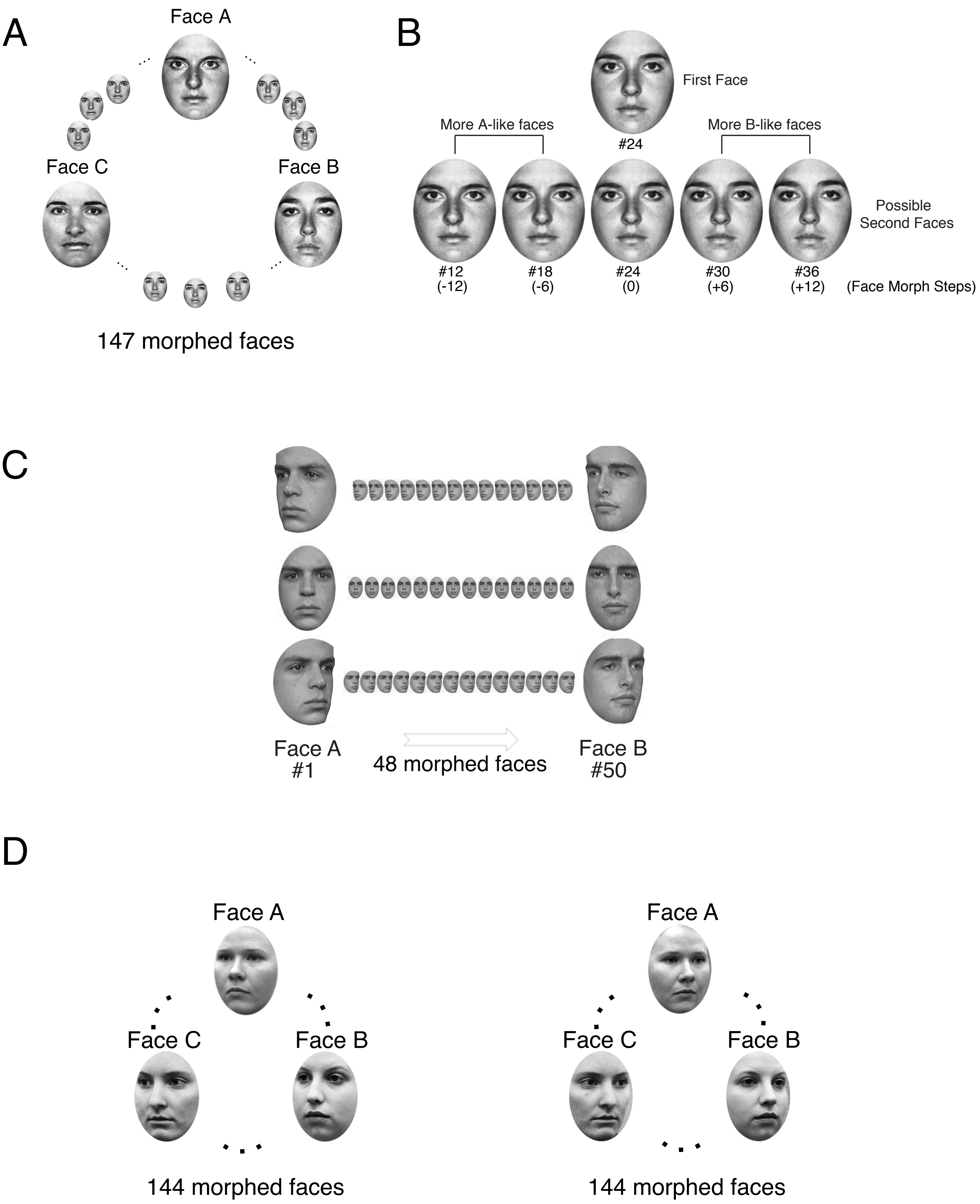
A
Face A
Face C
Face B
147 morphed faces
B
First Face
#24
More A-like faces
More B-like faces
Possible Second Faces
#12 (-12)
#18 (-6)
#24 (0)
#30 (+6)
#36 (+12)
(Face Morph Steps)
C
Face A #1
48 morphed faces
Face B #50
D
Face A
Face C
Face B
144 morphed faces
Face A
Face C
Face B
144 morphed faces

Figure S1. (**Related to Figure 1, 2, 3, and 4**) (A) Face morphs used in Experiments 1 and 2. The stimuli consisted of grayscale image morphs based on 3 original female Ekman faces [S1] with neutral expressions, cropped by an oval aperture to remove the hairline. A set of 48 morphs was created between these identities, resulting in a face morph continuum of 147 faces. (B) Experiment 2 trial structure. The faces used in this experiment were a subset of those in panel (A), including original face A (#1), original face B (#50), and the 48 face morphs in between. The first face presented in each trial sequence was drawn from a subset of 26 faces taken from the center of the morph continuum and could range from face morph #13 to face #38. The second face in the sequence could differ from the first face by ±12, ±6, or 0 face morph steps. Trials that fell in bins -12 and -6 had a first face that was more B-like relative to the second face, trials that fell in the 0 bin had identical first and second faces, and trials that fell in the +6 and +12 bin had a first face that was more A-like relative to the second face. (C) Face morphs used in Experiment 3. We used grayscale image morphs based on 2 original neutral male faces across three different viewpoints (frontal, left, right), cropped by an oval to remove the hairline. A set of 48 morphs was created within each of the viewpoints, resulting in three sets of 50 face morphs. (D) Experiment 4 stimuli. Grayscale image morphs were created from 3 original neutral female faces across two viewpoints (left and right), cropped by an oval to remove hairline. The identities shown here are similar to those actually used in the experiment, with permission obtained for reprint purposes. A set of 47 morphs was created within each of the viewpoints, resulting in two continuous sets of 144 face morphs.

