Memory across a short-delay: Systematic biases in memory for faces

Jihyang Jun*
Deborah H. Tan*
Roger W. Remington
Yuhong V. Jiang

Department of Psychology
University of Minnesota

Highlights

• When selecting a face to match one in memory, people tend to choose a more extreme face.
• This happens with celebrity faces, emotional faces, but not unfamiliar faces.
• The short-term memory error is reminiscent of the caricature effect.
• Distinctive features are rapidly extracted to bias face memory and decision.

Abstract

From geometric figures to human faces, many visual stimuli vary along a continuum in featural space, anchored at one end by a highly distinctive constellation of features, at the other by a neutral set. Here we used a continuum of morphed faces to test whether errors in visual short-term memory are symmetric in feature space around the target or systematically biased toward one or the other end of the continuum. Participants were shown a face for one second. After a brief delay, participants were asked to choose the face they had been shown among three face options, which consisted of the target face, one face that was slightly more distinctive, and one face that was slightly more neutral. Continuums of morphed faces ranged from an average, neutral face to different distinctive celebrity faces in Experiment 1, and from neutral facial expressions to highly emotional expressions in Experiment 2. Results showed that when participants made an incorrect response, they were more likely to incorrectly identify the more distinctive face than the more neutral or average face as the target face. This bias toward more extreme faces, however, was not observed for unfamiliar (non-celebrity) faces that were emotionally neutral. These findings suggest that visual memory encodes distinctive features of stimuli that lead to biases in later recognition.

Keywords: Visual short-term memory, face memory, memory precision
Introduction

A visual object, such as an apple, contains features similar to the features of most other apples as well as distinctive features. If asked to hold the image of an apple in memory, will visual short-term memory (VSTM) encode mostly features similar to or distinct from other apples? The nature of the encoding could produce systematic errors in later recognition as visual memories lose fidelity. For example, when presented with choice stimuli that vary along a continuum of distinctiveness within an object category during a memory test, errors could be biased in one of two directions. Errors would be biased toward the typical members of the object category if people preferentially encode features representative of the category or if memory decay makes the unique features of the object less salient. Alternatively, errors could be biased away from typical members if people preferentially encode features unique to the object. In this study, we report a series of forced-choice judgments on morphed faces to investigate the nature of short-term encoding for stimuli that vary along a featural continuum.

Extensive evidence suggests that faces are represented in a multi-dimensional space (Valentine, 1991; Rhodes & Jeffery, 2006; Tsao & Freiwald, 2006). Similar faces cluster in neighboring locations in the face space. Some models suggest that faces are represented relative to a norm – an average face – with a given face represented as a set of deviations from the average face (Leopold, O’Tool, Vetter, & Blanz, 2001; Rhodes & Jeffery, 2006). Two faces with opposite deviations in the face space, such as a face wider than the average and one narrower than the average, are considered anti-faces. Other models are exemplar-based. These models do not assume the presence of a norm at the origin of the face space. Rather, each face takes on certain values along various dimensions of the face space (Valentine, 1991). Research using face adaptation provides evidence for the norm-based models. For example, adapting to an anti-face makes an average face look more like the original face (Leopold et al., 2001; Rhodes & Jeffery, 2006; Tsao & Freiwald, 2006). However, because both norm-based and exemplar models compute similarity in a multi-dimensional space, compelling evidence in support of one versus the other model is rare (Lewis & Johnson, 1999; Valentine, 1991). In either case, these models assume that people use features, or constellations of features, to compute the similarity of a face to other faces. Given limitations in memory, biases in terms of the facial features one chooses to encode could introduce different types of systematic error in later recognition. For example, if visual short-term memory primarily encodes features that distinguish one face from another then memory errors should be biased toward faces that exaggerate those features, like a caricature (Rhodes, Brennan, & Carey, 1987). In contrast, if memory primarily encodes facial features that are typical of the norm, or memory decay erases distinctive features, then errors should be biased toward a more neutral face (Frowd et al., 2012).

Rhodes and colleagues’ earlier work on the caricature effect suggests that features of a face are exaggerated in representation (Rhodes et al., 1987; see also Benson & Perrett, 1991, 1994; Rhodes, 1997). They found that people more readily recognize line drawings of a famous individual when the drawing caricatures the person than when it is veridical. The caricature effect, however, does not apply equally to all faces. Benson and Perret (1991) found smaller caricature effects for less familiar faces, and Rhodes and Moody (1990) failed to find a caricature
effect for unfamiliar faces. These results suggest that people may encode faces by exaggerating distinctive aspects of the face, except when the distinctive features are difficult to extract.

Other studies have found an analogous effect using a short-term memory task. Mauro and Kubovy (1992) uncovered asymmetry in people’s similarity judgment of sequentially presented faces. Participants made a “same-different” judgment for two computer-generated faces presented sequentially. Each stimulus was either a face or its caricature. This produced four combinations. Two of these – face-followed-by-face and caricature-followed-by-caricature – yielded a “same” response. The other two combinations yielded a “different” response. Mauro and Kubovy found that participants were faster reporting “different” when the caricature preceded the face, than when the face preceded the caricature. This finding suggests that the first stimulus, the one briefly held in memory, may be slightly caricatured. However, the finding may also reflect a general effect in categorization where a typical member is judged less similar to an atypical member (e.g., a face caricature), more than the other way around.

In fact, Op de Beeck and colleagues extended this type of asymmetry in categorization to account for similarity judgments. Op de Beeck, Wageman, and Vogels (2003) created novel shapes, where some were closer to the average of the set than others. People gave higher similarity ratings when an atypical shape preceded a typical one, compared to the other way around. This finding extends to computerized faces (Rensbergen & Op de Beeck, 2014). For example, an extreme face followed by an average face was judged more similar than an average face followed by an extreme face. This is the opposite of what one might expect if the first item is remembered as more extreme than it was. Thus, it remains unclear whether errors in short-term memory for faces are biased in one direction or the other.

