studied separately (F(1,112) = 6.0, p<0.05). However, this effect was not found in Experiment 1 and so will not be discussed further. There was no effect of temporal separation for similar parents in Experiment 3b (F(1,112)< 1.0).

Discussion of Experiments 1-3

The results of Experiments 1-3 demonstrate that distractor items that are similar to studied parent faces can attract very high numbers of false alarms. The false alarm rate to the similar morphs is marginally higher that the hit rate to the parents in Experiment 1 and much higher in Experiment 3. In addition, in Experiment 2 subjects chose the similar morphs over the similar parents 54% of the time. The reverse is true for morphs constructed from dissimilar parent faces, although the false alarm rates to the dissimilar morphs is higher than the false alarms to the distractor faces.

Despite the high false alarm rates of the similar morphs, in Experiments 1 - 3 we find no evidence that the temporal separation manipulation affected the false alarm rate for either the similar or dissimilar morphs. This suggests that if a blending or abstraction mechanism underlies the high false alarm rates to the similar morphs, it does not have a temporal or a contextual component that differentiates between sequential presenations and those that differ by ~7 minutes.

The high false alarm rates to the similar morphs, and the tendency for the false alarm rates for even dissimilar morphs to be higher than the false alarm rates to distractors, suggests the possibility of a blending or binding mechanism at work that combines the features from studied faces to produce a false alarm to the morph composites. However, these data are also consistent with familiarity-based models that predict high false alarm rates because morph distractors are similar to many other faces in memory. Below we examine the ability of these models to account for the memory illusions evoked by the morphs.

Quantitative Models of Composite Stimuli False Alarms

The general strategy of our model testing will be to see whether an exemplar-based model can account for the high false alarm rates seen with the similar and dissimilar morphs. In a sense the exemplar-based model will be treated as a null hypothesis, in that if it can be rejected we will have evidence for an additional mechanism that somehow combines exemplars to create novel exemplars associated with the conjunction stimuli. Note that we will also have to propose a specific mechanism within the constraints of the model and demonstrate that it produces a significant increase in the model's ability to account for the morph false alarm rates. We will do this by extending the exemplar-based models to include novel traces that correspond to the morphs. These traces may result from the binding or blending mechanisms that have been proposed in the literature, and in the General Discussion we discuss the properties of such mechanisms.

The absence of temporal separation effects in our data makes model-testing much easier, since many models do not have a temporal component (although these could easily be added in many cases). Of course, the models must account for the false alarm rates to the morphs described above. However, any model of these data must account for other aspects as well. For instance, in Experiment 1 and 3 the dissimilar parent faces had significantly higher hit rates than the similar parent faces. This may result from the fact that when choosing parents that were very dissimilar from each other, we may have been forced to choose faces that were very distinctive. This distinctiveness may have made the faces more memorable and resulted in better activation of recollective processes at test or encoding processes at study. Thus any model of the data must also consider the effects of distinctiveness and typicality, in addition to accounting for the effects of blending or binding.

The Generalized Context Model (GCM) described previously is a reasonable model to use to explain the prototype effects seen with the morph and parent faces, since it can also account for the extant conjunction data in the literature. However, a variant of GCM, known as Identification, has also been used to explain face recognition data at a qualitative level Valentine (1991a, 1991b). In this version of GCM, distinctive faces are more likely to be encoded into memory, and therefore more likely to be identified at test. Below we fit both of these models to our recognition data.

Before describing the modeling procedures, a few words about the representational space are in order. When fitting the conjunction data from Kroll et al (1996), we had to simulate the properties between the faces. However, for the current data we have access to a representational space derived from a multidimensional scaling analysis (MDS) of similarity ratings between all pairs of faces. The details of these ratings are provided in Busey (1998), but a brief description of the procedures is described below.

The target, distractor, parent and morph faces used in Experiment 1-3 totaled 104 faces. This requires (104*103)/2 = 5356 different ratings in order to measure the similarity between all possible pairs. To obtain multiple replications on each pair, we asked 372 subjects to make 170 similarity ratings from this set. These ratings were performed after the subjects had undergone either Experiment 1 or 2 and so were already familiar with the range of faces[4]. Two faces appeared side-by-side on the monitor and subjects made a rating between 1 (most similar) and 9 (least similar). These ratings were converted to z-scores for each subject and combined so that each of the 5356 pairs of faces had at least 8 replications. The resulting similarity matrix was submitted to an multi-dimensional scaling algorithm that produced 6 interpretable dimensions.

After plotting the faces in their locations in MDS space, it became clear that the first four dimensions were age and race, facial hair and facial pudginess. To make these dimensions more interpretable for model-

[4] Twelve additional subjects participated in a pilot version of Experiment 1 that was functionally equivalent to the reported experiment. However, while their data was not included in Experiment 1, they also participated in the similarity ratings experiment and their data is also included in the MDS solution.

testing, independent ratings on age, race, facial hair and facial pudginess were obtained from new subjects for each face and used to rotate the MDS space so that these dimensions correspond to the veridical axes[5]. This does not change the locations of the faces in MDS space, only their appearance as two-dimensional projections. This rotation also allows different attentional allocations on each dimension during model-testing. For example, age may be more important for recognition than race, and a higher attentional weight on the age dimension will reflect this during model fitting.

Accounting for Typicality and Blending: GCM

Nosofsky's Generalized Context Model (GCM, Nosofsky, 1986) has previously accounted for prototype effects using only its exemplar-based structure in recognition memory experiments (Nosofsky, 1988; Shin & Nosofsky, 1992; Nosofsky & Zaki, in press). The basic functions of GCM are given in Eqs 1-5, which define how similarity is computed from distance between faces in MDS space, and describe how summed similarity is translated into a predicted probability of saying old to targets and distractors. In addition to the free parameter c and those used to convert familiarity to probability, GCM also includes attentional weights on each dimension that serve to enhance one set of features over another (Nosofsky, 1986, 1991). These weights, w_n, are used to stretch and shrink the dimensions of MDS space and therefore the distances computed in Eq 1 prior to the computation of similarity.

GCM was applied to the Experiment 1 data by assuming that 32 target faces and 32 parent faces were placed into memory[6]. For each test item (including targets, distractors, parents and morphs), the summed similarity for each item was computed to give a familiarity value via Eq 3 and then converted to a probability via Eq. 4, with a logistic function mapping familiarity to the probability of saying old. There were 8 free parameters: c, the six attention weight parameters (which sum to 1.0, giving 5 free parameters) and the two parameters of the logistic function mapping familiarity to probability. The number of parameters is quite small, considering that we are fitting 100 data points. Table 5 lists the estimated parameter values for all model fits.

Of particular interest is the ability of GCM to place the similar morph distractors above the similar parents. To test this, we fit the data to just the morph and parent data, and the results are shown in the top panel of Figure 2, which plots the probability of saying old from the data on the ordinate against the predicted probability of saying old on the abscissa. This particular fit emphasizes GCM's ability to account for the relation between the morphs and parents, since this is the motivation for applying GCM in the first place.

The scatter plot allows a direct comparison of the data to the model's predictions, since if the model can account for the data perfectly, the results should fall on the oblique line. Although the overall fit is quite poor, GCM does a reasonable job placing the morphs (which are distractors that happen to be similar to several studied faces) above the other distractors. In addition, the model

[5] This procedure was developed by Rob Nosofsky, and uses external ratings to rotate the MDS space such that the external ratings correlate with the locations of the stimuli along each axis.

[6] The original experiments had 104 faces, which included 36 target faces. However, all of our available MDS programs were limited to 100 stimuli (0-99), having apparently been programmed by the same software engineers responsible for the Y2K problem. As a result, we eliminated 4 target faces in such a way that neither the mean nor the variance of the overall target hit rate changed. To verify that this deletion did not affect our results, we actually did this twice, choosing a separate set of 4 faces to delete each time. The overall fits remained quite similar, and therefore we don't believe that the deletion of these faces affects our results.

places the similar morphs above the associated parent faces, which can be seen by comparing the solid diamonds with the open diamonds: the solid diamonds are shifted systematically to the right of the open diamonds. This is in accord with the experimental findings that the false alarm rates to the similar morphs are higher than the hit rates to their parent faces.

While GCM can account for the prototype effects seen with the morphs and parent faces, it cannot simultaneously account for other aspects of the data. When GCM is fit to the entire data from Experiment 1 as shown in the bottom panel of Figure 2, the overall model fit improves, but the model no longer predicts that morphs are chosen as old more often than the parents. Thus the model cannot account for the prototype effects seen with the morphs and simultaneously capture the other aspects of the data. Part of the difficulty with the fit comes from the fact that typical faces are still predicted to be called ’old’ more often than distinctive faces, which is contrary to the data from similar and dissimilar parents: dissimilar parents tended to be more distinctive and had a higher hit rate than similar parents, as shown in Table 2.

Figure 3 demonstrates why GCM may have had difficulty accounting for all aspects of the face recognition data. This figure shows the probability of saying old to targets (including parent faces) as a function of the typicality of the parent and target faces, which is defined as the summed similarity to all other studied faces. Similarity was computed directly from the MDS distances, using Eq 2 and a c value of 2.0. Included with this figure is the best fit of a second order polynomial, which demonstrates a U-shaped function that results from both very distinctive and very typical faces generating very high hit rates, with moderately typical faces generating somewhat lower hit rates. To see how the different models can account for distinctive and typical faces, we have circled two faces that have similar oldness ratings but lie at opposite ends of this scale. When these same points are circled in Figure 2, we see that GCM can account for the typical face but not the distinctive face, and places typical faces to the right of distinctive faces where the reverse is true in the similar and dissimilar parent data.

The U-shaped function of Figure 3 is consistent with results from Vokey & Read (1992) that suggest that much of face recognition appears to be driven by processes in addition to a familiarity mechanism. They describe a memorability component that is presumed to represent encoding and retrieval processes underlying an active search of memory and which could account for the finding that distinctive targets are easily recognized and distinctive distractors are easily rejected. The similarity of very typical faces to other faces in memory may result in saying old on the basis of what they describe as context-free familiarity, while very distinctive faces may facilitate active memory encoding and retrieval processes and result in an old response on the basis of recollective processes. GCM assumes that the more similar an item is to stored items in memory, the more likely the subject will say ’old’. As a result, GCM includes only the familiarity mechanism described by Vokey & Read (1992).