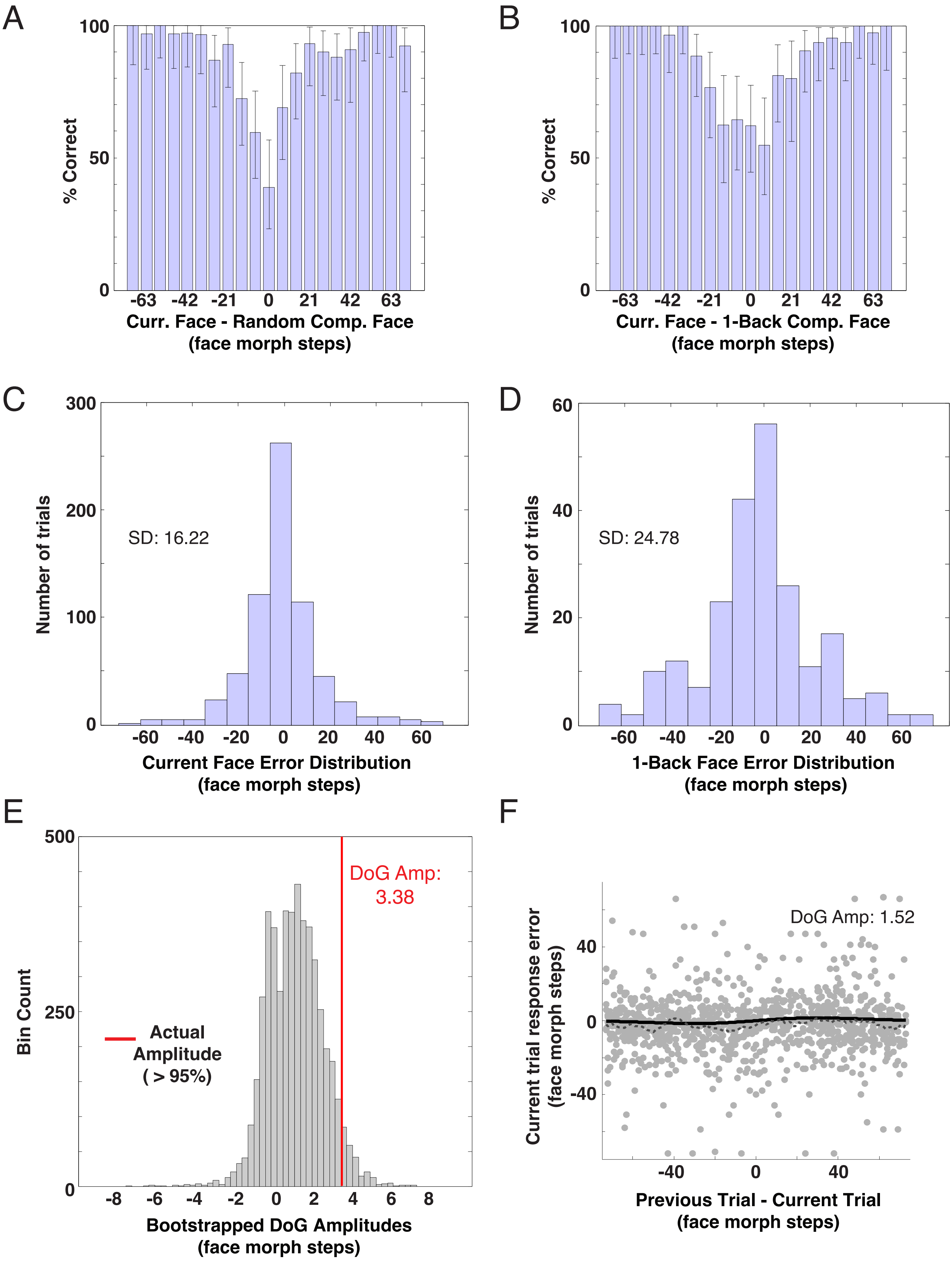
A
100
50
0
% Correct
-63
-42
-21
0
21
42
63
Curr. Face - Random Comp. Face
(face morph steps)
B
100
50
0
% Correct
-63
-42
-21
0
21
42
63
Curr. Face - 1-Back Comp. Face
(face morph steps)
C
300
200
100
0
Number of trials
SD: 16.22
-60
-40
-20
0
20
40
60
Current Face Error Distribution
(face morph steps)
D
60
40
20
0
Number of trials
SD: 24.78
-60
-40
-20
0
20
40
60
1-Back Face Error Distribution
(face morph steps)
E
500
250
0
Bin Count
DoG Amp:
3.38
Actual
Amplitude
(> 95%)
-8
-6
-4
-2
0
2
4
6
8
Bootstrapped DoG Amplitudes
(face morph steps)
F
Current trial response error
(face morph steps)
40
0
-40
DoG Amp: 1.52
-40
0
40
Previous Trial - Current Trial
(face morph steps)

