# The Ugliness-in-Averageness Effect: Tempering the Warm Glow of Familiarity

Evan W. Carr
Columbia University and University of California, San Diego

David E. Huber
University of Massachusetts, Amherst

Diane Pecher and Rene Zeelenberg
Erasmus University Rotterdam

Jamin Halberstadt
University of Otago

Piotr Winkielman
University of California, San Diego, University of Warwick, and SWPS University of Social Sciences and Humanities, Warsaw, Poland



*Mere exposure* (i.e., stimulus repetition) and *blending* (i.e., stimulus averaging) are classic ways to increase social preferences, including facial attractiveness. In both effects, increases in preference involve enhanced familiarity. Prominent memory theories assume that familiarity depends on a match between the target and similar items in memory. These theories predict that when individual items are weakly learned, their blends (morphs) should be relatively familiar, and thus liked—a *beauty-in-averageness effect* (*BiA*). However, when individual items are strongly learned, they are also more distinguishable. This "differentiation" hypothesis predicts that with strongly encoded items, familiarity (and thus, preference) for the blend will be relatively lower than individual items—an *ugliness-in-averageness effect* (*UiA*). We tested this novel theoretical prediction in 5 experiments. Experiment 1 showed that with weak learning, facial morphs were more attractive than contributing individuals (BiA effect). Experiments 2A and 2B demonstrated that when participants first strongly learned a subset of individual faces (either in a face-name memory task or perceptual-tracking task), morphs of trained individuals were *less* attractive than the trained individuals (UiA effect). Experiment 3 showed that changes in familiarity for the trained morph (rather than interstimulus conflict) drove the UiA effect. Using a within-subjects design, Experiment 4 mapped out the transition from BiA to UiA solely as a function of memory training. Finally, computational modeling using a well-known memory framework (REM) illustrated the familiarity transition observed in Experiment 4. Overall, these results highlight how memory processes illuminate classic and modern social preference phenomena.

*Keywords:* attractiveness, blending, familiarity, memory, mere exposure

*Supplemental materials:* http://dx.doi.org/10.1037/pspa0000083.supp

The origin of preferences is a central topic in social psychology (Allport, 1935; Berntson & Cacioppo, 2009; Schwarz, 2007; Zajonc, 1968, 1998). One key social preference is attractiveness, especially given that human behavior is implicitly and explicitly shaped by the beauty associated with a person, group, object, or idea (Reber, Schwarz, & Winkielman, 2004;

This article was published Online First April 3, 2017.

Evan W. Carr, Columbia Business School, Columbia University, and Department of Psychology, University of California, San Diego; David E. Huber, Department of Psychological and Brain Sciences, University of Massachusetts, Amherst; Diane Pecher and Rene Zeelenberg, Department of Psychology, Erasmus University Rotterdam; Jamin Halberstadt, Department of Psychology, University of Otago; Piotr Winkielman, Department of Psychology, University of California, San Diego, Behavioural Science Group, Warwick Business School, University of Warwick, and Faculty of Psychology, SWPS University of Social Sciences and Humanities, Warsaw, Poland.

Evan W. Carr conducted this research with government support under and awarded by the United States Department of Defense (DoD) and Army Research Office (ARO), via the National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. Piotr Winkielman was sponsored through the UCSD Academic Senate Grant and NSF grant BCS-1232676. We thank Ed Vul, Brad Love, Tim Brady, Gary Cottrell, Angela Yu, Christopher Oveis, David Barner, and John Wixted for insightful comments and feedback on this paper, as well as the many research assistants that helped to run these studies.

Correspondence concerning this article should be addressed to Evan W. Carr, Department of Management, Columbia Business School, Columbia University, 3022 Broadway, 7L Uris, New York, NY 10027. E-mail: ewc2138@columbia.edu


 http://dx.doi.org/10.1037/pspa0000083

Rhodes & Zebrowitz, 2002). Consequently, understanding such preferences not only helps to illuminate the mechanisms underlying affect and cognition, but it also informs practical applications.