Accounting for Distinctiveness: Identification

The inability of GCM to account for the high hit rates to distinctive target faces has been noted by Valentine (1991a, 1991b), who suggests that a version of GCM known as Identification (Nosofsky, 1986, 1987) may be more appropriate when accounting for face recognition. In this version of GCM, distinctive faces are more likely to be encoded into memory, and therefore more likely to be identified at test. Similarity is still constructed from the MDS space via Eq 1 and 2, but the probability of calling face i old is determined by,

$$P(\text{"old"} \mid i \;\; presented) = F\left[\frac{\underset{\substack{j \subset All\;\; Faces \\ In\;\; Memory}}{Max}\, [\, \eta_{i,j}\,]}{\sum\limits_{\substack{j \subset All\;\; Faces \\ In\;\; Memory}} \eta_{i,j}}\right]$$

Eq 6

where F is again the logistic function of Eq 5 that in this case maps the similarity ratio into a probability.

For target and parent faces, this model is the inverse of GCM, since the max rule in the numerator of Eq 6 will be 1.0 if a face was presented at study. This model makes an inverse prediction as well: whereas GCM predicted that more typical faces are more likely to be called old, the identification version predicts that distinctive faces are more likely to be called old. This results from the fact that a distinctive face will have a very low summed similarity and therefore a larger overall fraction. Thus this property of the model can account for the high hit rates to the distinctive targets seen in Figure 2. This model is not quite the inverse of GCM for distractors, since the numerator is no longer 1.0, and this makes the model potentially able to account for the high false alarm rates to the morphs as well. A very typical distractor will have a large numerator relative to a distinctive distractor, and thus it may potentially have a higher false alarm rate as well.

We fit the Identification model of GCM to the Experiment 1 data, using Eqs 1-2 and Eq 6 along with a logistic function F to map the ratio of the max similarity to the summed similarity into a predicted hit or false alarm rate. Figure 4 shows the resulting fit of this model, which includes 8 free parameters: c, 5 attention weights, and two logistic function parameters. Table 5 lists the estimated parameter values for all model fits.

While this model is doing a much better job accounting for the high hit rates to the distinctive targets, there are systematic deviations that suggest that the model is failing in interesting ways. First, the model cannot account for the high false alarm rates to the similar morphs: the filled diamonds are all shifted to the left of the open diamonds. Apparently the advantage in the numerator of Eq 6 for the morphs is dwarfed by the summed similarity in the denominator.

In addition to failing to predict the high false alarm rates to the morphs, the Identification version of GCM also has difficulty fitting data from old items. The open diamonds represent parent faces, which by virtue of their clean-shaven faces tend to be more typical than other target faces. These open diamonds appear shifted to the left of the crosses, suggesting that the model cannot account for typicality. As in Figure 2, the two circled points show two studied faces that have nearly identical hit rates in the data. The Identification version of GCM has trouble placing them near each other: as it tries to fit the typical face, it pushes the distinctive face above the prediction line.

The successes and failures of both versions of GCM are instructive for further model development, and demonstrate the need to account for both the high hit rates of distinctive targets, perhaps due to recollective processes that take advantage of distinctiveness, as well as the high hit rates and high false alarm rates of very typical parents and morphs, which may be mediated by a familiarity-based process. In the next section we propose just such a model, termed SimSample, and then show how it can simultaneously account for the effects of typicality and distinctiveness. As we will see, however, this model will also have difficulty accounting for the very high false alarm rates to the similar morphs, which suggests that exemplar-based models may not be sufficient to account for the prototype effects seen with the morphs and parents.

Accounting for both Typicality and Distinctiveness: The SimSample Model

To account for both typicality and distinctiveness that may contribute to familiarity and recollective processes respectively, below we develop a model that uses the similarity structure of the face space in conjunction with a sampling rule that samples items from memory. In designing this model, we rely as much as possible on proven mechanisms from categorization and memory research. In addition to adopting the MDS structure from GCM, we also adapt the sampling mechanism from the SAM model (Gillund & Shiffrin, 1984). This provides support for the various model assumptions, since the mechanisms that SimSample adopts have been quite successful in the categorization and memory domains.

The SimSample model samples from memory according to the similarity of a face in memory to the test face; thus the name SimSample. 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 at test, 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,

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

which is simply the Luce Choice Rule[7]. We then 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,

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

where the similarity criterion is a free parameter and $\eta_{i,k}$ is the similarity between test face *i* and sampled face *k*. 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

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

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,

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

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 subjects. 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 critSD,

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

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 $\theta(\eta_{i,k})$.

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 ver-

[7] This is similar to the sampling rule proposed in Gillund and Shiffrin's SAM model (1984), although SAM uses strengths rather than similarities to compute the Luce Choice ratio.

sions of SimSample were attempted, with little success.

The SimSample model has 8 free parameters (which is the same number as the two previous models): 1 generalization gradient parameter c, 5 attention weights, the response criterion and the standard deviation of the response criterion.

Summary of the model and predictions

The SimSample model embodies several principles. Faces are placed into memory at study and are represented in terms of their locations in the MDS face space. At test, a test face i is used to sample one face from memory. This sampling process is biased such that faces nearby in face space are more likely to be sampled. Once a face is sampled, it is compared with the test face and if they are similar enough, the observer says ’old’. Note that in terms of model fitting, targets and parents are studied items, while distractors and morphs appear only at test. Thus distractors and morphs do not have a trace in memory to be sampled, but we do know the locations of the distractors (including the morphs) in MDS space and can use this to compute the similarity to other items in memory.

In order to account for the data from Experiment 1-3, the SimSample model must account for four characteristics of the data:

1) As Figure 3 demonstrates, distinctive target faces have very high hit rates. This may result from recollective mechanisms taking advantage of the distinctive features of these faces.

2) Distractors (which tended to be distinctive) have very low false alarm rates.

3) Morphs created from similar parents have a false alarm rate that is higher than the hit rate to the associated parents. This prototype-like effect is consistent with a blending hypothesis, although an exemplar-based model may be able to account for this effect as described below.

4) As Figure 3 demonstrates, very typical parents have higher hit rates than moderately typical parents, and some as high as distinctive faces.

As we have seen, neither GCM nor the Identification version could account for all of these effects simultaneously. Below we describe how in principle the SimSample model can account for each of these aspects of the data.

Effects of Distinctiveness

The upper-left panel of Figure 5 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 $\eta_{i,k}$ < criterion and the observer will incorrectly say ’new’. Thus the SimSample model correctly predicts that more distinctive target faces will have higher hit rates than less distinctive targets.

The upper-right panel of Figure 5 demonstrates how SimSample accounts for the low false alarm rates to distinctive distractors. As with target faces, a distinctive distractor 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 distractors have above-zero false alarm rates, but the model predicts (correctly) that distinctive distractors will have the lowest false alarm rates.

Effects of Typicality

The bottom panels of Figure 5 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 distractors.

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.

Accounting for High Morph False Alarm Rates

The development of the SimSample model has two goals. First, we would like to account for the effects of distinctiveness and typicality within a single-process model, and the above discussion demonstrates how the model, at least at the qualitative level, can do this. The second goal is to investigate abstraction or blending mechanisms that may underlie the finding that observers are more likely to say old to a morph than to its parents, as long as the parents are similar. As described above, the SimSample model predicts higher false alarm rates to similar morphs than dissimilar morphs or other distractors. However, it is less clear that the model can account for the Experiment 2 finding that observers are *more* likely to say old to a morph (which they have never seen) than to its associated parents (which they have seen). In principle the model can account for this effect. Consider the logic described in the bottom panel of Figure 5. Suppose that the gray circle is a morph and the two black circles are its parents. In this case the morph has two opportunities to sample a nearby face that would produce an old response. However, if one of the parent faces was tested instead, the criterion circle may exclude the other parent, making the parent sample itself in order to produce an ’old’ response. This situation would produce predictions that are consistent with the data: the morph would have a higher false alarm rate than the hit rate to either parent. However, the criterion is constrained by other aspects of the data, and therefore it may be possible to falsify this version of SimSample when fitting the oldness ratings to all of the data.

Figure 6 shows the fit of the SimSample model to the Experiment 3 data. Overall the fit is much better than either GCM or the Identification version of GCM, with a root-mean squared error (RMSE) of 0.1462. This good fit is obtained with no increase in the number of free parameters over the previous two models. Our first goal in developing the SimSample model was to account for both the effects of typicality and distinctiveness. The two circled data points in Figure 6 are the same typical and distinctive faces shown circled in Figures 2 and 4. The previous model fits placed these two points far apart, despite their almost identical hit rates. However, the SimSample process successfully places both faces quite near each other and near the diagonal. Similar examples can be found throughout the data, demonstrating that the model can correctly account for the high hit rates to both very typical and very distinctive faces. Thus effects attributable to both familiarity and recollective mechanisms can be accounted for by the single-process SimSample model.

In addition to incorporating the effects of both typicality and distinctiveness, the SimSample model also correctly predicts the low false alarm rates to the distractors, and accounts for the higher false alarm rates to the morphs than the distractors, even though neither was seen at study. Thus of the four characteristics of the data described above, the SimSample model can account for the first, second and fourth characteristics. There are several systematic deviations in the data, most notably the fact that for studied items there appears to be more variability in the data than in the predictions, which gives the overall fit a curvilinear shape. This might result from aspects of memory that are not captured by the similarity ratings. For example, a small blemish on a face might make it very memorable, but subjects may ignore this when making similarity ratings. In addition, very typical faces may be judged against some other more global aspects of faces (Levin, 1996) rather than just the similarity to studied faces. While these suggest future work, overall the model captures most of the qualitative and quantitative aspects of the data.

What is more important to our central question is whether the model can account for the fact that observers label the similar morphs as old more often than the associated similar parents. Despite the success with distinctive and typical targets and distractors, the SimSample model places the similar morphs systematically below the parents. Table 6 shows the mean oldness ratings for targets and distractors, as well as both similar and dissimilar morphs and the parents, along with the SimSample model predictions for all six types of stimuli.

Clearly this exemplar-based version of the model has trouble accounting for the high false alarm rates to the similar morphs. There are two possible reasons for this. First, our current exemplar-based model, which accounting for many of the effects of typicality and distinctiveness, may not be formulated correctly to handle the behavior of faces that are very similar to other faces. Various extensions to this model were attempted in the hopes of saving the exemplar based version of the model. However, none of the extensions had much of an improvement over the current version of the SimSample model, and none could push the morph false alarm rate over the parent hit rate for the similar morphs.