Figure S2. (Related to Figure 1). (A-B) 2AFC experiment testing for memory confusion. It is possible that memory confusion, mistakenly reporting the 1-back face rather than the current face, could have contributed to the pattern of results in Experiment 1 (e.g., Figure 1). To determine whether subjects might have experienced memory confusion, and whether the probability of confusing the target and 1-back face changed with increasing similarity between those faces, we ran a control experiment that measured how often subjects mistakenly reported the 1-back target face rather than the current target face. The stimulus set, trial sequence, and timing were similar to that of Experiment 1b: on each trial, subjects saw a random face for 750 ms, followed by a 1000 ms noise mask and 250 ms fixation cross. Subjects then saw a blank screen for 2000 ms, followed by a screen displaying two faces—a target and a lure—one was randomly assigned to the left and the other to the right of fixation. Subjects indicated which face they had last seen (L or R) using the keyboard arrow keys. One of the comparison faces was always the current face (the target, a correct response), and the second comparison face (the lure) was either a random face (50% of trials) or the 1-back face (50% of trials; picking this 1-back face would constitute a memory confusion). The random lure faces served to establish a baseline for how often subjects made discrimination errors or made lapses. A subject may have mistakenly picked the 1-back lure face on a given trial (rather than the current target face) due to a discrimination error rather than a memory confusion; thus, it is important to have a baseline measure of discrimination error. We tested three subjects in this 2AFC experiment (all of whom had also participated in Experiment 1), and collapsed their data for subsequent analyses. Panel A shows percent correct (with 95% binomial confidence intervals) for trials with a random lure face, binned by the morph step difference between the target and lure faces (discrimination errors and lapses). Panel B shows percent correct for trials with a 1-back lure face, binned by the morph step difference between the target and lure faces (memory confusions). There was no significant difference in the distribution of errors obtained for trials with a random lure versus trials with a 1-back lure (panels A versus B; P = 0.25, permutation test), indicating that memory confusions per se were not common and that most errors were attributable to discrimination errors or lapses. (C-D) The precision of memory for current and 1-back faces. To measure the precision of subjects' memory, we ran the same three subjects from the 2AFC experiment above (and Experiment 1) in a modified version of Experiment 1a. The procedure was identical to that of Experiment 1a, except that on a random and unpredictable 25% of trials, subjects were asked to match the adjustment face to the 1-back target face rather than the current target face. The histogram in panel C shows the error distribution (collapsed across subjects) for judgments of the current target face (75% of all responses), and the histogram in panel D shows the collapsed error distribution for judgments of the 1-back target face (25% of all trials). The standard deviation of the 1-back face error distribution is significantly larger than the standard deviation of the current face error distribution, indicating that subjects were less precise in their overall recollection of the 1-back faces (P < 0.001, permuted null distribution and Mann-Whitney U test). One might argue that, since subjects had to report the 1-back face on only 25% of the trials, the task became more difficult and resulted in broader error distributions for reporting the 1-back face. However, it is critical to note that in Experiment 1 and 4 (Figure 1 and 4), subjects were not required to recall or report the 1-back face at all. Thus, by having subjects report the 1-back face on 25% of trials, the 1-back error distribution we measure here is conservatively narrow—likely more precise than the error distribution in Experiments 1 and 4 for the 1-back face. (E) Memory confusion model. Although there was no significant bias to picking the 1-back face over a random face (no overall difference between panels A and B), we examined whether memory confusion could contribute to the pattern of results in Figure 1. For any bin in which subjects were more likely to choose a 1-back lure than a random lure (higher error rate in panel B than panel A), we attributed the excess errors to memory confusions (a lenient criterion for what counts as memory confusion), and used those values to constrain the memory model. Based on the difference between panels A and B, a total of 2.9% of responses could potentially be classified as memory confusions. To directly model the potential influence of memory confusions, we ran 5000 bootstrapped iterations where we randomly chose 1000 target faces per iteration and simulated method of adjustment responses to each target face based on the empirical error distribution in panel C. For the same 1000 random target face trials, we also simulated responses to each 1-back target face based on the 1-back error distribution in panel D (simulating responses subjects would give if they had experienced memory confusions). We then used this model to generate response data in which subjects experienced memory confusions on a portion of the trials. To do this, we took the 1000 simulated current target face responses and replaced a percentage of those with 1-back target face responses, based on the frequency of memory confusions per bin in the 2AFC experiment. That is, we used the memory confusion error rate to choose a corresponding proportion of randomly sampled trials from the 1-back error distribution. We then treated the resulting simulated data just as we had treated the empirical data from Figure 1, fitting a derivative of a Gaussian (DoG) curve to the combined simulation data and estimating the amplitude of the DoG curve. We repeated this fitting procedure for all iterations of the memory confusion model, with the amplitude estimate from each iteration reflecting the apparent serial dependence that might arise as a result of memory confusions. The histogram in panel E shows DoG amplitudes for 5000 simulation iterations. Without constraining the width of the DoG fits to that of the empirical fits from Experiment 1, 95% of the simulated DoG amplitudes were smaller than the empirical amplitude (red line). Constraining the width of the DoG simulations to within +/- 5 face morph steps of peak serial dependence (testing whether memory confusions could produce serial dependence of the same width and amplitude as we observed in Experiment 1), over 99% of the simulated amplitudes were lower than the empirical amplitude. Finally, even when an unrealistically large proportion of errors are classified as memory confusions (50% of the errors in panel B), over 97% of the simulated amplitudes were smaller than the empirical amplitude. Thus, although there could be a small contribution of memory confusion to our results, the constellation of results here shows that memory confusion cannot completely account for the serial dependence of face perception. (F) One example memory confusion simulation. One example simulated data set, with a DoG amplitude of ~1.5 - significantly less than the empirical amplitude. This simulated amplitude is one of the 5000 iterations in panel E.