Among the classic determinants of preferences in psychology, two have been broadly discussed: *mere exposure* (i.e., stimulus repetition) and *blending* (i.e., stimulus averaging). Both effects occur (at least partially) because familiarity increases preferences. Here, we use these classic phenomena to shed light on the memory mechanisms linking exposure, blending, and preference. More specifically, we explore predictions generated by modern memory models, which link familiarity (and thus, preference) to the degree of match of the target to memory representations. These memory models predict a nuanced relationship between exposure and preference for individuals and their blends, which depend on the amount of learning. To preview the key idea, when individuals are weakly learned (low exposure), their blend has relatively higher familiarity (and thus, preference). In contrast, when individuals are strongly learned (high exposure), their blend has relatively lower familiarity and preference. Overall, using five experiments and computational memory modeling, we find ample support for our general claim that familiarity contributes to preferences for individuals and blends. Critically, we confirm our seemingly nonintuitive prediction that the relative preferences for individuals and their blends reverses with increasing prior exposure to the stimuli used to create the blend. Next, we offer some background on mere exposure, blending effects, and modern memory models.

## Mere Exposure, Blending, and Social Preferences

Among the most well-known psychological phenomena is the *mere exposure effect*—or increased preference from unreinforced stimulus repetition—which dates back at least to Titchener's (1915) observations about the "warm glow of familiarity." Zajonc (1968) has renewed the field's interest in mere exposure, and since then, it has been investigated and applied across psychology and business settings (Baker, 1999; Balogh, & Porter, 1986; Obermiller, 1985; Pettigrew & Tropp, 2008; Thompson, 2017; Tremblay, Inoue, McClannahan, & Ross, 2010; Zajonc, 2001). The effect is robust across a wide range of stimuli (e.g., faces, words, sounds, and images) and modalities (e.g., vision, audition, touch, and smell), though subject to important boundary conditions (Bornstein, 1989).

Theoretically, the mere exposure effect offers an important window into emotion-cognition links and processes underlying implicit memory. The connection between repetition and preference could occur for many reasons (for reviews, see Fang, Singh, & Ahluwalia, 2007; Moreland & Topolinski, 2010), but much evidence suggests that repetition facilitates processing and elicits an implicit sense of familiarity (Bornstein & D'Agostino, 1992; Butler & Berry, 2004; Klinger & Greenwald, 1994; Winkielman, Schwarz, Fazendeiro, & Reber, 2003). Although the mere exposure effect is tied to the subjective sense of familiarity, it does not depend on the explicit recognition that the stimulus is "old" (Whittlesea & Price, 2001). Importantly, mere exposure effects on preferences generalize to stimuli that are similar to ones seen previously yet objectively *new* (Whittlesea, 2002), and this generalization follows a similarity gradient between the original and test stimulus (Gordon & Holyoak, 1983). Such generalization effects have also been obtained for social stimuli such as faces (Rhodes, Halberstadt, & Brajkovich, 2001), and exposure to other-race faces can increase liking for objectively new faces within that same race group (Smith, Dijksterhuis, & Chaiken, 2008; Verosky & Todorov, 2010; Zebrowitz, White, & Wieneke, 2008). Generalization effects also offer a path toward changing real-world social preferences that extend beyond the specific individuals engaged in personal interactions (e.g., intergroup contact; Pettigrew & Tropp, 2008). Therefore, it is important to understand the nature, mechanisms, and limitations of mere exposure effects and their generalization.

Another classic phenomenon in the domain of preferences is *blending* (or stimulus averaging). Since the original observations by Galton (1879) on composite portraits, psychologists have documented that averaging makes stimuli more attractive across a variety of different modalities and stimuli. This effect occurs for abstract dot patterns, colors, birds, cars, watches, fish, voices, and gestures (Bruckert et al., 2010; Halberstadt & Rhodes, 2003; Winkielman, Halberstadt, Fazendeiro, & Catty, 2006; Wöllner et al., 2012), but it is especially robust for faces (Halberstadt, 2006; Langlois & Roggman, 1990; Rhodes & Tremewan, 1996). Many explanations have been proposed for this *beauty-in-averageness (BiA) effect.* Some authors invoke evolutionarily shaped "mutant-detector" mechanisms, where morphed faces signal greater fitness, due to greater symmetry and a lack of unusual features (Thornhill & Gangestad, 1993). However, as with the mere exposure effect, the dominant explanations are cognitive. Langlois and Roggman (1990) point out that blending several faces makes the average face more similar to the central tendency of a local population of faces encountered by the participants. In fact, the attractiveness of averaged faces varies as a function of exposure to different populations of faces, suggesting the importance of learning processes (Dotsch, Hassin, & Todorov, 2016; Principe & Langlois, 2012; Rubenstein, Kalakanis, & Langlois, 1999). Consistently, the attractiveness of average faces is also associated with their implicit familiarity (Peskin & Newell, 2004; Rhodes et al., 2001). This fits with many studies that use abstract patterns (e.g., random dots) which are derived from a category average. The average (even when not studied itself) is familiar and preferred because of its similarity to exemplars in memory, as reflected in liking judgments and physiological measures (Winkielman et al., 2006).