Accounting for Blending: Prototype mechanisms

When faced with the apparent failure of an exemplar-based model to account for prototype effects such as the high false alarm rates to the morphs, a traditional approach within the categorization literature is to extend an exemplar-based account with hypothesized prototypes that are assumed to exist in memory as the result of some abstraction or blending process. When working in MDS space, an exemplar model is altered to include the assumption that at study some form of binding or blending mechanism creates faint traces called prototypes at the locations of the morphs in MDS space. For our present purposes we will not specify the exact nature of this mechanism, only indicate that its output creates a new exemplar (the prototype) in MDS space at the locations of the morphs. In the General Discussion we will consider possible process-based models that would produce such a new exemplar. Evidence for or against the existence of such a mechanism is found by adding a weighting parameter to the prototype component and fitting this as a free parameter. In some cases the estimated value is zero, implying no prototyping or abstraction mechanisms (e.g. Shin & Nosofsky, 1992). In other cases, the estimated parameter value is above zero, which has been interpreted in the categorization literature as evidence for some form of abstraction mechanism (e.g. Homa, et al., 1993).

With this logic in mind we introduced a prototype mechanism into the SimSample model. This was done by assuming that as the two parent faces are studied, a combination of the two faces is also created in memory. The location of this combination is assumed to be near the average of the two parent faces in MDS space, although for the purposes of the present model fits we assumed that the prototype was created at the location of the morph in the MDS face space. Previously Busey (1998) demonstrated that the morph face was quite close to the expected location in MDS space, with the exception of two small biases produced by the morphing operation. Thus there appears to be a close correspondence between the actual and expected location of the morph in MDS space, although since we know (and use) the exact location of the morph in MDS space this is less important.

A prototype mechanism for the morphs was introduced by modifying Eq 7 as follows. We assume that the prototype acts as a weak or faint trace in memory that is possible to sample via the SimSample process. The strength of the prototype trace is a free parameter, and the strength 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,

$$P(sample\ \ k\,|\,i\ \ presented) = \frac{\eta_{i,k}}{\sum\limits_{\substack{j\subset All\ \ Faces\\ In\ \ Memory}} \eta_{i,j} + \sum\limits_{j\subset prototypes} pw\ \ \eta_{i,j}}$$

Eq 12

for faces actually studied, and,

$$P(sample\ \ k\,|\,i\ \ presented) = \frac{\eta_{i,k}}{\sum\limits_{\substack{j\subset All\ \ Faces\\ In\ \ Memory}} \eta_{i,j} + \sum\limits_{j\subset prototypes} pw\ \ \eta_{i,j}}$$

Eq 13

for the morphs. In both Eq 12 and Eq 13, pw is the prototype weight that is freely estimated. 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[8]. 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,

$$P("old"\,|\,i\ \ presented) = \sum_{\substack{k\subset\ faces\ \ in\\ memory}} P(sample\ \ k\,|\,i\ \ presented)\Theta(\eta_{i,k}) + \sum_{k\subset\ prototypes} P(sample\ \ k\,|\,i\ \ presented)\Theta(pw\ \ \eta_{i,k})$$

Eq 14

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 proto-

[8] 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.

types 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. In a later section we discuss versions of the model that assumes that prototypes are created for all possible pairs of faces, not just between parent faces.

The addition of prototypes to the SimSample model reduces the RMSE by a modest amount, with the RMSE reducing to 0.1441, which not quite a significant improvement in error[9] (F(1,90) = 3.67; p=0.06). Inspection of the fits, however, suggests that an alternative mechanism proposed by Metcalfe (1990) might be a more appropriate mechanism. She suggests that blending is more likely to occur between similar items. Knapp & Anderson (1984) propose a similar model in which prototypes only form between similar exemplars. In our data, the fits to the similar morphs do increase with the addition of the prototype, but so do the dissimilar morphs by an equal amount. Inspection of the original SimSample model fit in Figure 6 reveals that the SimSample model was already fitting most of the dissimilar morphs, and the addition of a prototype actually hurts the fit of these faces. This suggests that different prototype values should be assigned to the similar and dissimilar morphs. We fit such a model, and found that it did a much better job than a single prototype version. We then generalized this rule such that the prototype value for each morph was proportional to the similarity between the two parents. For morph i, this was computed as,

$$pw_i = (\eta_{i,p_1} + \eta_{i,p_2})\ \rho \qquad \text{Eq 15}$$

where pw_i is the prototype weight for morph i for use with Eqs 12, 13 and 14, η_{i,p_1} is the similarity of the morph to parent 1 and η_{i,p_2} is the similarity of the morph to parent 2. The free parameter ρ is the constant of proportionality that relates the similarity of the parents to the appropriate value for the prototyping effect. The parameter pw_i is no longer freely estimated, but instead depends upon the free parameter ρ and the similarity of the two parent faces p_1 and p_2.

The resulting model fit of the SimSample model with proportional prototypes is shown in Figure 7. The fit to the morphs is much improved over Figure 6 (SimSample), with a RMSE of 0.1411. This reduction in error relative to the original SimSample model is statistically significant despite the free parameter[10] (F(1,90) = 7.70; p<0.05) and it is better than the previous prototype model which had the same number of free parameters. More importantly, this model now correctly predicts that similar morphs will have a higher oldness rating than their parents, as shown in Table 6. This model also does a much better job accounting for the dissimilar morphs as well, although it still has a trouble accounting for the dissimilar parents, which tend to be more distinctive. Surprisingly, this model improves the fit to all 16 morphs with the addition of only 1 free parameter, and produces a qualitative change in the predictions. Although the model is only accounting for about 57% of the variance, and an hypothesis test would certainly reject the model, it is correctly ordering the means in Table 6 and the model's predictions are much improved over previous models based on GCM.

The prototype extension implied by Eq

[9] The formula for computing the F ratio is: $F^* = \dfrac{SSE(R) - SSE(F)}{df_R - df_F} \div MSE(F)$ where SSE = RMSE2 * (n-p) for the full and reduced models (n is the number of data points and p is the number of free parameters). This statistic is distributed approximately $F(df_R - df_F, df_F)$ when Ho holds.

[10] In all model fits the RMSE is corrected for the number of free parameters, such that the sum of squared errors is divided by the number of data points minus the number of free parameters.

15 was only computed for the morph locations. This is an arbitrary assumption, and although we were only probing these locations at test, any blending mechanism must be assumed to act in some capacity between all pairs of faces, not just between the 16 arbitrarily-chosen parent pairs. We implemented a model that assumes that prototypes are created between all pairs of faces, the faintness of which was proportional to the similarity of the pair of faces. Since there are 64 faces in memory (32 targets and 32 parents), this implies the creation of ((64-1)*64)/2 = 2016 prototypes. The faintness of the prototypes was computed via Eq 15, and if a prototype was sampled, the probability of saying old was given by the second half of Eq 14. This model produces a RMSE of 0.1421, which is comparable to that produced by the original SimSample with Proportional Prototypes model and uses the same number of free parameters. It also produces a statistically significant reduction in error relative to the original SimSample model (F(1,91) = 6.36; p<0.05). Computationally this model is expensive, since it must compute the sampling process across all 2080 items in memory, making this model somewhat difficult to work with. A version that included a pruning mechanism that allowed prototypes to form between pairs of faces that were more similar than a criterion was also fit, which produced a RMSE of 0.1377. This is a significant error reduction relative to the original SimSample model despite the fact that this model has two additional free parameters (F(2,90) = 6.82; p<0.05). Together the success of all three prototype models demonstrate that a prototype mechanism improves the fit of the SimSample model despite the additional free parameter or parameters.

Converging Evidence: Fits to Forced Choice Data

No new model, especially one that includes prototypes, should be based on the evidence of only one paradigm or experiment. In the following section we use the data from Experiment 2 and Experiment 3a to provide converging evidence that 1) current exemplar-based models are insufficient to handle face recognition data, 2) the SimSample model does a much better job fitting data from face recognition experiments, and 3) a version of SimSample that includes prototypes still does a better job of accounting for the high false alarm rates to the similar morphs than a version without prototypes. All of these conclusions are entirely consistent with the results of the Experiment 1 SimSample model fits.

A forced-choice task differs in important ways from old/new recognition. First, the subject does not have to set a criterion for saying old or new. Their task is to simply choose the face that they believe occurred during the study session. This eliminates criterion shifts that may selectively affect the morphs. Second, when considering the rate at which the subjects choose face a over face b (which is our dependent measure in forced choice) must consider the relative evidence for a over b. One formulation, adapted from categorization tasks, takes the form of a relative comparison involving an exponential,

$$P(Choose\, a \mid a \;\; and \;\; b) = \frac{e^{\zeta A}}{e^{\zeta A} + e^{\zeta B}} \qquad \text{Eq. 16}$$

where A represents the evidence for face a, B represents the evidence for face b. The free parameter ζ 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 ζ, virtually all choosing probabilities will be close to 0.5, since the exponents will all be close to 0. However, for large ζ, this emphasizes the impact that A and B can have, such that if A dominates B only slightly, the subject will be very likely to say “a”. Thus ζ can be thought of as a confidence parameter that indicates how much the evidence of a over b influences the resulting choosing rate for face a.

Eq. 16 can be used to adapt all three

major models to produce predictions for the forced-choice recognition paradigm if we assume that the subject computes the probability of saying old to faces a and b via the equations appropriate to each model, which provides the values A and B for Eq 16. In situations where A and B are about equal (that is, subjects are equally likely to say old to faces a and b), the probability of choosing face a will be close to 0.5. However, as one face tends to dominate, Eq 16 will get closer to 1.0.

The fits to the three major models are shown in Figure 8. The conclusions mirror those from Experiment 1. Neither GCM nor its Identification version fit very well, with a tendency to bunch predictions together for the various stimulus types. In addition, neither model succeeds in placing the similar morphs above the similar parents. The SimSample fit is much better, with significantly lower RMSE (0.1120) despite having the same number of parameters. However, it too has difficulty placing the similar morphs above the similar parents, as shown in Table 6. Best fitting parameters and error values for all models are shown in Table 5.

The prototype extensions that were applied to the SimSample model in Experiment 1 were also implemented for Experiment 2. A single prototype model produced a significant decrease in RMSE, which was 0.1060 (F(1,90) = 11.48; p<0.05), while the proportional prototype version produced an even better fit, with an RMSE equal to 0.0968 (F(1,90) = 30.75; p<0.05). In addition, the prototype models were the only versions of SimSample that could place the similar morphs above their parents, as shown in Table 6. A version in which prototypes were created between all possible pairs of faces (conditioned on a threshold similarity) accurately predicts that similar morphs would be chosen slightly more often than the similar parents, and produced a RMSE of 0.1066, which is significantly better than the original SimSample fit. These results confirm the Experiment 1 data and demonstrate that even with a different test procedure, a prototype mechanism still provides a better fit of the exemplar-based SimSample model.