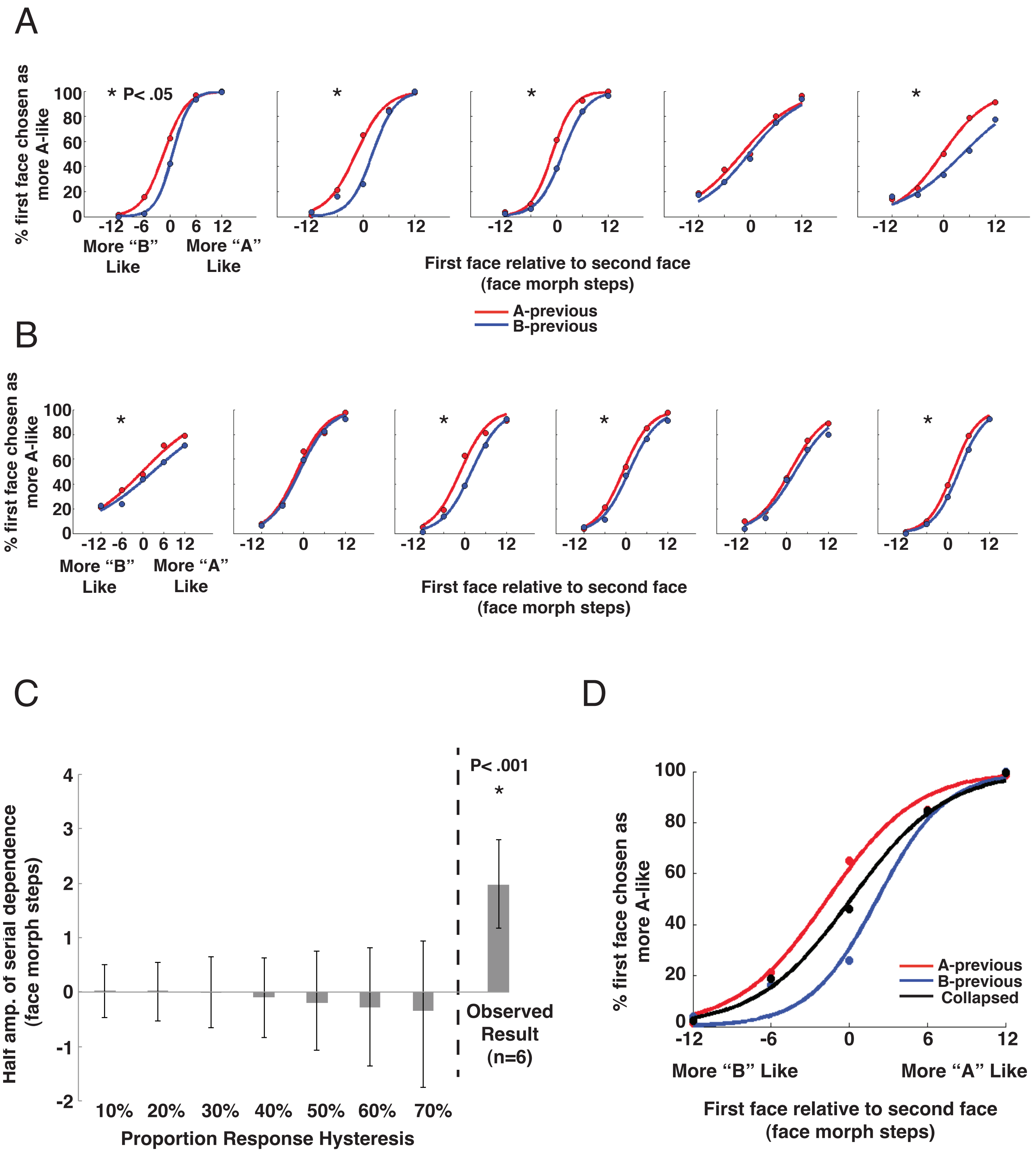
A
% first face chosen as more A-like
100
80
60
40
20
0
* P< .05
*
*
*
-12 -6 0 6 12
More "B" Like
More "A" Like
-12 0 12
First face relative to second face (face morph steps)
A-previous
B-previous
B
% first face chosen as more A-like
100
80
60
40
20
0
*
*
*
*
-12 -6 0 6 12
More "B" Like
More "A" Like
-12 0 12
First face relative to second face (face morph steps)
C
Half amp. of serial dependence (face morph steps)
4
3
2
1
0
-1
-2
P< .001
*
Observed Result (n=6)
10%
20%
30%
40%
50%
60%
70%
Proportion Response Hysteresis
D
% first face chosen as more A-like
100
80
60
40
20
0
-12
-6
0
6
12
A-previous
B-previous
Collapsed
More "B" Like
More "A" Like
First face relative to second face (face morph steps)