## Memory Models (and How Familiarity Works)

The above discussion highlights the importance of understanding the mechanisms of familiarity for social psychological theories of preference. We argue that the relevant memory literature not only helps explain why these classic preference phenomena occur, but it also helps us to identify the boundary conditions under which they disappear (and even reverse). For simplicity, we only briefly review the core assumptions that informed our reasoning behind the current experiments. However, other important and relevant aspects of the memory literature, including its quantitative, computational, neuroscientific, and applied components, are available across several reviews (Gillund & Shiffrin, 1984; Mandler, 1980; McClelland & Chappell, 1998; Wixted & Mickes, 2014; for a specific application of the computational or connectionist perspective to key questions in social psychology, refer to a review by Smith [1996]).

To first review our terminology, *objective familiarity* refers to the actual exposure history (i.e., how many times the stimulus was encountered), s*ubjective familiarity* refers to a "sense of knowing" for the stimulus, whereas *recognition* refers to a judgment about a previous encounter with the stimulus. These distinctions are important because, as mentioned above, the relation between familiarity and preference primarily concerns subjective familiarity. Incidentally, it is worth noting that in the memory models discussed here, subjective familiarity is often (though not always) linked to *fluency,* or the ease of stimulus processing. This is because a previous encounter with an item is thought to increase the activation, reprocessing efficiency, and thus retrievability of its trace. For most of this paper, we will focus on subjective familiarity, but we will revisit the issue of fluency in the General Discussion.

What elicits subjective familiarity? Prominent memory theories suggest that familiarity of a probe depends on the "global match" between the probe and the set of items in memory to which it is compared (Gillund & Shiffrin, 1984; Hintzman, 1986). These models assume that memory contains a vast array of separate memory traces for all previous events. When presented with a memory probe (e.g., a question asking "Have you seen this face before?"), the probe item matches a subset of the memory traces, and this subset is tallied up to provide a "global match" value, specifying the degree of familiarity for the probe. In short, familiarity is a measure of how well a stimulus matches everything in memory.

According to these global match memory models, familiarity for a previously encountered item reflects the summation of one strong match value (a match to the actual memory trace of the probe item) and a large number of smaller values owing to partial matches to similar memory traces. If the actual memory trace is weak (because only a few item features were stored), the corresponding memory trace will only be weakly active, owing to a small number of matching features, as compared to a stronger memory. Thus, familiarity will be higher for strongly learned items than for weakly learned items. However, familiarity can also be greatly influenced by the other memory traces, particularly if some of those memories are similar to the item used to probe memory, resulting in potential false memories for highly similar, prototypical, and/or "central" items (Roediger & McDermott, 1995; Shiffrin, Huber, & Marinelli, 1995), including composite faces (de Fockert & Wolfenstein, 2009). In these memory models, retrieval strength for each memory trace is calculated from the number of matching features between the probe and the memory trace (Hintzman, 1986; Murphy, 2002). This helps to explain the mere exposure effect, given that stronger memory traces for actually studied items will result in a better match (and thus, higher familiarity values and greater preference). Global matching models were first developed to explain episodic recognition memory, and they assumed that the memory decision for whether a probe was old or new was based on the global familiarity of the probe (Hintzman, 1986; Nosofsky, 1986). If the familiarity of the probe exceeds a threshold, it is judged to be old; otherwise, it is judged to be new (for our purposes in the current paper, these models also apply to identification and categorization).

Global match memory models explain the BiA effect from the following process: First, participants are incidentally exposed to many exemplars using minimal exposure, which results in the formation of very weak individual traces. Later, participants are presented with the blend probe (or morph) that is similar to many of these memory traces by virtue of being a blend of the stored memories. The more similarity the blend has to all other face traces, the more familiar (and preferred) it will be compared with the weakly learned individual faces. Consistent with this account, traditional BiA paradigms use only single incidental exposure to individuals. Further, evidence shows that the BiA effect increases with the number of faces that compose the blend. In fact, the classic Langlois and Roggman (1990) paper only observed a clear BiA effect when averaging eight or more individuals, which may make the morph appear very familiar (compared to a morph that averages only two individuals).