To examine the nature of coding in visual short-term memory, we examined the pattern of errors when participants attempted to identify a previously presented face when it was presented among morphs along a continuum from an individual to an average face, or facial expressions morphed between an emotional face and a neutral face (Figure 1b). One morphed face was presented for encoding. After a brief delay, three choices were presented: the encoded (target) face, a foil closer to the average face, and a foil closer to the extreme. A bias would appear as an asymmetric pattern of errors for foils flanking the target along the morph continuum. If people encode distinctive features they should more frequently err in selecting the face morph closer to the extreme, whereas encoding of features common to all faces should produce inflated selection of the foil closer to the average face.

Our experiments tested three types of faces: celebrity faces, unfamiliar emotional faces, and unfamiliar neutral faces. The difficulty people have in recognizing unfamiliar faces suggest that distinctive features of unfamiliar faces are difficult to extract (Benson & Perrett, 1991; Rhodes & Moody, 1990). In addition, familiarity is known to warp the face space – people judge familiar faces to be less typical and more distinctive than their anti-faces (Faerber, Kaufmann, Leder, Martin, & Schweinberger, 2016). These findings raise the possibility that, much like the caricature effect, the direction of the bias toward one end of the featural continuum in short-term memory may differ for different types of faces.

Experiment 1
This experiment tested the nature of memory errors for celebrity faces. We took 8 male celebrity faces; for each stimulus, we created faces morphed between a given celebrity face (100%) and an average male Caucasian face (0%), in steps of 10%. Participants viewed a face (target) for 1s. After a 1s delay, they saw three choices – a face that matched the target (T), a foil closer to the average than the target (T-10%), and a foil closer to the original (T+10%) – and had to choose the target. If memory errors are random, then participants should be equally likely to choose either of the two foils. If memory errors are more norm-like, then choices should be biased toward the T-10% foil than the T+10% foil. Alternatively, if the distinctive features of a face are exaggerated in representation, then errors should be more biased toward the T+10% foil than the T-10% foil.

Experiment 1A tested the main hypothesis using morphs of celebrity faces. Experiment 1B tested memory for inverted faces. Given that face inversion impairs recognition (Rossion & Gauthier, 2002; Yin, 1969), any caricature-like effect should be reduced or eliminated with inverted faces. In Experiment 1C the encoding face was not shown to participants. Instead, we asked participants to guess which of the three choices the computer would have displayed. While this may come across as an odd task, it allows us to determine whether, in the absence of memory, people would be biased in choosing the more or less extreme option.

**Method**

**Sample size determination.** Sample size in each of Experiment 1A, 1B, and 1C was 16, a value chosen to be comparable to previous studies on VSTM and adequate to detect effects of memory load observed in previous studies (e.g., Luck & Vogel, 1997; estimated power was greater than .80 with a sample size of 16). However, the direction of memory error has not been investigated before. We therefore used the effect size obtained in Experiment 1 to determine sample size in subsequent experiments. Based on Experiment 1’s effect size, G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) estimated that 5 or more participants should be tested to obtain a power of .80. All subsequent studies tested 16 participants per experiment.

**Participants.** Participants in this study were students at the University of Minnesota between 18 and 35 years of age. All participants had normal or corrected-to-normal visual acuity, were naïve to the purpose of the study, provided informed consent, and received extra course credit for participation. Each participant was tested in only one experiment.

Forty-eight participants completed Experiment 1. Experiment 1A had 10 females and 6 males with a mean age of 19.4 years. Experiment 1B had 12 females and 4 males with a mean age of 20.4 years. Experiment 1C had 9 females and 7 males with a mean age of 20.6 years.

**Equipment.** Participants were tested individually in a room with normal interior lighting. The experiment was programmed using Psychtoolbox (Brainard, 1997; Pelli, 1997) implemented in MATLAB (www.mathworks.com). Stimuli were displayed on a 19” monitor with a spatial resolution of 1024 x 768 pixels and a vertical refresh rate of 75 Hz. Viewing distance was unconstrained at approximately 40cm. Visual angles estimated based on this distance were approximate.
**Stimuli.** We used 8 front-view Caucasian male celebrities (see also Jiang, Lee, Asaad, & Remington, 2016), including Adrien Brody, George Bush, Jim Carrey, Johnny Depp, Jimmy Kimmel, Daniel Radcliffe, Charlie Sheen, and Ben Stiller. Faces had a neutral or slightly happy expression, without glasses. The faces were converted into grayscale and uploaded into FaceGen Modeller (www.facegen.com), which removed hair and other external features. Each face was morphed with FaceGen’s average Caucasian male face (generated based on a dataset of 273 faces, see “FaceGen Frequently Asked Questions”, https://facegen.com/faq.htm). To create face morphs, the computer-aided morphing technique detects features in corresponding locations of two faces and blends the images together with different proportions of the two faces. At one end of the morphing spectrum was the average male face. At the other end was a celebrity’s face. For each celebrity face, the stimulus set contained morphs that differed by 10%, from 0% (the average face), to 10%, 20%, … and 100%. Figure 1a displays sample morphs for one of the celebrity faces.

![Sample stimuli used in Experiments 1, 2, and 4.](image)

![Sample stimuli used in Experiment 3 (stimuli were from Bediou et al., 2012).](image)

**Iterative procedure.** Our first attempt at addressing the research question was to examine how people’s choices migrated over time in the feature space. To this end the encoding faces changed from block to block using an iterative procedure (Griffiths, Christian, & Kalish, 2008). Specifically, after each target face was shown, three choices appeared: T, T+10%, and T-10%. Whichever participants chose as the one matching the target face became the target face in the next block of trials. Choosing the +10% face on a trial where the target was 40%, for instance, would lead to the presentation of a 50% face as the target in the subsequent block,
approximately 24 trials later. The iterative procedure was chosen to allow errors to multiply over time and the target face to “evolve” (Griffiths et al., 2008). However, the “evolution” was curtailed when the evolving target face hit the boundary of our stimulus space (e.g., the 0% or 100% face). Because our stimulus set did not contain antifaces and caricatured faces, we could not accurately assess the end result of the “evolution” (Appendix A). Therefore, we focused on the nature of memory errors. Subsequent experiments used fixed stimuli across blocks rather than the dynamic iterative procedure.