Figure S3. (**Related to Figure 2 and 3**) (A) Data from each subject for Experiment 2. Abscissa shows the identity of the first face relative to the second face. Trials that fell in bins -12 and -6 on the x-axis had a first face that was more B-like relative to the second face, trials that fell in the 0 bin had identical first and second faces, and trials that feel in the +6 and +12 bins had a first face that was more A-like relative to the second face. The ordinate shows the proportion of first faces that were chosen as being more A-like. The red data consists of all trials with 1-back first faces that were more "A"-like and the blue data consists of all trials with 1-back first faces that were more "B"-like. Asterisk indicates significance at the 0.05 level; P-values are based on each subject's permuted null distribution. (B) Data for each subject from Experiment 3. The format of the graphs is the same as in panel (A). Experiment 3 was identical to Experiment 2, except sequential trials always contained a different viewpoint. The first and second face within a trial were always viewed from the same angle. Four of six individual subjects showed a significant shift in their psychometric functions based on the identity shown in the previous trial, with the remaining 2 subjects showing a trend in the same direction. (C) Response hysteresis simulation for Experiment 3. The 2IFC design of Experiments 2 and 3 was designed to disentangle perceptual serial dependence from hysteresis in subjects' responses. Nonetheless, we conducted a simulation to test whether repetition of responses could have produced the serial dependence we observed in Experiments 2 and 3. For each subject's trial sequence presented in Experiment 3, we simulated responses with various proportions of response hysteresis. For example, to generate 10% response hysteresis, 90% of trials (randomly chosen) had correct responses and the remaining 10% of trials repeated the 1-back response. For correct trials where subjects saw the same first and second face, we assigned a response at random. We then computed the amplitude of serial dependence within the simulated response sequence, and repeated this analysis 1000 times in order to generate bootstrapped confidence intervals at each proportion of hysteresis. At 80% and greater response hysteresis, the simulated responses became too noisy to reliably fit with psychometric functions. No amount of response hysteresis produced significant serial dependence, indicating that response hysteresis could not be responsible for the perceptual serial dependence we observed in Experiments 2 and 3. (D) Perceptual serial dependence reduces sensitivity. Here we show example data from subject 2 in Experiment 2 with A-previous and B-previous psychometric functions collapsed (black curve). By definition, when the psychometric functions in Experiments 2 and 3 that are separated by trial type (e.g., blue and red functions) are collapsed, they yield a shallower single function compared to the average slope of the two logistic functions separately. That is, subjects' sensitivity in Experiments 2 and 3 was higher when computed separately for trials with different 1-back face identities versus when computed on collapsed data from all trial types ($p < 0.001$, permuted group null, n=12). This consistent shift in PSE when separating responses based on trial-type indicates that the reduced sensitivity is an important perceptual consequence of serial dependence. While perceptual priming leads to an improvement in sensitivity and performance, our results actually show the opposite effect: the perceived identity of a face was misperceived as being more similar to a previous face, which actually reduced overall sensitivity to stimulus difference.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

General Methods

For all experiments, faces were centered on a white background and overlaid with a central fixation cross. Subjects viewed stimuli at a distance of 56 cm on a monitor with a resolution of 1024 x 768 and a refresh rate of 100 Hz. Subjects used a keyboard or mouse for all responses.

All experimental procedures were approved the by UC Berkeley Institutional Review Board. Participants were affiliates of UC Berkeley and provided written informed consent before participation. All participants had normal or corrected-to-normal vision, and all except one were naïve to the purpose of the experiment.

Experiment 1a & 1b

Subjects. Five subjects (4 female; age = 24-32 years) participated in Experiment 1a and four subjects (4 female; age = 18-31) participated in Experiment 1b. One of the subjects in Experiment 1b was not naïve to the experiment, and one of the subjects participated in both Experiment 1a and 1b.

Stimuli and procedure. We used a set of 147 Caucasian female faces with neutral expressions (**Figure S1A**), which were generated from three original Ekman identities [S1] using Morph 2.5 (Gryphon Software). Each presented face subtended 5.9 x 7.3 degrees of visual angle. During the experiment, subjects were tested on their ability to identify randomly chosen target faces with a method of adjustment (MOA) task. We measured subjects' identification errors on the MOA task to determine whether a subject's perception of each target identity was influenced by previously seen target identities.