Additional support for the SimSample model comes from fits to Experiment 3a, which is a partial replication of Experiment 1. This experiment was identical to Experiment 1 except that it did not include the parents at test. Nevertheless, the data still provide a test for the various models, since the rest of the data still constrain the predicted morph false alarm rates. Thus these fits, in addition to providing a replication, demonstrate how inter-related the various stimuli are, such that the morphs may not be fit even in the absence of constraining factors from the parent hit rates.

The fits mirror those from Experiment 1, and the best-fitting parameter values are shown in Table 5. SimSample produced a better fit than either GCM or the Identification version. In addition, both versions of the prototype mechanism produced significantly better fits that the original SimSample model, despite the additional free parameter for each prototype version. For a single prototype version, the RMSE was 0.1397 (F(1,91) = 4.95, p<0.05), and for the proportional prototype version, the RMSE was 0.1391 (F(1,91) = 5.77, p<0.05).

Summary of Modeling

Across three experiments and two paradigms, we find support for the SimSample model over previous exemplar-based models. Versions of SimSample that included a prototype extension produced significant decreases in error in all three experiments, and the version that included a mechanism that produced prototypes on the basis of the similarity between parent faces produced the best fits in all three experiments.

Summary and Conclusions

The major goal of the present work is to investigate the possible contributions of blending or binding mechanisms in face recognition. Previous work using conjunction stimuli composed of parts of studied faces found evidence for high false alarm rates to the conjunction stimuli. However, these authors fail to consider similarity or familiarity as an alternative explanation. Simulations of the GCM model suggests that while GCM could account for effects that a binding explanation cannot, the model produced poor quantitative fits. However, these poor fits may have resulted from our simulation of the similarity relations, not from any inadequacy in the model itself. The current experiments provide enough data to test existing exemplar-based models.

To address the role of blending or binding mechanisms in a robust paradigm that would allow for model testing, we constructed physical blends that were used as distractors at test. The data from three experiments demonstrate extremely high false alarm rates to morphs constructed from similar parents, although we never find effects that can be tied to whether the parent faces are studied sequentially or separately.

Existing exemplar based models based on summed similarity could not simultaneously account for the high morph false alarm rates and the high discriminability of distinctive faces. To explain these effects we developed a model, termed SimSample, which could place a very typical face at the same predicted oldness rating with a very distinctive face. This model can therefore account for much of the effects of the similarity structure between the faces, including possible contributions of familiarity and recall-based mechanisms. However, the SimSample model could not place the similar morphs above their parents.

Two different versions of prototype models were fit to the data from the three experiments, which include both old/new recognition and forced-choice paradigms. A prototype mechanism is assumed such that the locations of the morphs in MDS space act as very faint traces in memory, and contribute partially to the SimSample process. In general, the prototype versions of SimSample performed significantly better than the original version of the model. In all three experiments, the best fitting model included a prototype mechanism in which the contribution from each prototype is not fixed for all morphs, but instead is related to the similarity of the morph to its parents. Note that these different prototype weights for each morph that appear to be required are above and beyond the effects of similarity of the morph to its parents, which the exemplar-based version already assumes. The original SimSample process can correctly place similar morphs above dissimilar morphs due to the similarity between the similar morphs and their parents. It also appears that similar morphs need proportionately more prototype weight than the dissimilar morphs. This is consistent with the idea that blending between two faces is more likely to occur for faces that are closer together.

Finding a prototype model that can account for the high false alarm rates to the morphs does not, of course, rule out all exemplar-based models. However, we have fit the leading exemplar-based models, including many different variants of a model we developed, and found no exemplar model that could account for the data from the similar morphs. These models have successful accounted for a wide range of phenomenon in categorization and recognition, and there was every expectation that they would provide a good account of the present design as well. These models have previously been applied to stimuli such as random-dot patterns and schematic faces, which has the advantage of reducing the idiosyncratic responding on the basis of the subjects. However, this behavior may have contributed to our distinctiveness effects seen with naturalistic faces, and therefore requiring the SimSample model to provide an account of distinctiveness. In addition, in most categorization and recognition experiments using prototypes, a single prototype is created from a set of relatively homogenous exemplars such as polygons or random dot patterns, while we had 16 different prototype stimuli which placed more constraints on the model.

Any exemplar model that attempts to account for these data must have mechanisms that can simultaneously account for the high false alarm rates to the morphs and the excellent discriminability of the distinctive faces. Thus, an exemplar-based model must include some additional machinery other than summed similarity to account for these effects. We attempted a variety of model extensions, all of which were unsuccessful at improving the fit of an exemplar-based version of SimSample to the similar morph false alarm rates. These models include mechanisms such as an expanding criterion that depends upon the summed similarity between the test face and all faces in memory, a clustering mechanism, and other extensions. None of these exemplar-based extensions could account for the high false alarm rates seen with the similar morphs.

These results are somewhat at odds with the conclusions of Homa et al. (1993), who suggest that it is unlikely that subjects are able to produce abstractions when the training patterns are seen only once at study. However, their random polygons may have been less familiar to subjects than our faces, and our experience with faces may make abstractions such as blending easier to accomplish.

An alternative explanation for the prototype effects seen in the data is that the SimSample model is incorrect in its account of items that are similar to other items in memory. This effect of typicality was handled within the SimSample model by favoring nearby items during the sampling process. If a nearby item is sampled and is also inside the decision criterion, the model predicts the observer will say ’old’ even though the wrong item was sampled. While this process was successful at accounting for the hit rates of both very distinctive and very typical faces, it was insufficient to account for the morphs unless an explicit prototype mechanism was assumed. Rather than requiring a prototype mechanism, the model’s account of typicality might be wrong, and some other mechanism applied to the exemplars in memory could account for the prototype effects seen in the morphs. Although this could be the case, we would like to stress that all of the assumptions underlying the SimSample model come from models based on a vast literature involving categorization and recognition memory. The generalization gradient relating distance to similarity comes from Shepard (1974), the geometric model comes from the categorization literature (see Nosofsky, 1992 for a review), and the sampling and decision rules are adapted from the recognition memory literature, including the SAM model (Gillund and Shiffrin, 1984; Raaijmankers & Shiffrin, 1981). It is interesting to note that the sampling and decision rules have usually been applied to recall rather that recognition, and the success of these rules in the present case suggest that similar mechanisms may be at work in areas as diverse as word recall and face recognition.

A second alternative explanation for the prototype effects seen with the morphs is that there is something special about morphs, such that observers may treat morphs differently. A detailed analysis of the location of the morph stimuli in MDS space was performed on the current faces by Busey (1998), in which potential biases introduced by the morphing operation were examined. This analysis revealed that while morphs appear more typical than the parent faces, such effects were due solely to the geometry of MDS space, since the average of any two points in a high-dimensional space is likely to be closer to the centroid than either two points. Two other biases were identified: morphs tend to appear younger and pudgier than predicted on the basis of the parent faces, which biases the morph in MDS space. However, all of these biases were taken into account by the location of the morphs in MDS space and therefore cannot affect the resulting model fits in the present case. It remains possible, however, that some aspects of the morphing operation may not show up in the similarity ratings (and therefore not be represented in the MDS solution), such as the tendency for the morph to appear smoother than real faces. This in turn may have increased the probability of saying 'old' to the morphs. One way to address this issue is to perform the model fits using the raw similarity values rather than the output of the MDS, mapping rated similarity to computed similarity using a power function. We fit several versions of SimSample with the raw similarity ratings to both the Experiment 1 and 2 data, and found that a) overall the fits were somewhat worse than the MDS-based fits and b) none of the models could account for the prototype effects seen with the morphs. Thus it appears that the MDS algorithm may have reduced the noise seen in the raw similarity ratings, but doesn't do so by removing critical information that otherwise would allow SimSample to account for the prototype effects. Even if this explanation is correct, it would not explain *why* subjects exhibit the bias to respond 'old' to the morphs.

Other models from the memory literature might also account for the effects of typicality and distinctiveness seen in the present data, and these models might also predict the false alarm rates to the morphs. Hintzman's Minerva 2 model has previously been shown to be able to account for abstractions (Hintzman, 1986). The challenge with such models is to construct a set of input vectors that represents the similarity relations between the faces as expressed by their locations in MDS space. One possibility that we begun to explore in related work is to use the coordinates of the six MDS dimensions as input to Minerva 2. In preliminary investigations, the model could not account for the high false alarm rates to the similar morphs. However, it is not clear that the MDS coordinates represent the best input vectors to this model, and we are continuing to pursue these investigations.

Characteristics of a Prototype Mechanism

In proposing a prototype mechanism, we have specified only its output (new exemplars in MDS space) rather than its process. If such a mechanism does exist, we might specify its likely characteristics.

First, such a mechanism appears not to have a temporal component, or if it does the temporal window is longer than the 5 minute study period of each experiment. This might be consistent with a blending mechanism that works without conscious input from the subjects, since any control processes that may have been active during encoding appear not to have much of an influence. For example, blending did not seem to be more likely to occur between two sequential faces. Despite this lack of temporal separation effects, one would certainly expect a blending mechanism to have some kind of window; one would not expect blending or mis-binding to occur between faces learned decades apart.

Second, the prototype mechanism appears to combine features according to an averaging of features or some other central tendency mechanism that would correspond to the mathematical blending performed on our images. This is in contrast to the binding mechanism proposed in the memory illusion literature in which features are combined from different faces. The difficulty in distinguishing between these two prototype mechanisms is that both versions predict high conjunction errors since the composite face is similar to two faces in memory. Thus it would be difficult to empirically distinguish between the two theories.

Whatever the method of prototype formation, it appears that such a mechanism is more likely to construct a prototype from stimuli that are similar than between those that are dissimilar. In all three experiments in the present article, a proportional prototype mechanism fit better than a single prototype mechanism, despite the fact that they have the same number of free parameters. This effect is above and beyond the effects of similarity in the model, since the model can already account for the fact similar morphs have higher false alarm rates than dissimilar morphs.

Implications for Memory Illusions

This work was motivated in part from findings in the memory illusion literature that demonstrate high errors to conjunction distractors. As we point out in the introduction, the existing memory illusion data are somewhat difficult to interpret, and some of the observed effects are explained by a familiarity-based model but not by a binding explanation. To support the contribution of a familiarity-based mechanism like GCM, we find strong evidence that similarity affects the conjunction error rates: similar morphs produce higher error rates than dissimilar morphs. Thus, memory illusion work needs to acknowledge the contributions of a familiarity mechanism to account for many of their effects.