Figure 2. Trial sequence used in Experiment 1A. In this example, the target face was at the 60% morph level. Choice #3 on the test display was the exact match. Choice #2 was a face toward the norm (50% morph level). Choice #1 was a face away from the norm (70% morph level).

Experiment 1A. This experiment displayed upright faces. After 8 trials of practice, participants completed 15 blocks, each consisting of 24 trials. In Block 1, each of the 8 celebrity identities was presented once at each of three morph levels for encoding: 40%, 60%, and 80%. These 24 trials were presented in a random order. On each trial, one face (approximately 6.4°x10.2°) was presented at the center of the screen for 1s and disappeared. After a delay of 1s, three choices were presented at equidistant locations on an imaginary circle, corresponding to the 4 o’clock, 8 o’clock, and 12 o’clock positions. The center of each choice was approximately 11.2° from the display center. The choices included the target face (T), a face 10% more toward the original (T+10%), and a face 10% more toward the average face (T-10%). The position of the three faces was randomized from trial-to-trial. The numbers 1, 2, and 3 were presented at the bottom of the faces: number 1 at the 4 o’clock position, 2 at the 8 o’clock position, and 3 at the 12 o’clock position, in all trials. Participants pressed the number corresponding to the face that matched the target (Figure 2). The choices were presented until participants made a response, after which three rising tones indicated the completion of the trial. The tones were the same regardless of whether the response was correct. The next trial commenced after 1s. At the completion of the block, the computer informed participants of the number of trials that they got correct in that block and encouraged them to be accurate.
The choice that participants made for each of the 24 faces displayed in Block 1 became the target face for Block 2. The choices were again one face that matched the new target and two faces that were +10% or -10% relative to the new target. One exception was made. If the target face was a 100% face, the choices did not exceed 100%. Instead, the three choices contained a 90% face, and two 100% faces. Conversely, if the target face was a 0% face, the choices did not drop below 0%. The three choices contained a 10% face and two 0% faces. These trials were flagged out and removed from subsequent data analysis. The choices people made for Block 2 became the target face for Block 3, and so on, until the completion of all 15 blocks.

Participants were not informed of the fact that the faces were modified from famous faces, nor were they informed about how their choices would influence what they would see in subsequent blocks. At the completion of the experiment, we assessed recognition of the celebrity faces. Participants could recognize many of the un-edited celebrity faces, but spontaneous recognition of the morphs was infrequent (Appendix B).

**Experiment 1B**. Experiment 1B was identical to Experiment 1A, except that all faces, including the encoding face and the three choices on each trial, were presented in an inverted orientation.

**Experiment 1C**. Experiment 1C was identical to Experiment 1 except that a gray oval was presented for 1 sec instead of the target face. After a 1s delay, participants viewed three choices that differed in morph levels in steps of 10%: T, T-10%, and T+10%. Participants were told that the computer had a face to show, but instead displayed an oval. Their task was to choose the face that they thought the computer would have displayed. The procedure for adjusting the (hidden) faces was the same as in Experiment 1A. Specifically, the computer selected a target face for encoding (e.g., 40% face A), and later displayed three choices (e.g., 30%, 40%, and 50% face A). Participants chose one of them, and their choice became the hidden target face used by the computer in the next block. The only difference between Experiments 1A and 1C was that whereas the chosen target face was displayed for encoding in Experiment 1A, it was covered up by an oval so participants did not see it in Experiment 1C.

Each of Experiments 1A, 1B, and 1C took approximately 45 minutes to complete.

**Effect sizes**. Whenever possible, we report effect size along with statistical significance. In ANOVA, Cohen’s $\hat{f}^2$ of 0.02, 0.15, and 0.35 correspond to small, medium, and large effects. In t-tests, Cohen’s $d$ of 0.2, 0.5, and 0.8 correspond to small, medium, and large effects.

**Results**

Even though the encoding faces in Block 1 covered just three morph levels, the iterative procedure led to the presentation of nearly all morph levels in later blocks. Table 1 displays the mean number of trials (out of 360) for which a specific morph level was encoded.

In the following analyses, we excluded target faces at 0% or 100% (the choices could not contain both +10% and -10% foils). We also excluded faces at 10% or 20% morph levels because the average number of trials was typically less than 10, the arbitrary cutoff used to ensure data reliability. The pattern of the results was the same when 20% morphs were included. Figure 3 shows the results.
Table 1. Mean number of trials for which a specific morph was the target face, out of 360 trials in total. Standard error (S.E.) of the mean across participants is shown in parenthesis.

<table>
<thead>
<tr>
<th>Morph</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp.1A</td>
<td>3 (1)</td>
<td>3 (1)</td>
<td>9 (1)</td>
<td>16 (2)</td>
<td>35 (3)</td>
<td>35 (5)</td>
<td>53 (3)</td>
<td>61 (3)</td>
<td>73 (4)</td>
<td>44 (3)</td>
<td>30 (3)</td>
</tr>
<tr>
<td>Exp.1B</td>
<td>4 (1)</td>
<td>9 (1)</td>
<td>18 (2)</td>
<td>29 (3)</td>
<td>47 (3)</td>
<td>43 (3)</td>
<td>53 (2)</td>
<td>50 (2)</td>
<td>50 (2)</td>
<td>34 (3)</td>
<td>22 (3)</td>
</tr>
<tr>
<td>Exp.1C</td>
<td>7 (1)</td>
<td>14 (1)</td>
<td>32 (8)</td>
<td>37 (4)</td>
<td>49 (3)</td>
<td>44 (3)</td>
<td>45 (3)</td>
<td>38 (4)</td>
<td>42 (3)</td>
<td>29 (2)</td>
<td>23 (3)</td>
</tr>
</tbody>
</table>

1. Proportion of correct responses

Accuracy was above chance for upright faces (42% in Experiment 1A), $t(15) = 5.91, p < .001$, Cohen’s $d = 1.48$, above chance for inverted faces (36% in Experiment 1B), $t(15) = 3.88, p = .001$, Cohen’s $d = 0.97$, and marginally below chance for the guessing task (31% in Experiment 1C), $t(15) = 1.91, p = .075$, Cohen’s $d = 0.48$. The use of highly similar probes achieved what we intended – an induction of high error rates. It is to the nature of errors that we turn to next.