Experiment 1a

On each trial, a random target face was presented for 750 ms, followed by a 1000 ms noise mask of randomly shuffled black and white pixels, to reduce afterimages, and then a 250 ms fixation cross prior to the response (**Figure 1A**). Subjects then saw a test screen containing a random adjustment face, which they adjusted to match the target face. After picking a match face, subjects saw a 1000 ms noise mask followed by a 1000 ms fixation cross before the next trial began. Here, we use the terms "target face" to mean the face that subjects tried to match, "adjustment face" to denote the randomly-selected face used as the starting point for matching the target, and "match face" for the face that subjects selected as most similar to the target face. The experiment was self-paced and subjects were allowed to take as much time as necessary to respond. We recorded responses based on the numerical value of the match face along the morph continuum, with possible values ranging from 1 to 147. Four subjects each completed 540 trials, and one subject completed 624 trials.

Experiment 1b

In order to rule out any potential biases due to previous motor responses, four additional subjects completed a version of the experiment where half of the trials, selected randomly, did not require a response. Subjects instead saw a surprise blank screen for 2000 ms during the response period, followed by the next trial. Each of the subjects in the no-response condition completed 2382 total trials over 5-6 sessions.

Memory confusion experiments
2AFC
To determine whether subjects might be experiencing memory confusion, and whether the probability of confusing the target and 1-back face changes with increasing similarity between those faces, we ran a control experiment that measured how often subjects mistakenly reported the 1-back target face rather than the current target face (**Figure S2A & B**). The stimulus set, trial sequence, and timing were similar to that of Experiment 1b: on each trial, subjects saw a random face for 750 ms, followed by a 1000 ms noise mask and 250 ms fixation cross. Subjects then saw a blank screen for 2000 ms, followed by a screen displaying two faces—a target and a lure—one was randomly assigned to the left and the other to the right of fixation. Subjects indicated which face they had last seen (L or R) using the keyboard arrow keys. One of the comparison faces was always the current face (the target, and thus, a correct response), and the second comparison face (the lure) was either a random face (50% of trials) or the 1-back face (50% of trials; picking this 1-back face would constitute a memory confusion). The random lure faces served to establish a baseline for how often subjects made discrimination errors or lapses. We tested three subjects in this 2AFC experiment (all of whom also participated in Experiment 1), and collapsed their data for subsequent analyses. Two subjects completed 500 trials over two sessions, and one subject completed 300 trials in one session.

MOA
To measure the precision of subjects' memory for the current and 1-back faces, we ran the same three subjects from the 2AFC experiment above (and Experiment 1) in a modified version of Experiment 1a **(Figure S2C & D).** The procedure was identical to that of Experiment 1a, except that on a random and unpredictable 25% of trials, subjects were asked to match the adjustment face to the 1-back target face rather than the current target face. The histogram in Figure S2C shows the error distribution (collapsed across subjects) for judgments of the current target face (75% of all responses), and the histogram in Figure S2D shows the collapsed error distribution for judgments of the 1-back target face (25% of all trials). Two subjects completed 250 trials in one session, and one subject completed 300 trials over two sessions.

Analysis. Identification error was computed as the shortest distance along the morph wheel between the match face and the target face. Identification error was compared to the difference in target face identities between the current and previous trial, computed as the shortest distance along the morph wheel between the previous target face (1-back) and the current target face. Trials were considered lapses and excluded if errors exceeded +/- 60 morph units (3.5 standard deviations from mean on average, less than 5% of data excluded) or if the response time was longer than 10 seconds. We fit a simplified Gaussian derivative (DoG) to each subject's data of the form:

$$y = abcxe^{-(bx)^2}$$

where parameter y is identification error on each trial (match face – current target face), x is the difference along the wheel between the current and 1-back target face (1-back target face – current target face), a is half the peak-to-trough amplitude of the derivative-of-Gaussian, b scales the width of the Gaussian derivative, and c is a constant, $\sqrt{2}/e^{-0.5}$, which scales the curve to make the a parameter equal to the peak amplitude (**Figure 1B**). We fit the Gaussian derivative using constrained nonlinear minimization of the residual sum of squares.

For each subject's data, we generated confidence intervals by calculating a bootstrapped distribution of the model-fitting parameter values by resampling the data with replacement 10,000 times [S2]. On each iteration, we fit a new DoG to obtain a bootstrapped half-amplitude and width for each subject. We used the half amplitude of the DoG, the a parameter in the above equation, to measure the degree to which subjects' reports of face identity were pulled in the direction of n-back face identities. If subjects' perception of face identity was repelled by the 1-back face (e.g., because of a negative aftereffect; [S3, S4]) or not influenced by the 1-back face (because of independent, bias-free perception on each trial), then the half-amplitude of the DoG should be negative or close to zero, respectively.