Despite a contribution of similarity, a summed-similarity mechanism appears not to be the entire story, at least if naturalistic faces are used as stimuli. GCM could not provide good quantitative fits in simulations of extant memory illusion data nor our face recognition data. In our own data, familiarity-based models cannot simultaneously account for both the high morph false alarm rate and the good discriminability of distinctive faces. Even a model that can account for the data from both typical and distinctive faces (SimSample) could not account for the similar morphs without prototypes. In addition, our work was done with blended rather than conjunction prototypes, which have been suggested as providing stronger evidence for a blending mechanism (Schooler & Tanaka, 1991). Thus our findings can be taken as support for the conclusions derived from memory illusion data. There may indeed be some sort of binding or blending mechanism at work with the faces in memory that results in such high errors to conjunction stimuli.

Future Work

Inspection of the fits of the SimSample model to the data indicates that there are some deviations that suggest future model development. There is a fair amount of noise around the diagonal, which may be in part attributable to the noise of the MDS data. In addition, the data show a somewhat curvilinear shape, which is due to the fact that among old items, there is more variability in the data than the model can account for. This is perhaps due to aspects of distinctiveness that are not captured by the similarity ratings. For example, a feature such as a mole may not overly affect the similarity ratings between two otherwise quite similar faces. However, remembering this mole may greatly enhance future recognition of that face. Speeded sequential same-different judgments may provide better MDS spaces than similarity ratings for use with recognition memory data, and we are currently collecting such responses.

Clearly these deviations are sufficient to reject the model using a hypothesis-testing technique. For example, the RMSE of the SimSample model with proportional prototypes is 0.1411, while the average standard error of the hit rates was around 0.047. A complete model must include not only the similarity relations captured by the MDS space, but also non-perceptual effects such as those described by Levin (1996), which include the environment in which the faces are viewed, as well as social aspects such as familiarity with a particular race.

Despite these deviations, we believe that the SimSample model, with the addition of proportional prototypes, qualitatively accounts for many aspects of the data, and can quantitatively account for much of the individual hit rates. The SimSample model accounts for the data much better than previous models based on GCM, and does so with the same number of parameters. This model combines assumptions derived from models of categorization and recognition, and therefore has considerable empirical support in addition to the present successes. In addition, the model accounts for effects previously attributed to separate familiarity and recollective processes with a single sampling mechanism. Despite the fact that an exemplar-based version could not account for the false recognition effects seen with the morphs, these effects can be accounted for by a model extension that adds only one free parameter. These extensions support the idea that faces are blended together in memory, as long as they are similar to each other. We believe the success of the SimSample model demonstrates the utility of the face-space approach as applied to face recognition data, and suggests that this framework can be used to answer interesting questions about distinctiveness, typicality and blending.

Author Notes

The authors wish to thank Richard Shiffrin and Robert Nosofsky for many helpful discussions about the quantitative modeling. We would also like to thank Nelson Cowan and the two anonymous reviewers for their assistance in revising this article.

References

Ashby, F.G., Prinzmetal, W., Ivry, R., & Maddox, W.T. (1996). A formal theory of feature binding in object recognition. *Psychological Review, 103,* 165-192.

Bartlett, J., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition, 12,* 219-228.

Busey, T. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science, 9,* 476-482.

Busey, T. A, Tunnicliff, J., Loftus, G., & Loftus, E.(submitted). Accounts of the confidence-accuracy relation in recognition memory. Submitted to *Psychonomic Bulletin & Review*.

Byatt, G. & Rhodes, G. (in press). Recognition of own-race and other-race caricatures: Implications for models of face recognition. In press, *Vision Research.*

Franks, J. J., & Bransford, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology*, *90*, 65-74.

Gillund, G, & Shiffrin, R. (1984). A retrieval model for both recognition and recall. *Psychological Review, 92*, 1-38.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers, 26*, 381-386.

Hintzman, D. (1986). "Schema Abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Homa, D., Goldhardt, B., Burruel-Homa, L., & Smith, J.C. (1993). Influence of manipulated category knowledge on prototype classification and recognition. *Memory & Cognition*, *21*(4), 529-538.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Psychology: Human Learning and Memory, 7,* 418-439.

Kayser, A. (1985). Heads. New York: Abbeville Press.

Knapp, A., & Anderson, J.A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory & Cognition, 10*, 616-637.

Kroll, N., Knight, R., Metcalfe, J., Wolf, S., & Tulving, E. (1996). Cohesion failure as a source of memory illusions. *Journal of Memory and Language*, *35*, 176-196.

Levin, D. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 1364-1382.

Loftus, G.R. & Masson, M.E.J. (1994). Using confidence intervals in within-subjects designs. *Psychonomic Bulletin & Review, 1,* 476-490.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Metcalfe, J. (1990). Composite Holographic Associative Recall Model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119,* 145-160.

Neumann, P. G. (1977). Visual prototype formation with discontinuous representation of dimensions of variability. *Memory & Cognition, 5*, 187-197.

Nosofsky, R. M. & Zaki, S. R. (1998). Dissociations between Categorization and Recognition in Amnesics and Normals: An Exemplar-Based Interpretation. In press: Psychological Science.

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. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 700-708.

Nosofsky, R.M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3-27.

Nosofsky, R.M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43,* 25-53.

Pelli, D. G. & Zhang, L (1991). Accurate control of contrast on microcomputer displays. *Vision Research, 30,* 1033-1048.

Raaijmankers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93-134.

Reinitz, M., Lammers, W., & Cochran, B. (1992). Memory-conjunction errors: Miscombination of stored stimulus features can produce illusions of memory. *Memory & Cognition*, *20*(1), 1-11.

Reinitz, M., Morrissey, J. & Demb, J. (1994). Role of attention in face encoding. *Journal of Experimental Psychology: Learning, Memory & Cognition, 20,* 161-168.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika, 39,* 373-421.

Shin, H.J., & Nosofsky, R.M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General, 121*, 278-304.

Schooler, J. W., & Tanaka, J.W. (1991). Composites, compromises, and CHARM: What is the evidence for blend memory representations? *Journal of Experimental Psychology: General, 120,* 96-100.

Solso, R. L., & McCarthy, J. E. (1981). Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology, 72*, 499-503.

Tanaka, J., Giles, M., Kremen, S., & Simon, V. (Submitted). Mapping attractor fields in faces space: The atypicaly bias in face recognition.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14,* 107-141.

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.

Vokey, J. & Read, J (1992). Familiarity, memorability and the effect of typicality on the recognition of faces. *Memory & Cognition, 20* 291-302.

Tables

	Test Condition			
Study Condition	Old	Conjunction	Feature	New
Reinitz et al. (1994)				
Full Attention (control)	.67	.23	.09	.01
Divided Attention (deficit)	.40	.32	.19	.09
Kroll et al. (1996)				
Older Adults (control)	.97	.31	.06	.00
Right Hemisphere Lesion (deficit)	1.00	.58	.13	.01
GCM Predictions for Kroll et al. (Euclidean distances)				
Older Adults (control)	1.00	.25	.13	.03
Right Hemisphere Lesion (deficit)	1.00	.44	.26	.09

Table 1. Example Memory Conjunction Data and predictions from Nosofsky's Generalized Context Model (GCM, Nosofsky, 1986). Cells contain the probability of identifying a particular stimulus as an old face. The Reinitz et al. study used a forced-choice technique, which required subjects to pick 2 faces out of 8 candidate faces. Kroll et al. used an old/new recognition paradigm. GCM predictions used the following parameters: control: $c = 1.54$, $\beta = 4.00E+4$, $\theta = 18.14$; deficit: $c= 1.12$, $\beta = 1.01E+34$, $\theta = 154.6$. Versions using a city block metric produced only marginal increases in the fits.

False alarms for morphs	Similar	Dissimilar
Sequential	.614	.467
Separate	.622	.469

Hit Rates for Parents	Similar	Dissimilar
Sequential	.547	.675
Separate	.592	.647

Table 2. False alarm rates for morphs and hit rates for parent faces in Experiment 1. The common standard error of the mean estimated from a within-subjects ANOVA is 0.0256 (see Loftus and Masson, 1994).

Probability of choosing the parents over the morph	Similar	Dissimilar
Sequential	.471	.653
Separate	.452	.657

Table 3. Experiment 2 data. Probability of choosing the two parents over the morph in a forced-choice comparison. The common SEM = 0.0198.

Experiment 3a: False alarms for morphs	Similar	Dissimilar
Sequential	.710 (.019)	.473 (.027)
Separate	.690 (.022)	.475 (.031)

Experiment 3b: Hit rates for parents	Similar	Dissimilar
Sequential	.589 (.031)	.633 (.042)
Separate	.542 (.039)	.744 (.035)

Table 4. Experiment 3 means and standard errors for the false alarm rates to the morphs and hit rates to the parents. Individual SEMs for each condition are shown in parentheses.

Experiment 1												
Model	Parameters										RMSE	RSQ
GCM fit to parents,	c	w_1	w_2	w_3	w_4	w_5	w_6	β	θ			
morphs	2.6e-5	.24	.02	.15	.03	.00	.00	.75	100		.2060	.031
GCM	c	w_1	w_2	w_3	w_4	w_5	w_6	β	θ			
	6.68	.03	.18	.29	.42	.07	.01	1.16	2.79		.1587	.46
GCM- Identification	c	w_1	w_2	w_3	w_4	w_5	w_6	β	θ			
	0.18	.13	.13	.25	.13	.13	.25	457	6251		.1800	.30
SimSample	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD			
	3.50	.06	.12	.64	.07	.08	.02	.14	.07		.1462	.54
SimSample-	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD	pw		
Prototypes	3.91	.02	.11	.71	.07	.08	.02	.13	.04	.41	.1441	.56
SimSample- Prop.	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD	ρ		
Prototypes	3.69	.06	.14	.62	.07	.09	.02	.14	.06	2.77	.1411	.57

Experiment 2													
Model	Parameters											RMSE	RSQ
GCM	c	w_1	w_2	w_3	w_4	w_5	w_6	β	θ	ζ			
	7.92	.16	.01	.50	.00	.17	.16	22.64	7.44	2.02		.1312	.76
GCM- Identification	c	w_1	w_2	w_3	w_4	w_5	w_6	β	θ	ζ			
	2.30	.13	.07	.02	.45	.21	.12	48.88	10000	1.6		.1205	.73
SimSample	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD	ζ			
	3.48	.08	.10	.12	.55	.08	.08	.01	.19	6.9		.1120	.77
SimSample-	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD	ζ	pw		
Prototypes	2.83	.09	.14	.08	.44	.04	.20	.31	.05	4.63	.54	.1060	.80
SimSample- Prop.	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD	ζ	ρ		
Prototypes	4.35	.12	.14	.00	.31	.20	.23	.11	.55	4.36	5.32	.0971	.83

Experiment 3a												
Model	Parameters										RMSE	
GCM	c	w_1	w_2	w_3	w_4	w_5	w_6	β	θ			
	5.91	.03	.30	.19	.32	.16	.01	1.58	4.08		.1488	
GCM- Identification	c	w_1	w_2	w_3	w_4	w_5	w_6	β	θ			
	16.87	.16	.33	.09	.03	.28	.11	2.40	6.70		.1876	
SimSample	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD			
	3.53	.03	.12	.60	.09	.12	.03	.11	.03		.1427	
SimSample-	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD	pw		
Prototypes	3.82	.02	.12	.63	.08	.11	.03	.11	.02	.39	.1397	
SimSample- Prop.	c	w_1	w_2	w_3	w_4	w_5	w_6	Crit	CritSD	ρ		
Prototypes	3.59	.05	.17	.54	.13	.09	.02	.15	.10	2.81	.1391	

Table 5. Estimated Parameter values for model fits associated with graphed data for Experiments 1-3.