2. Nature of errors

The presence of a systematic bias can be assessed by testing whether people made more errors in the +10% or the -10% direction. For each experimental version, we ran a repeated measures ANOVA on error type (+10% or -10%) and target morph level. Note in Experiment 1A data were missing from one participant at the 30% morph level, hence the analysis was performed on 15 participants.
Figure 3. Results from Experiment 1. The gray area in each data figure highlights where the more extreme option was chosen more often than the less extreme option. Error bars show ± 1 S.E. of the mean. (a) Proportion of the three types of responses in Experiment 1A. (b) Proportion of the three types of responses in Experiment 1B. (c) A schematic illustration of the procedure used in Experiment 1C. (d) Proportion of the three types of responses in Experiment 1C.

With upright celebrity faces (Experiment 1A), participants made more +10% errors than -10% errors (Figure 3a), yielding a significant main effect of error type, $F(1, 14) = 26.67, p < .001$, Cohen’s $f^2 = 1.91$. Accuracy differed for target faces of different morph levels, $F(6, 84) = 2.82, p = .015$, Cohen’s $f^2 = 0.20$ for the main effect of morph level. Trend analysis showed a significant linear trend of morph level, $F(1, 14) = 7.89, p = .014$, Cohen’s $f^2 = 0.56$, suggesting that memory accuracy declined with increasing target morph level; the quadratic and other higher-order trends were not significant, smallest $p = .13$. [footnote1]. The degree of bias toward the T+10% choice was not identical across morph levels, $F(6, 84) = 8.66, p < .001$, Cohen’s $f^2 = 0.62$, for the interaction between error type and morph level. Trend analysis on the interaction term showed a significant linear trend, $F(1, 14) = 7.35, p = .017$, Cohen’s $f^2 = 0.52$, and a significant quadratic trend, $F(1, 14) = 18.88, p = .001$, Cohen’s $f^2 = 1.35$, indicating that the bias toward the T+10% choice was stronger for target faces at intermediate morph levels than those at the high or low levels.

---

1 Owing to the relatively small number of trials at each morph level, we performed a polynomial trend analysis rather than pairwise comparisons, to reduce the number of multiple comparisons.
When faces were presented in an inverted orientation (Experiment 1B), there was no longer any bias toward either type of foil (Figure 3b). Neither the main effect of error type, $F < 1,$ nor the main effect of morph level was significant, $F(6, 90) = 1.28, p = .273, \text{Cohen's } f^2 = 0.09,$ and the two factors did not interact, $F(6, 90) = 1.50, p = .188, \text{Cohen's } f^2 = 0.10.$ A cross-experiment comparison [footnote 2] between Experiments 1A and 1B showed a significant interaction between face orientation and error type, $F(1, 29) = 16.27, p < .001, \text{Cohen's } f^2 = 0.56.$ Upright faces, but not inverted faces, were associated with a positive memory bias.

To what degree was the positive bias a memory error rather than a decision bias? When confronted with three choices, T-10%, T, and T+10%, might people simply be biased in choosing the most extreme option, the one farthest from the average face? If so, they should show the same pattern when making a choice in the absence of memory. Results in Experiment 1C did not support this possibility. Here, participants made significantly fewer choices of T+10% than T-10%, $F(1, 15) = 7.46, p = .016, \text{Cohen's } f^2 = 0.49$ for the main effect of error type (Figure 3d). The main effect of target morph level, $F < 1,$ and the interaction between error type and morph level, $F(6, 90) = 1.19, p = .317, \text{Cohen's } f^2 = 0.08,$ were not significant. This pattern of choice bias was the opposite of what we found in Experiment 1A. Direct comparisons between the two experiments showed a significant interaction between experimental version (1A vs. 1C) and error type, $F(1, 29) = 26.28, p < .001, \text{Cohen's } f^2 = 0.91.$

Discussion

Using morphs created from celebrity faces, Experiment 1 shows that memory errors for faces in a short-term memory task are biased away from the average face. Memory errors in the direction of +10% foils are more likely than memory errors in the direction of -10% foils. This bias was specific to upright faces and was absent when faces were presented in an inverted orientation. The lack of a memory bias for inverted faces is anticipated by norm-based models of face representation. Though norms exist for inverted faces, deviation from the norm is more readily extracted from upright than from inverted faces (Leopold et al., 2001). By this account, memory bias is related to the efficiency with which people extract featural deviations from the norm. Because the same stimuli were used for upright and inverted faces, the findings also argue against an account based on low-level differences in stimuli. For example, one might argue that even though the morphing procedure divided the stimuli into steps of 10%, it may not have done a good job and instead created stimuli that were not linear. Morph level T% may have been more similar to morph level T+10% than to morph level T-10%. If such low-level differences existed, they should have been maintained when faces were inverted. Because results changed when faces were inverted, the bias toward T+10% for upright faces is likely a psychological effect rather than a stimulus effect.