In order to calculate significance, we also generated a null distribution of half amplitude (a) values for each subject using a permutation analysis. We randomly shuffled each subject's response errors relative to the difference between the current and 1-back target face and recalculated the DoG fit for each iteration of the shuffled data. We ran this procedure for 10,000 iterations in order to generate a within-subject null distribution of half amplitude values. P-values were calculated by computing the proportion of half amplitudes in each subject's null distribution that were greater than or equal to the observed half amplitude. To test significance at the group level, we chose a random a parameter value index (without replacement) from each subject's null distribution and averaged those values across all five subjects. We repeated this procedure for 10,000 iterations in order to generate a group null distribution of average half amplitude values, and calculated the p-value as described above.

Experiment 2

Subjects. Six subjects (3 female; age = 23-30 years) participated in the experiment. One of the subjects was not naïve to the experiment, and two of the subjects had participated in Experiment 1. One subject became unavailable after one run of the experiment and was only included in the group analysis.

Stimuli and procedure. We used a subset of 50 female faces from the 147 female face morphs used in Experiment 1. This subset of faces consisted of two original female identities, Face A (#1) and Face B (#50), and the 48 morphs between them (**Figure S1A**). Each face was presented in an oval aperture to mask out the hairline and subtended 5.9 x 7.3 degrees of visual angle. Noise was added to the faces to increase difficulty by randomly replacing 5-10% of the pixels in the entire image with black or white pixels.

Before beginning the experiment, subjects were trained to recognize Face A and Face B. During training, subjects were initially shown each face, with a label, and then tested on their recall through 30 randomized trials (15 trials per face). During the randomized trials, subjects were shown either Face A or Face B and had to respond by correctly identifying which face they were shown. If subjects did not get at least 90% of trials correct (27/30), they had to repeat the training phase.

Immediately after training, subjects began the main experiment. Participants were shown a sequence of two faces in each trial, and they had to decide which of the two looked more similar to Face A (much like a two-interval-forced-choice [2IFC] task). The initial face presented in each trial sequence, "first face," was drawn from a subset of 26 faces from the 50 faces used in this experiment. These 26 faces were taken from the center of the morph continuum and could range from face morph #13 to face #38. The following face in the trial sequence, "second face," could differ from the first face by ±12, ±6, or 0 face morphs. Trials that fell in bins -12 and -6 had a first face that was more B-like relative to the second face, trials that fell in the 0 bin had identical first and second faces, and trials that fell in the +6 and +12 bin had a first face that was more A-like relative to the second face (**Figure S1B**). The second face could range anywhere along the morph continuum between Face A and Face B, while the first face was limited to the center of the continuum. Within a trial, the likelihood of the first face being A- or B-like relative to the second face was randomized.

The first face was presented for 1000 ms, followed by a 1000 ms noise mask and 250 ms fixation cross. The second face was presented for 500 ms followed by a 1000 ms noise mask (**Figure 2A**). Subjects responded by identifying which of the two faces looked more similar to Face A by pressing "1" for the first face or "2" for the second face, followed by a 1500 ms fixation cross before the next trial. Four subjects completed 820 trials over two runs, one subject completed 1230 trials over three runs, and one subject became unavailable after one run of 410 trials, but the inclusion of their data had no impact on the group effects.

Analysis. Trials were sorted into one of two groups: B-previous or A-previous. Group membership was determined by comparing the position in the morph continuum of the current trial first face to that of the 1-back first face (**Figure 2A**). Trials for which the 1-back first face was closer to Face A along the morph continuum were labeled as "A-previous" trials and trials for which the 1-back first face was more similar to Face B were labeled as "B-previous" trials. Each subject saw an equal number of A-previous and B-previous trials, but presentation order was shuffled. Once we separated a subject's data into two groups, we fit a separate psychometric function to A-previous and B-previous trials using the following logistic equation:

$$P(respond\ A\ on\ trial\ t) = \frac{1}{1 + e^{-a(x_t - b)}}$$

where x_t is the difference between the first and second face for trial *t* (i.e., -12, -6, 0, 6, or 12), parameter *a* scales with the slope, and *b* is the point of subjective equality (PSE) .

We generated confidence intervals by calculating a bootstrapped distribution of model-fitting parameter values. Within each trial type and bin, we resampled the data with replacement for 10,000 iterations and fit a new psychometric function on each iteration [S2]. We then computed the difference between "A-previous" and "B-previous" bootstrapped *b* (PSE) values in order to generate a distribution of PSE differences. To test for significance, we ran a permutation analysis where we shuffled the 'A-previous' and 'B-previous' labels within each of the five bins. We then recalculated logistic curve fits for the new, randomly assigned A-previous and B-previous trials and computed the difference in PSE between the new parameters. We ran this procedure for 10,000 iterations in order to generate a within-subject null distribution of difference scores. We calculated a p-value by computing the proportion of difference values in each subject's null distribution that were greater than or equal to the observed difference between curves.

To calculate the just noticeable difference (JND) for this set of face morphs, we collapsed each subject's A-previous and B-previous psychometric functions (**Figure S3A**) into one set of data and fit that data with a single logistic function. We then found the x values at which the logistic curve passed through 25% and 75% on the y-axis, and calculated the JND by taking half of the absolute difference between these two x values. The resulting JND was about 4.5 face morph steps between these female faces.