Figures

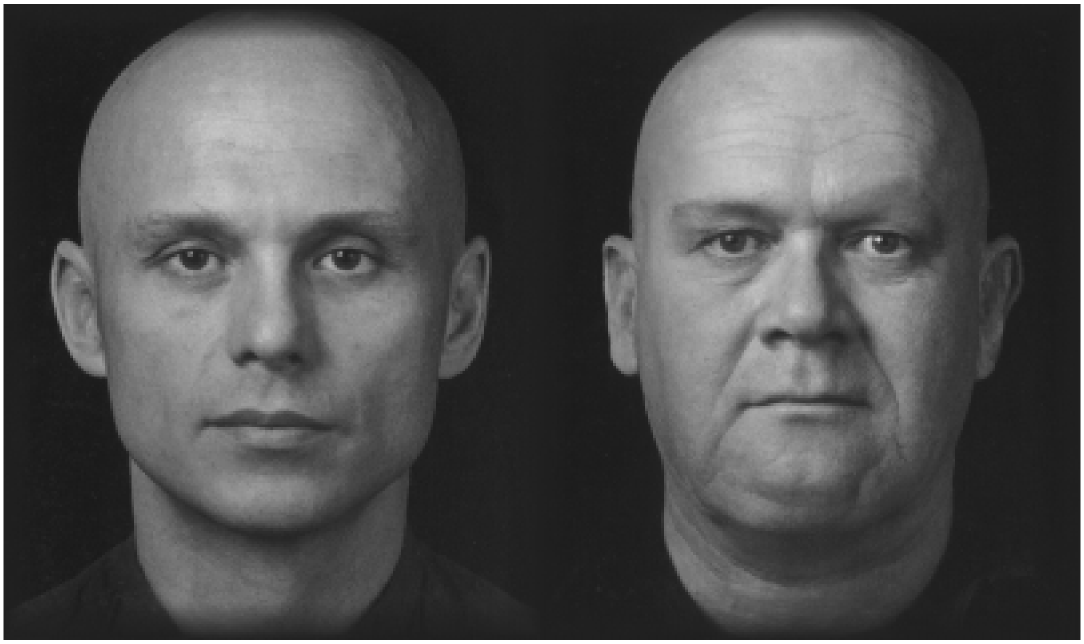


Figure 1. Example morphs constructed from similar (right) and dissimilar (left) parents.

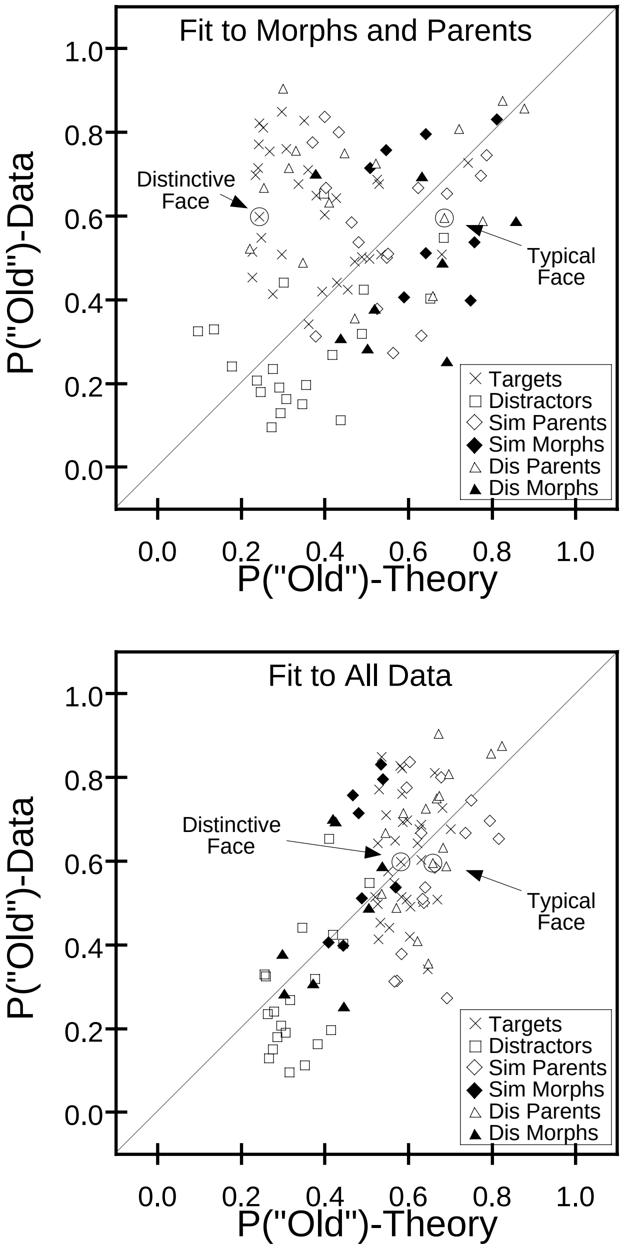


Figure 2. Fits of Nosofsky's Generalized Context Model (GCM) to the probability of saying old to targets and distractors for Experiment 1. Sim = Similar; Dis = Dissimilar. Top Panel: Fit of the model to just the Morph and Parent data. The diagonal represents perfect prediction. The model correctly places similar morphs above the similar parents. However, two faces (circled) with virtually identical hit rates in the data are placed far apart by the model as a result of the inability to account for the high hit rates to distinctive faces. Bottom Panel: Fit of the model to all data. Although the overall fit is better, the model now places the similar morphs below the similar parents, and typical faces are still predicted to be higher than distinctive faces, contrary to the similar and dissimilar parent data.

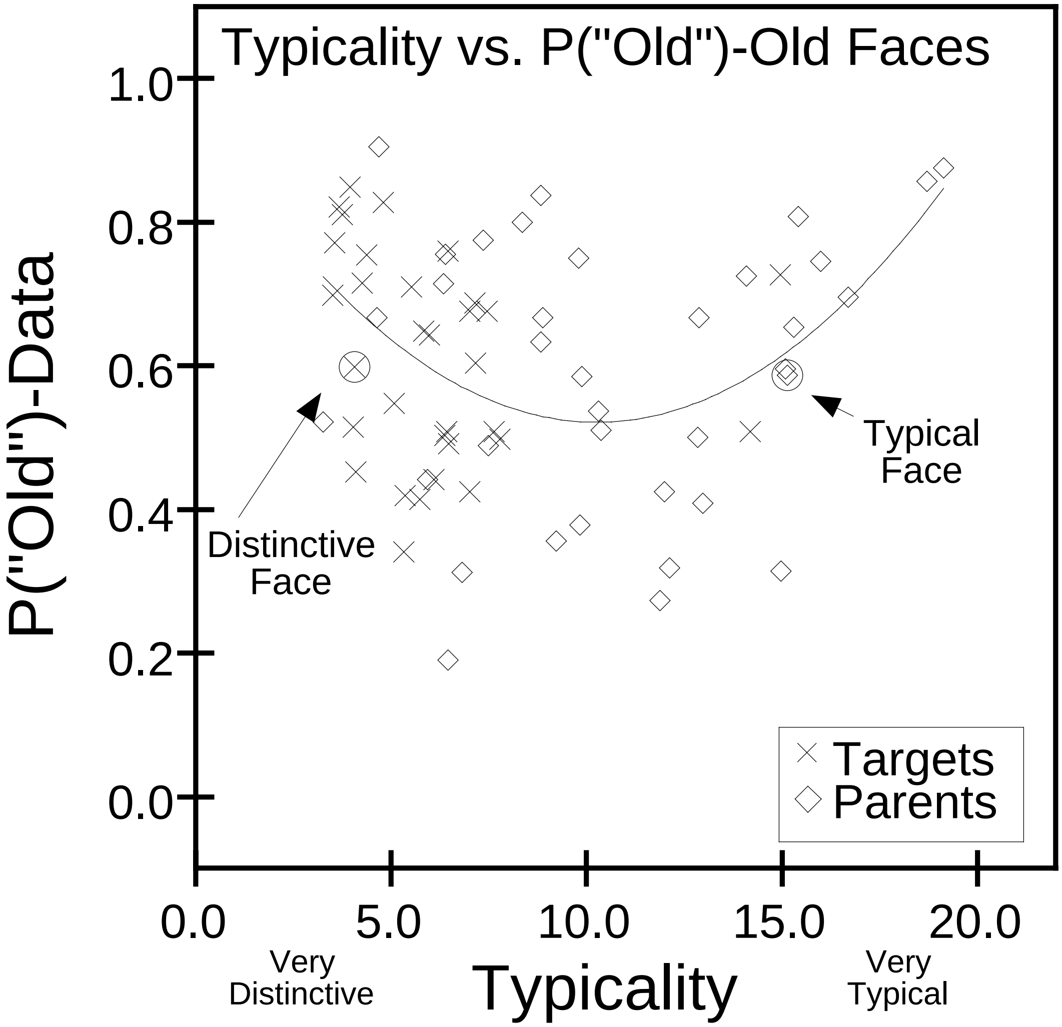


Figure 3. Computed typicality of each studied face compared with the probability of calling each face old in the data for Experiment 1, along with the fit of a second-order polynomial. The data form a curvilinear relationship, with very distinctive and very typical faces receiving high hit rates, while moderately typical faces receiving fewer old responses.