Experiment 1 also showed that when choices were not based on memory, people were more likely to choose an option closer to the norm (T-10%) rather than one farther from the

---

2 Note that this is a between-experiment comparison. The interaction between experiment and error type should be further validated in future studies that use a between-subject, rather than between-experiment, comparison.
norm (T+10%). Although the task in Experiment 1C may seem somewhat odd, participants did spend time pondering their response. Response time in Experiment 1C – approximately 2s – was similar to that of Experiment 1B (2.1s; RT was not recorded in Experiment 1A). The positive bias shown in Experiment 1A thus reflected a bias in memory rather than a choice bias. The opposite bias shown in Experiment 1C may reflect a tendency to choose faces that are more attractive, i.e., those closer to the average face (Langlois & Roggman, 1990). Experiment 1A, along with the two control versions, provide strong evidence that people rapidly extract featural deviations from morphs of celebrity faces. This representation biases subsequent memory and decisions in a short-term memory task.

**Experiment 2**

This experiment aimed to replicate the findings with celebrity faces from Experiment 1A using a more standard discrimination task than the adaptive target paradigm of Experiment 1. Here we adopted a standard experimental design in which a fixed set of stimuli was presented in all blocks. All stimuli, from 10% to 90% morphs, were presented equally often as the target face; their presentation was unrelated to participants’ choices.

**Method**

**Participants.** Sixteen participants, including 15 females and 1 male with a mean age of 21.9 years, completed Experiment 2.

**Design and procedure.** Similar to Experiment 1, participants completed 360 trials, divided equally and randomly into 9 target morph levels (10%, 20%, 30%, … 90%) and 8 face identities. The task stopped every 24 trials to give participants a break and to provide feedback on their accuracy. Unlike Experiment 1, this experiment did not use an iterative procedure – participants’ choices had no bearing on what they would see in future blocks. Other aspects of the experiment were the same as in Experiment 1A.

**Results and Discussion**

Results from this experiment (Figure 4) replicated those of Experiment 1A. Accuracy (43%) was above chance, t(15) = 7.96, p < .001, Cohen’s d = 1.99, comparable to the level observed in Experiment 1A, t(30) = 0.70, p = .491, Cohen’s d = 0.25.

Similar to Experiment 1A, participants were more likely to commit T+10% than T-10% errors. In fact, when the analysis was restricted to morph levels 30%-90%, the range used in Experiment 1, the statistical results showed the same pattern as those found in Experiment 1. Here we report statistical results that include the entire range of morph levels.
A repeated measures ANOVA on error type and target morph level revealed a significant main effect of error type, $F(1, 15) = 14.85, p = .002,$ Cohen’s $f = 0.99$. Accuracy varied across target morph levels, $F(8, 120) = 17.16, p < .001,$ Cohen’s $f = 1.15$. Trend analysis on morph level showed a significant linear trend, $F(1, 15) = 20.45, p < .001,$ Cohen’s $f = 1.36$, and a significant quadratic trend, $F(1, 15) = 26.64, p < .001,$ Cohen’s $f = 1.78$. Numerically, accuracy was poor at the 10% morph level; it then rose for the 20% and 30% morph levels, before declining at higher levels. The interaction between error type and morph level was significant, $F(8, 120) = 2.51, p = .015,$ Cohen’s $f = 0.17$. As seen in Figure 4, the bias toward T+10% was stronger at intermediate morph levels than at low or high levels.

The first two experiments used the same 8 celebrity faces as stimuli. This allowed us to examine item effects, treating the 8 celebrities as another factor in the analysis. The item-level analysis showed that the observed memory bias was not influenced by individual items, in that there was no interaction between error type and items, $F(7, 105) = 1.64, p = .132,$ Cohen’s $f = 0.11$ in Experiment 1A; $F(7, 105) = 1.63, p = .134,$ Cohen’s $f = 0.11$ in Experiment 2.

Experiment 2 successfully replicated the main finding of Experiment 1. These two experiments suggest that features of a celebrity face are exaggerated after a short memory interval.

Experiment 3
Experiment 3 attempted to extend findings from the first two experiments to a different underlying continuum, emotional expression. In place of the individual celebrity face morphs, Experiment 3 used face morphs created by Bediou et al. (2012), which contained two male and two female individuals in five expressions: neutral, happy, fearful, sad, and angry. For each individual, the face morphs were created by morphing the neutral expression with that individual’s face in full-blown expression (e.g., 100% happy; Figure 1b). We tested whether memory for facial expressions showed the same bias as memory for celebrity face identities.

Method

Participants. Thirty-two participants completed Experiment 3. There were two versions: upright (Experiment 3A) and inverted (Experiment 3B). Experiment 3A had 14 females and 2 males with a mean age of 19.0 years. Experiment 3B had 4 females and 10 males with a mean age of 20.0 years.

Stimuli. Bediou et al. (2012) generously provided their face morphs. The stimulus set contained 4 unfamiliar individuals – 2 females and 2 males – with a front view. Each individual depicted 5 expressions: neutral, happy, angry, sad, and fearful. Bediou et al. morphed each emotional face (e.g., 100% happy) with the neutral face of the same individual to create a range of faces that differed in the intensity of an emotion (Figure 1b). All faces were in grayscale, without hair and other external features. Each face subtended approximately 8.5° x 11.2° presented against a black background. Informal observations suggest that it was difficult to identify the expression of morph levels below 50%. Therefore, in Experiment 3 we restricted the target face’s morph levels to 60%, 70%, 80%, and 90%.

Experiment 3A. In Experiment 3A faces were presented in an upright orientation. The design and procedure were similar to Experiment 2, except that a new set of stimuli were used. The target face could be one of 4 expressions (happy, angry, sad, and fearful) at one of 4 morph levels (60%, 70%, 80%, and 90%). The choice stimuli contained 3 options: T, T-10% in emotional strength, and T+10% in emotional strength. Each participant completed 384 trials, divided equally and randomly into four expressions, four target morph levels, and four individuals.

Experiment 3B. Experiment 3B was identical to Experiment 3A except that faces were presented in an inverted orientation (rotated 180°).

Results
Memory accuracy in Experiment 3 was low. It was marginally above chance for upright faces (34.8% in Experiment 3A), \( t(15) = 1.83, p = .087, \) Cohen’s \( d = 0.46, \) and not different from chance for inverted faces (34.7% in Experiment 3B), \( t(15) = 1.64, p = .122, \) Cohen’s \( d = 0.41. \) The low accuracy raised questions about whether participants simply made random guesses. If so, memory errors should be symmetrical around the target. As shown next, this was not the case.