In order to determine whether the 1-back second face also pulled subjects' perception, we fit several lagged logistic regression models to each subject's data and determined which model best predicted subjects' responses. Each successive model tested whether considering another face further back in the past explained significantly more variance in subjects' responses compared to a model without that face. Each subject's data was fit using the following logistic function:

$$P(respond\ A\ on\ trial\ t) = Logit^{-1}(\alpha + \beta_0 \times Offset_t + \sum_{i=1}^{m} \beta_i (X_{t-i} - X_t))$$

where α is a constant, X_t is the position on the linear face continuum of the first face of trial *t*, $X_{(t-i)}$ is the position on the face continuum of the *ith*-back first or second face, and $Offset_t$ is the difference between the first and second face for trial *t* (i.e., -12, -6, 0, 6, or 12). Each lag was calculated as the difference between the current trial first face and the *ith*-back face, i.e. Lag 1: (1-back second face – current first face); Lag 2: (1-back first face – current first face); Lag 3: (2-back second face – current face), and added to the model in a stepwise manner. A negative lag value indicated that the *ith*-back face was closer to Face A (#1) compared to the current first face. To select the best model for each subject's data, we implemented a stepwise procedure where we tested significance of each additional lag using AIC [S5]. Once the highest order lag no longer decreased the model AIC, we stopped adding additional *ith*-back face differences.

We ran an F-test to determine whether the model with the highest order significant lag had a significantly better fit to the data compared to a model with no lags added (i.e. only including the difference between the first and second face on the current trial as a predictor).

Experiment 3

Subjects. Six subjects (3 female; age = 24-36 years) participated in the experiment. One of the subjects was not naïve to the experiment, and three of the subjects had participated in Experiment 2.

Stimuli and procedure. We used grayscale image morphs based on 2 original neutral male faces across three different viewpoints (frontal, left, right), cropped by an oval to remove the hairline (**Figure S1C**). Each presented face subtended 5.64 x 7.47 degrees of visual angle. We randomly replaced 5% of the pixels in each image with black or white pixels in order to increase difficulty by reducing small features that might be diagnostic markers of a given identity.

The procedure for Experiment 3 was identical to that of Experiment 2, except subjects were trained on the two original male face identities, Face A and Face B, within each of the three possible viewpoints (**Figure S1C**).

During the main experiment, subjects were shown the first face for 1000 ms, followed by a 1000 ms noise mask and a 250 ms fixation cross. They then saw the second face for 500 ms followed by a 1000 ms noise mask (**Figure 3A**). Subjects had to indicate which of the two faces looked more similar to male Face A, after which they saw a fixation dot for 1500 ms. No two sequential trials contained the same viewpoint, but the target and comparison face (within a single trial) were always viewed from the same angle. Six subjects completed 820 trials over two runs.

Analysis. Experiment 3 had identical logistic equation fitting, bootstrap, and permutation analysis as in Experiment 2.

Experiment 4

Subjects. Five subjects (4 female; age = 18-37 years) participated in the experiment. One of the subjects was not naïve to the experiment, and three of the subjects had participated in Experiment 1.

Stimuli and procedure. We used a continuum of 144 Caucasian female faces with neutral expressions (**Figure S1D**), which were generated from three original identities across two different viewpoints (left- and right-facing profile), cropped by an oval to remove the hairline. Each presented face subtended 5.9 x 7.3 degrees of visual angle. During the experiment, subjects were tested on their ability to identify randomly chosen target faces with a method of adjustment (MOA) task. The procedure for Experiment 4 was identical to that of Experiment 1, except subjects were trained on the three original female face identities within each of the two possible viewpoints (**Figure S1D**).

During training, subjects were familiarized with Face A, B, and C in each of the two viewpoints. Subjects were initially shown each face turned to the right, with a label, and then tested on recall through 30 randomized trials (10 trials per face). During the randomized trials, subjects were shown one of the three faces and had to respond by correctly identifying which face they were viewing. If subjects did not get at least 90% of trials correct (27/30), they had to repeat the training phase. The same training procedure was then repeated for the left-facing identities.

During the main experiment, a random target face was presented for 750 ms, followed by a 1000 ms noise mask of randomly shuffled black and white pixels, to reduce afterimages, and then a 250 ms fixation cross prior to the response (**Figure 4A**). Subjects then saw a test screen containing a random adjustment face, which they adjusted to match the target face. Importantly, no two sequential trials contained the same viewpoint, but the target and adjustment face (within a single trial) were always viewed from the same angle.

Analysis. Experiment 4 had identical DoG fitting, bootstrap, and permutation analysis as in Experiment 1.

SUPPLEMENTAL REFERENCES

S1. Ekman, P., and Friesen, W. V. (1976). Measuring facial movement. J Nonverbal Behav *1*, 56–75.

S2. Efron, B., and Tibshirani, R. (1993). An introduction to the bootstrap.

S3. Webster, M. A., and MacLeod, D. I. A. (2011). Visual adaptation and face perception. Philosophical Transactions of the Royal Society B: Biological Sciences *366*, 1702–1725.

S4. Clifford, C. W., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., and Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. Vision Research *47*, 3125–3131.

S5. Akaike, H. (1974). A new look at the statistical identification model. IEEE Trans. Auto. Control *19*, 716–723.