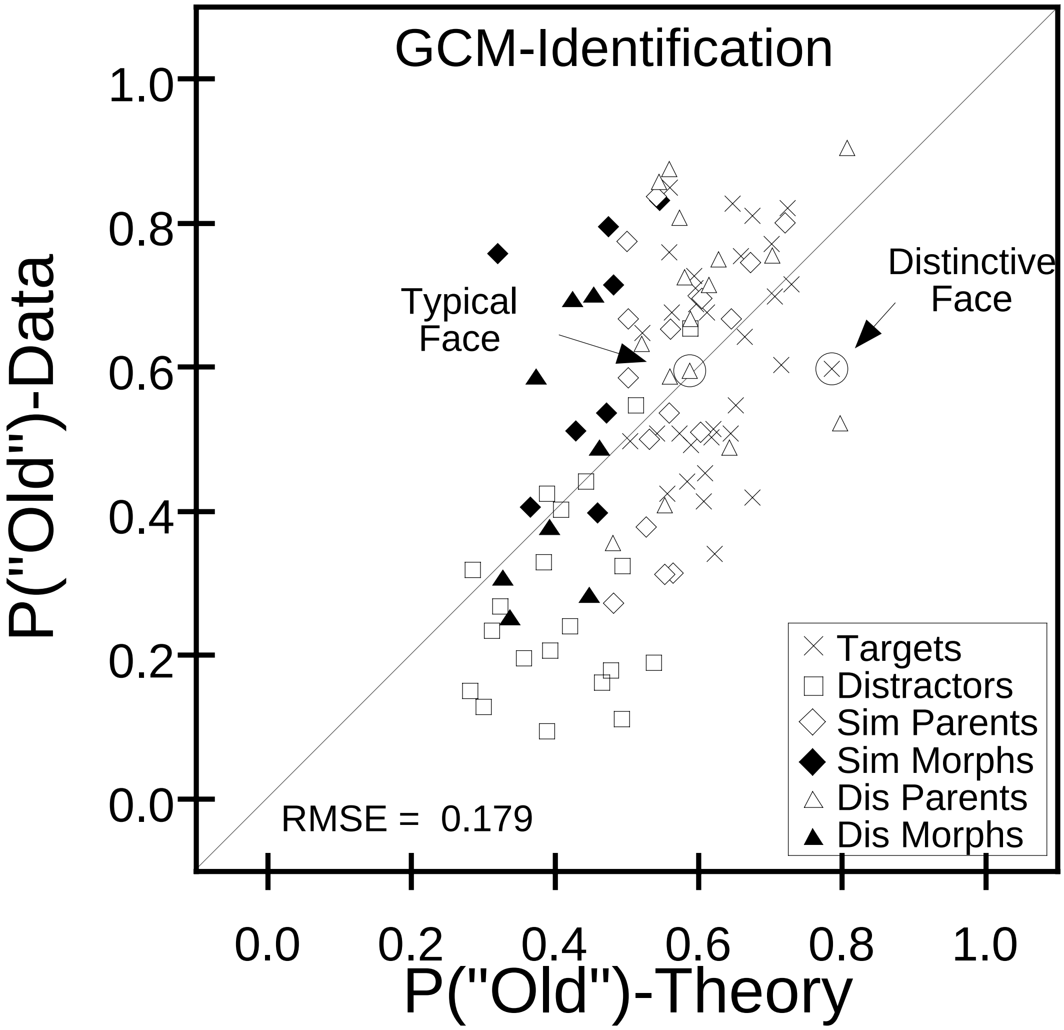
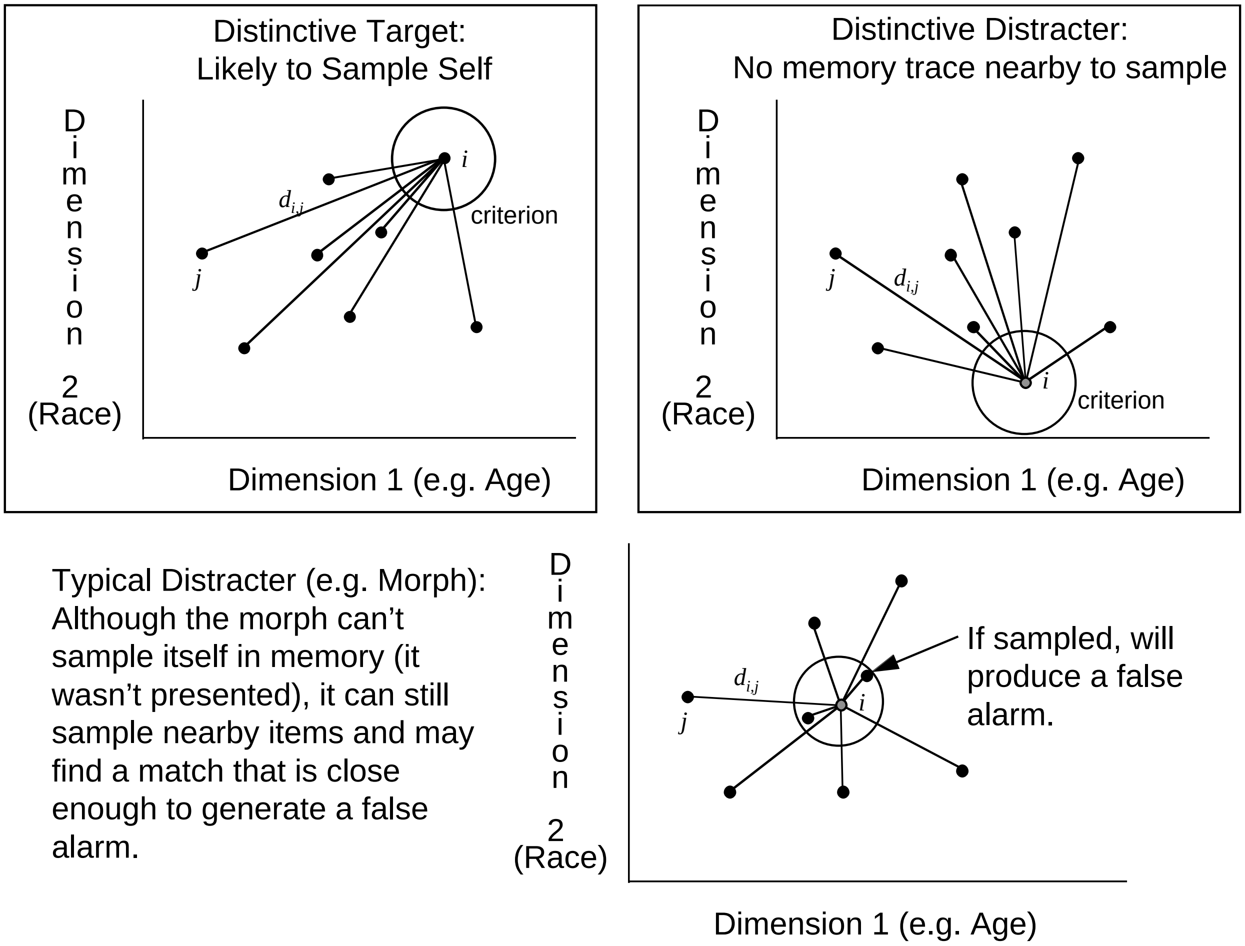


Figure 4. Fit of the Identification version of GCM the probability of saying old to targets and distractors for Experiment 1. This model does a good job of account for the high hit rates to distinctive targets and low false alarm rates to distinctive distractors. However, the model has difficulty account for the high false alarm rates to the very similar morphs, placing them below the parent faces (open diamonds). As with GCM, this model places the two faces with similar hit rates (circled faces) far apart, although they are reversed relative to Figure 2.



Distinctive Target: Likely to Sample Self

Distinctive Distracter: No memory trace nearby to sample

Typical Distracter (e.g. Morph): Although the morph can't sample itself in memory (it wasn't presented), it can still sample nearby items and may find a match that is close enough to generate a false alarm.

If sampled, will produce a false alarm.

Figure 5. Predictions of the SimSample model to Distinctive Targets, Distinctive Distractors, and Typical Distractors (morphs). Upper Left: A distinctive target is very likely to sample itself and thus has a high hit rate. Upper Right: A distinctive distractor cannot sample itself and may not have any nearby faces that could produce a false alarm if sampled. Bottom Panel: A very typical distractor 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.

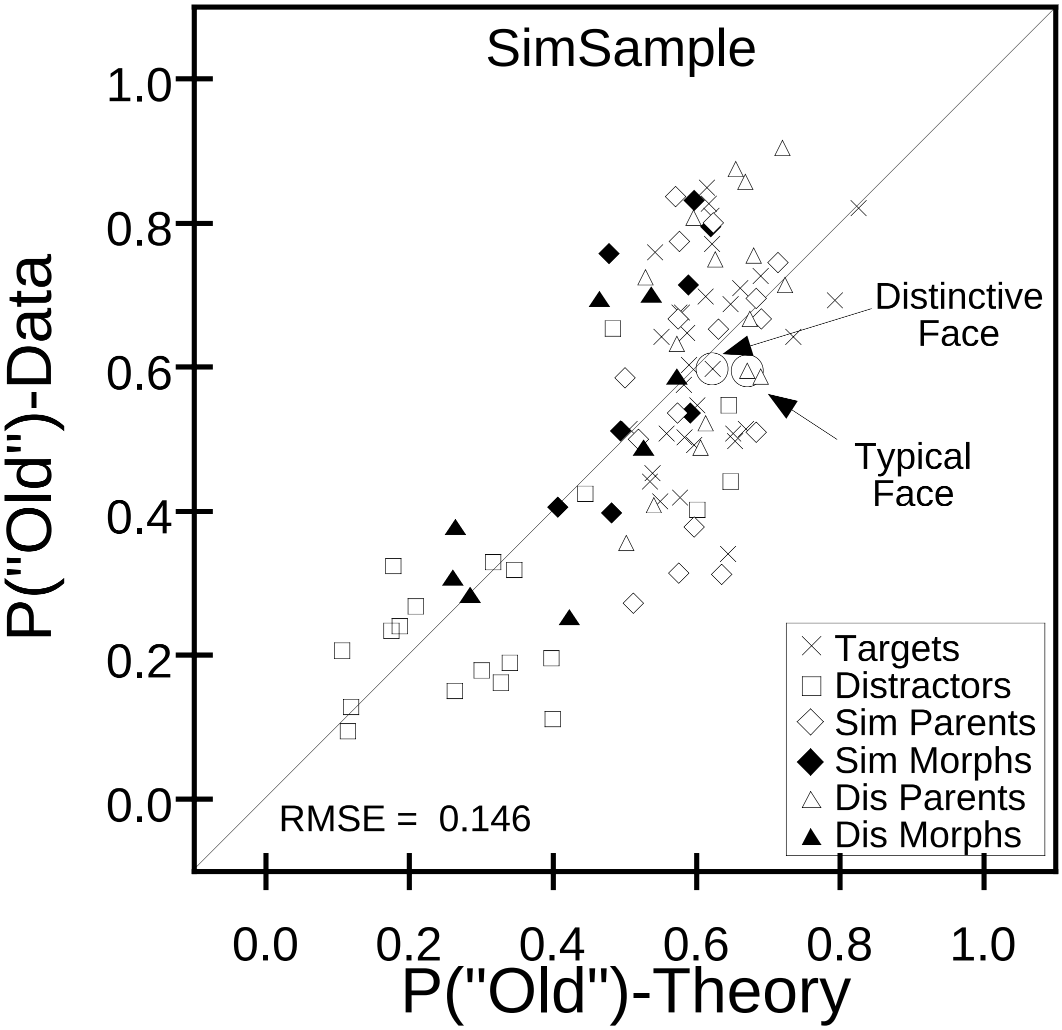


Figure 6. Fit of the SimSample model to the Experiment 1 data.