Participants in Experiment 3A more often erred in the +10% direction than the -10% direction (Figure 5). A repeated measures ANOVA on error type, type of facial expression, and target morph level revealed a significant main effect of error type, \( F(1, 15) = 32.62, p < .001, \) Cohen’s \( f^2 = 2.17. \) Neither the type of facial expression, \( F < 1, \) nor target morph level, \( F < 1, \) had an effect. The interaction between the type of facial expression and error type was marginally significant, \( F(3, 45) = 2.61, p = .063, \) Cohen’s \( f^2 = 0.17, \) but none of the other interactions were significant, all \( ps > .05. \)

To examine the generality of the findings across the four types of expression, we computed an index of error direction by taking the difference between the proportion of T+10% errors and T-10% errors (Figure 6a), separately for each facial expression. All four types of facial
expression showed more T+10% errors than T-10% errors, ps < .01. The effect declined moderately from happy (0.19), fearful (0.15), angry (0.13), to sad (0.09). To understand the order of decline, we asked another group of 16 participants to classify the facial expression of the stimuli used in Experiment 3A. These participants were shown each expression at each of the 4 morph levels, and pressed a button to indicate which expression was shown. There were 512 trials. Accuracy for expression classification declined from happy (99%), fearful (90%), angry (89%), to sad (85%). The order showed a remarkable correspondence with the size of positive memory bias in Experiment 3A. Expressions that people are more sensitive to are associated with greater memory biases in the +10% direction.

When faces were inverted in Experiment 3B, we no longer observed a positive bias (Figures 5b, 6b). The main effect of error type was not significant, \( F(1, 15) = 2.08, p = .170 \), Cohen’s \( f^2 = 0.14 \). Neither the type of facial expression, nor target morph level influenced the results, \( F_s < 1 \). None of the interactions were significant, smallest \( p = .165 \). Thus, inverted emotional faces showed no systematic memory biases.

The difference between upright and inverted emotional faces was supported by a cross-experiment comparison in an ANOVA using orientation as a between-subject factor and error type as a within-subject factor. Orientation and error type interacted significantly, \( F(1, 30) = 18.27, p < .001 \), Cohen’s \( f^2 = 0.61 \).

Discussion

Experiment 3 extended the findings from celebrity faces to emotional faces. We showed that memory for facial expression was also biased toward more extreme options, namely, faces with a stronger expression than what people encoded. This bias applied to all four expressions tested here. It was strongest for happy and weakest for sad. When another group of participants categorized the morphs into the four expression types, they were most accurate with happy and least accurate with sad. The correspondence between memory bias and categorization accuracy suggests that expressions people are more sensitive to are also ones that show stronger positive memory biases. Similar to Experiment 1, the positive memory bias disappeared when faces were inverted. These findings suggest that the memory bias may be related to how easy it is to extract distinctive features of a face.

Experiment 4

Both familiar faces and emotional faces are more memorable than neutral, unfamiliar faces (Franklin & Adams, 2010; Megreya & Burton, 2006, 2008). This raises questions about whether the positive memory bias shown so far generalizes to neutral, unfamiliar faces. That is, the nature of the encoding into VSTM may depend on the distinctiveness of the stimulus, which is enhanced for celebrity and emotional faces. Given that the positive memory bias is stronger for stimuli that people are more sensitive to (e.g., happy rather than sad faces), it is possible that the positive memory bias would weaken for faces that are less memorable.
Previous studies suggest that familiarity mediates encoding in face processing. Buttle and Raymond (2003) found superior change detection performance for celebrity faces, but not for faces recently learned, relative to unfamiliar faces. In fact, recognition errors for unfamiliar faces are surprisingly high (Megreya & Burton, 2008). Other differences between familiar and unfamiliar faces are also notable. The caricature effect is stronger for familiar faces, and in some studies, it is absent for unfamiliar faces (Hagen & Perkins, 1983; Rhodes & Moody, 1990). In addition, Halberstadt, Pecher, Zeelenberg, Ip Wai, and Winkielman (2013) showed that whereas morphing with an average face increased the attractiveness rating of an unfamiliar face, the same procedure reduced the attractiveness rating of a celebrity face. Faerber et al. (2016) further demonstrated that familiarity distorts the norm-based face space – a famous face is evaluated as more distinctive and trustworthy than its antiface and unfamiliar faces. These findings raise the possibility that the positive memory bias observed so far may be less pronounced with unfamiliar, neutral faces.

Method

Participants. Sixteen participants, including 11 females and 5 males with a mean age of 20.3 years, completed Experiment 4.

Stimuli. The face stimuli consisted of 16 Caucasian males’ front-view face photographs: 8 celebrities and 8 unfamiliar faces. The 8 celebrities’ images were the same as in Experiments 1 and 2. The 8 unfamiliar faces were obtained from Glasgow Unfamiliar Face Database (GUFD; Burton, White, & McNeill, 2010) or from images found on the internet. To ensure consistency in stimulus creation, we ran FaceGen Modeller on all 16 photographs in a single batch, in which each individual face was morphed with FaceGen’s average Caucasian male face in steps of 10%.

Design and Procedure. This experiment was similar to Experiment 2 with the following exceptions. First, the target face could be based on one of 16 individuals, 8 celebrity and 8 unfamiliar faces, and each target face could be at one of six morph levels ranging from 30% to 80%. Second, on the test display we included three options: a face identical to the encoded face (T), a more extreme face (T+20%), and a less extreme face (T-20%). Note that the foils differed from the target by 20% rather than 10%. This was chosen because pilot data (N=14, from the same general participant pool) yielded chance-level accuracy for unfamiliar faces when ±10% options were used. Participants completed 384 trials; they were given a break every 24 trials and received feedback about mean accuracy of the preceding 24 trials. Target faces of different familiarity status and morph level were randomly intermixed in trial order.

Results
Figure 7. Results from Experiment 4. Proportion of the three types of response for (a) celebrity faces and (b) non-celebrity faces. Error bars show ± 1 S.E. of the mean.

Accuracy for both celebrity trials (47%) and unfamiliar faces (45%) was above chance, $t(15) = 6.92, p < .001$, Cohen’s $d = 1.73$ in the celebrity condition; $t(15) = 5.68, p < .001$, Cohen’s $d = 1.42$ in the unfamiliar face condition.

To examine whether participants showed a positive memory bias for both types of faces (Figure 7), we entered familiarity, error type, and target morph level as factors in a repeated measures ANOVA. This analysis showed that familiarity altered the nature of memory errors. Although neither the main effect of error type, $F < 1$, nor that of familiarity, $F(1, 15) = 1.27, p = .278$, was significant, there was a significant interaction between error type and familiarity, $F(1, 15) = 18.16, p = .001$, Cohen’s $f^2 = 1.21$. Memory errors showed opposite trends for celebrity and unfamiliar faces: more T+20% errors (29%) than T-20% errors (25%) on celebrity trials, $F(1, 15) = 1.97, p = .181$, Cohen’s $f^2 = 0.13$, and fewer T+20% errors (25%) than T-20% errors (30%) with unfamiliar faces, $F(1, 15) = 2.61, p = .127$, Cohen’s $f^2 = 0.17$. The statistical results held when data were combined across all morph levels.

Discussion

The unfamiliar faces failed to show a positive bias in short-term memory. In fact, whereas memory was biased toward a more extreme option for celebrity faces, it trended toward the less extreme option for unfamiliar faces. The contrast between celebrity and unfamiliar faces is consistent with several previous findings showing that familiarity modulates face perception (Buttle & Raymond, 2003; Megreya & Burton, 2006), including a reduction in the size of the caricature effect for unfamiliar faces (Rhodes & Moody, 1990). Features of unfamiliar faces may be difficult to extract; after a short delay, its distinctiveness further reduces, producing a tendency to choose a less extreme option.

Our study used faces in which the external features were removed from the stimuli, leaving only internal features to distinguish target from distractors. Previous work has shown that internal features likely play a more significant role in the recognition of familiar compared to unfamiliar faces (Ellis, Shepherd, & Davies, 1979; Nachson, Moscovitch, & Umilta, 1994; Young, Hay, McWeeny, Flude, & Ellis, 1985). Our results are consistent with the different roles
played by internal and external features. Future studies should be conducted to examine whether similar results are found when external features are included.

**General Discussion**

Four experiments examined the nature of encoding in visual short-term memory by examining whether errors in discriminating stimuli along a feature continuum are symmetrically centered around the encoding stimulus, or are systematically biased to accentuate or attenuate the distinctive properties of that stimulus. Using faces as stimuli, we explored the direction of memory errors by creating a continuum of morphed faces ranging from specific exemplars of celebrity faces, unfamiliar faces, facial expressions, and inverted faces to neutral or average exemplars. We found that for celebrity faces and emotional expressions, errors in a short-term memory task are systematically biased toward options more extreme than the target face, with no bias for inverted faces and a trend in the opposite direction for unfamiliar faces. This finding is reminiscent of the caricature effect (Lewis & Johnston, 1999; Mauro and Kubovy, 1992; Rhodes et al., 1987; Rhodes & Moody, 1990). This pattern suggests that the information encoded into VSTM may depend on the distinctiveness of the stimuli. Support for this comes from the analysis showing that the bias effect seen for emotional faces was stronger for expressions to which people showed greater sensitivity.

Might the positive memory bias for celebrity and emotional faces reflect an inaccuracy in face perception? That is, perhaps face perception is not veridical – a given morph level A is perceived as A plus x%. This alternative interpretation is logically unsound, analogous to the Greco’s fallacy (Firestone & Scholl, 2014). If face A is perceived as A plus x%, then this should happen both when people view the target face, and when they view the test faces. The same stimulus – A – should yield the same biased perception during encoding and during testing and therefore should be judged the same. The fact that this cancelation did not happen in our data suggests that by the time testing occurs, representation of the target face in memory is more extreme than the percept induced by the matching test face.

We have further evidence that the positive memory bias for celebrity faces is related to judgment involving memory. When choosing an option to match one’s memory (Experiment 1A and Experiment 2), participants exhibit a positive bias toward the extreme option. But when choosing an option to guess the computer’s choice (Experiment 1C), participants exhibit a negative bias toward the least extreme option. The positive bias therefore only occurs when people try to match a choice with an existing memory representation.

However, judgment errors in a VSTM task can reflect processes at multiple stages. Previous studies on change blindness, for instance, show that failure to detect a change can result from multiple levels, including poor retention and errors in comparison (Simons & Rensink, 2005). Similarly, the judgment errors observed in our study can also arise from a bias in memory representation or a bias in memory comparison. On the former account, participants erroneously remembered the encoding face T as T+x%. Consequently, they were more likely to choose the T+x% option than the T-x% option as matching the target. On the latter account, participants may have remembered the encoding face as T. But when the choices were shown,
the one with $+x\%$ deviation may have been more readily recognized as possessing features of the encoded face, biasing people to choose it as the match. Both accounts are compatible with our data and may both contribute to the observed outcome. In fact, these two accounts were consistent with the two models that Rhodes et al. (1987) considered in interpreting the caricature effect. Preferential encoding of critical features exaggerates their presence, producing a bias in the representation. This bias itself will tend to propagate into decision, as the match will depend on the ability to extract these same critical features from the test set. Regardless of whether the bias occurs during memory retention or during comparison, it is clear that the positive decision bias emerges only in the service of matching the test face with a memory representation. It does not occur when there is no memory representation to begin with. In addition, the decision account shares with the memory account the idea that encoding represents a face in terms of a set of critical features that distinguish it from the norm.