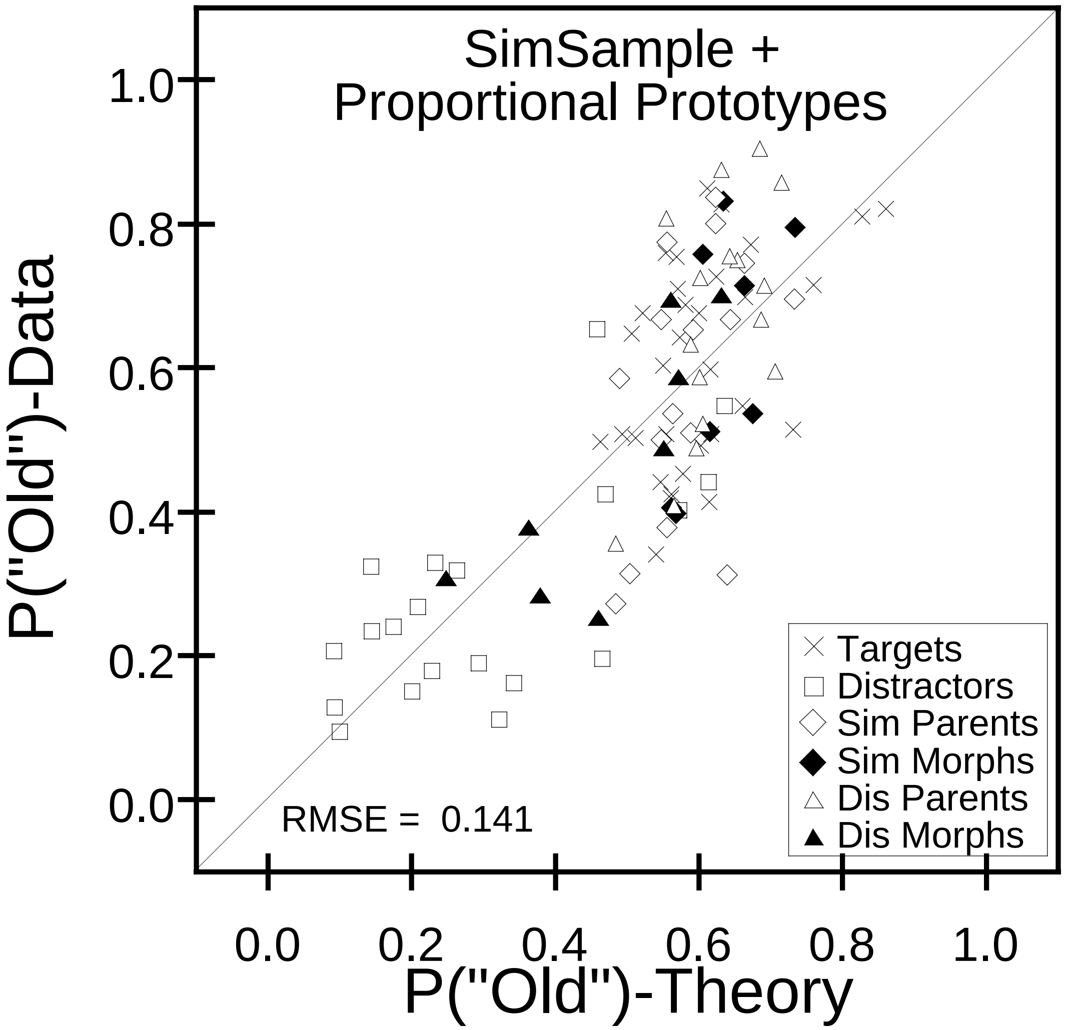
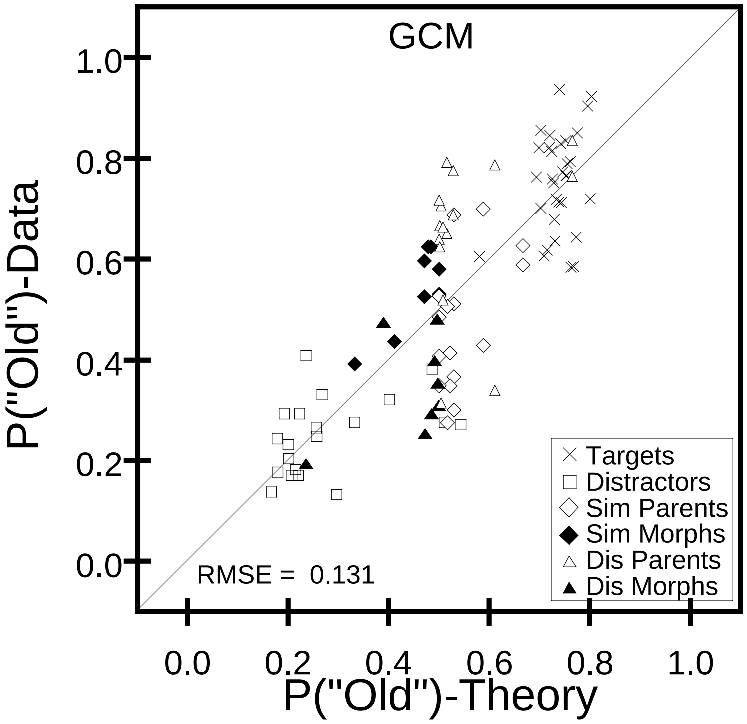
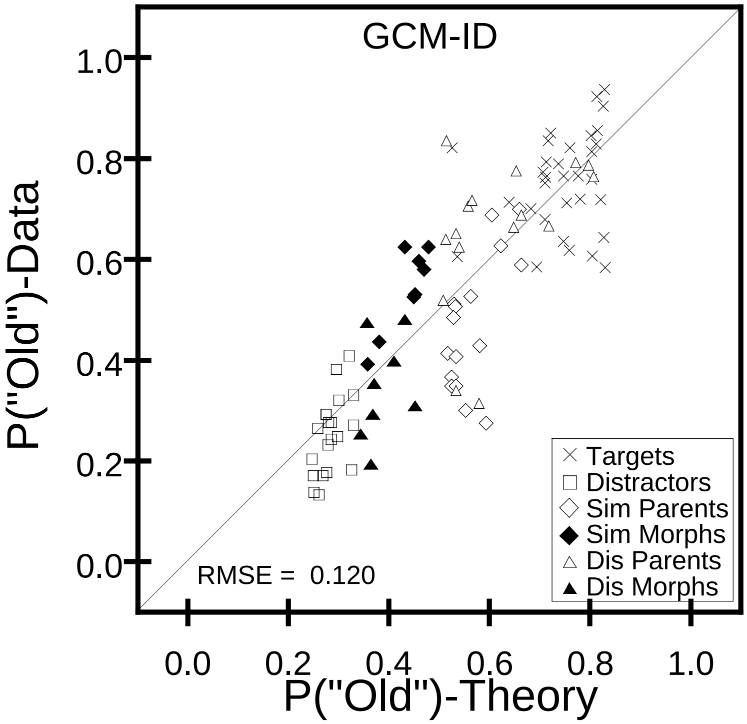


Figure 7. Fit of the SimSample model with the addition of prototypes that are proportional to the similarity of the parent faces.


GCM
P("Old")-Data
1.0
0.8
0.6
0.4
0.2
0.0
RMSE = 0.131
× Targets
□ Distractors
◇ Sim Parents
◆ Sim Morphs
△ Dis Parents
▲ Dis Morphs
0.0
0.2
0.4
0.6
0.8
1.0
P("Old")-Theory


GCM-ID
P("Old")-Data
1.0
0.8
0.6
0.4
0.2
0.0
RMSE = 0.120
× Targets
□ Distractors
◇ Sim Parents
◆ Sim Morphs
△ Dis Parents
▲ Dis Morphs
0.0
0.2
0.4
0.6
0.8
1.0
P("Old")-Theory

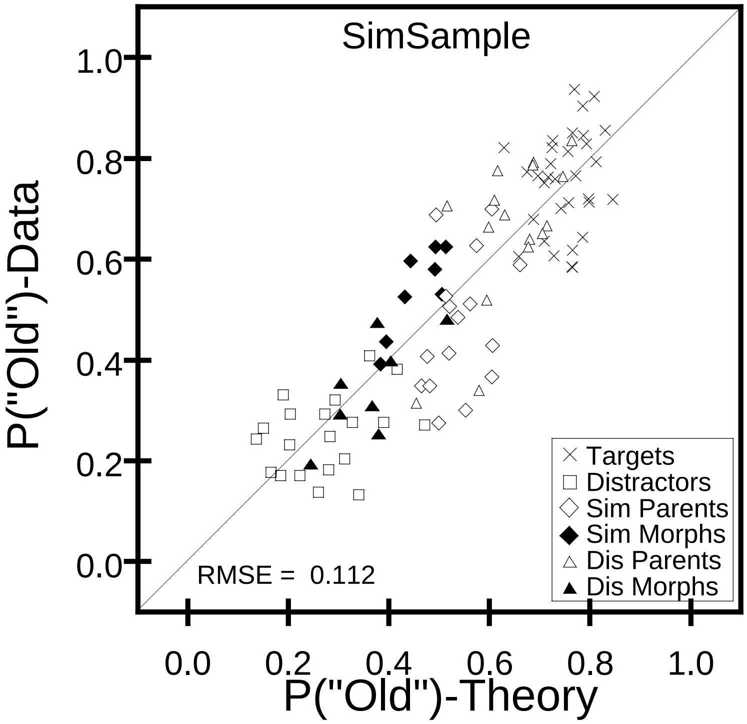


Figure 8. Fit of the three major models to the forced-choice data of Experiment 2.