The positive memory bias may be related to the phenomenon of anchoring, where mental representation of an object is anchored to a reference point (Jacowitz & Kahneman, 1995). If such anchoring is occurring in face memory, it is clear from our study that the reference point for celebrity faces and emotional faces is not the “norm”, but deviation from the norm. Rapid extraction of deviation from canonical representations appears to be a broad phenomenon in vision. In visual search, for example, people more readily detect a target defined as lacking a canonical feature (e.g., finding a tilted line among vertical lines, or finding a mirror-reflected letter among normal letters), than one that possesses a canonical feature (Wolfe, 2001). Future studies are needed to test whether the positive errors seen for celebrity and emotional faces also extend to stimuli in other domains, particularly those that are likely coded relative to a canonical shape.

Our study also provides suggestive evidence that the degree of positive memory bias is related to the ease with which distinctive features can be extracted from the stimuli. This bias was absent when faces were turned upside-down, and trended in the opposite direction for unfamiliar faces. The difference between celebrity and unfamiliar faces is consistent with other findings in the literature, which have demonstrated differences in face processing of familiar and unfamiliar individuals. In our study, a puzzling aspect was that spontaneous recognition rates of the celebrity faces were low. Although participants could recognize around 70% of the un-edited celebrity faces, spontaneous recognition rate of the face morphs used in the study was around 20% (Appendix B). This level of recognition, however, may be enough to boost the distinctiveness of the celebrity faces. Future studies are needed to experimentally quantify distinctiveness of various facial features (G. Yovel, personal communication, May 2018). To further clarify whether the positive memory bias was driven by familiarity, future studies may also include personally familiar faces, such as the faces of friends or family of the participants. Previous studies suggest that personally familiar faces lead to wider and stronger neural activations, compared with celebrity faces (Caharel et al., 2002; Herzmann et al., 2004; Taylor et al., 2009). If familiarity plays a strong role in the positive memory bias, then the bias should also be observed when personally familiar faces are used.

If face memory is somewhat caricatured for some but not other stimuli, then recall from memory, such as when a witness reproduces the face of a perpetrator, may introduce errors.
Whereas reproduction of familiar faces is likely slightly caricatured, that for unfamiliar faces may be slightly “normalized”. Knowledge about which direction the errors fall may help police or the public to identify individuals based on sketches. Because faces used in the current study are morphed in a holistic manner, sketches that consider configuration (rather than just face features; Frowd et al., 2012) are more likely subjected to the type of memory bias described in our study.

Conclusion

This study shows that short-term memory errors for celebrity faces and emotional faces are biased away from the average face, toward a more extreme representation of the face that had just been seen. This trend was absent or reversed for unfamiliar faces or inverted faces. This finding suggests that distinctive features are extracted to bias subsequent memory and decision. Future studies that experimentally quantify distinctiveness are needed to further examine the role of distinctiveness in modulating memory errors.

Acknowledgements

Jihyang Jun and Deborah Tan made equal contributions to this study, which was supported in part by a graduate research award from the Department of Psychology, University of Minnesota. We thank Galit Yovel for insightful discussions and suggestions, Thomas Busey, Diane Pecher, and Caitlin Sisk for suggestions. Correspondence should be directed to Jihyang Jun or Deborah Tan, 75 East River Road, S504 Elliott Hall. Email: junxx082@umn.edu or tanxx541@umn.edu.

References


Experiment 1A used an iteration procedure (Griffiths et al., 2008) across blocks. Specifically, the target faces started at a particular morph level, such as 40% morph, in Block 1. Whatever participants chose as the match became the next block’s target face, and so on. For example, if participants chose the 50% option as one that matched the target face of Block 1, then in Block 2 they would encode the 50% morph as the target face. In principle, systematic biases (e.g., toward T+10%) should yield new target faces more and more toward the extreme option. Plotting the new morph level shown in subsequent blocks can chart out this “evolution” across multiple iterations of this procedure. Unfortunately, the iteration procedure required the use of stimuli without an upper limit. In our study, we did not have faces above 100% or below 0% in morph level, so when the evolving face hit the boundary (0% or 100%), its evolution was artificially thwarted. Therefore, the procedure could not accurately reveal the final product of the evolution. However, we did observe substantial “evolution” for stimuli not contaminated by hitting the boundary, specifically faces that started at 40% morph level. The following figure (Appendix A, Figure 1) displays the mean morph level used as the target face across the 15 blocks. It is apparent that the morph level increased over multiple iterations. This is an alternative approach to demonstrating the positive memory errors reported in Experiment 1A.

Appendix A, Figure 1. The mean morph level of subsequent blocks’ target faces when the morph level started at 40% in Block 1. Data were from Experiment 1A. Error bars shows ± 1 S.E. of the mean.
Appendix B. Famous face recognition procedure and results

**Procedure.** In experiments involving upright faces (Experiments 1A, 1C, 2, and 4), at the completion of the experiment, participants were asked whether they thought the faces used in the experiment were modified from famous people. They responded with a 4-point scale: 1-all faces, 2-about half of the faces, 3-about a quarter of the faces, and 4-none of the faces were modified from famous faces. Next, we presented the unedited color photograph of each famous individual used and asked participants to name them; a response that named the key role of an actor was considered correct (e.g., “the guy that played Harry Potter” was considered a correct response for Daniel Radcliffe). An experimenter coded each response as correct, partially correct, or incorrect. For each 100% face, participants also reported whether, when performing the memory task, they had recognized this individual.

**Results.** The rates for spontaneous recognition of the faces used in the memory task were low. On average, participants in Experiments 1A, 1C, 2, and 4 respectively reported that 22%, 8%, 28%, and 34% of the faces used in the memory task were from famous people. Recognition rates were much higher when shown the 8 celebrity photographs; participants in Experiments 1A, 1C, 2, and 4 respectively recognized an average of 71%, 64%, 73%, and 66% of the faces.