Even though global match memory models are among the most popular and widely accepted (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997), one observation that challenged the global match assumption was the null *list-strength effect* in recognition memory (e.g., Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990). A list-strength effect occurs when memory retrieval becomes more difficult by strengthening competing memories (usually other items on the study list). As predicted by global match models, there is a positive list-strength effect when actively *recalling* something from memory (e.g., by practicing your new phone number, it becomes difficult to recall your old phone number). However, the same prediction does not apply for *recognition* (e.g., practicing your new phone number has no effect on your ability to recognize your old phone number), and there was also some evidence for a negative list-strength effect in recognition (e.g., practicing your new phone number *helps* you to recognize your old phone number). Subsequently, global match models were revised by considering not just how well a probe item matches memory traces but also the extent to which a probe item *mis*matches memory traces. Thus, when a memory trace is strengthened through additional exposure, its representation becomes more complete, increasing the chance that the critical mismatching features are stored. If mismatching features are stored, the memory traces containing these mismatching features are "differentiated" from the probe item, and these traces contribute less to the summed global match familiarity signal. In the memory literature, this process is referred to as *differentiation*—where stronger encoding of a stimulus clarifies the differences between it and the test item (Shiffrin et al., 1990). Most importantly for our purposes, differentiation predicts that with increasing memory strength for actually encountered individual items, highly similar but actually new probe items (like a morph) will be become less familiar than the individual items.

## How Do Exposure and Blending Effects Interact to Drive Familiarity and Preferences?

With the above principles in mind, we derived several predictions regarding the combined effects of exposure and blending on familiarity and preferences. Our central prediction is that the effects of blending two faces should depend on the larger memory context—and more specifically, on the amount of prior exposure to individual faces contributing to the blend. First, when participants have no memory traces for any related individual exemplars, there should be no BiA effect, since the blend is not similar to anything. Next, when participants have weak, undifferentiated memory

traces for individual exemplars, there should be a traditional BiA effect because the blend will at least partially match multiple faces in memory, producing greater familiarity for the blend compared to a particular exemplar face. Finally, when participants have strong memory traces for individual exemplars, those exemplars become well differentiated from the blend. In turn, the blend will partially match the stored exemplar while also *mis*matching it, resulting in relatively less familiarity for the blend. Thus, with strong memories for the exemplars, our theoretical perspective predicts an *ugliness-in-averageness (UiA) effect*. Importantly, note that the morph should still benefit from some similarity (or partial match) to the exposed exemplars, and thus have greater familiarity and liking than completely unfamiliar stimuli. As such, "ugliness" is defined here as a relative difference in preference compared to the components, instead of an absolute decline. In other words, blending highly familiar exemplars should reduce the benefits of their exposure, but it should not bring the morph below the original attractiveness level of unfamiliar face blends.

Although this novel prediction has never been tested, it is consistent with studies where participants judge stimuli that are objectively new but include features of previously learned exemplars. For example, in one memory paradigm, participants first studied words like "blackmail" and "jailbird," and then were asked about the word "blackbird," as well as the original and control items (Jones & Jacoby, 2001). Another paradigm instructed participants to first study word pairs (e.g., *table-clock*, *fish-computer*, etc.) either only once (weak pairs) or several times (strong pairs). Next, they were asked about intact pairs, rearranged pairs, and control items (Kelley & Wixted, 2001). In both cases, participants showed an elevated false alarm rate to the "blended items" (e.g., "blackbird" or *fish-clock*). Crucially, the false alarm rate was lower than the recognition of actually presented items and was further reduced (but not eliminated) when participants had a stronger memory of the initially studied items. Again, the theoretical interpretation is that "blended items" create a sense of familiarity, but strong memory traces for their individual components increase differentiation.

Our memory-based prediction is also distinct from other alternative accounts. The most intuitive alternative prediction is that the effects of exposure and blending are *additive*—that is, preferences from mere exposure and blending should combine in a positive fashion, making the morph of familiar individuals very attractive. This prediction is similar to the additive pattern observed from combining subliminal affective priming with smiling faces and mere exposure on liking of ideographs (Monahan, Murphy, & Zajonc, 2000), which follows from assumptions that mere exposure and blending involve separate mechanisms. Other accounts make the complete opposite *mismatch* prediction, where mere exposure and blending combine negatively, making the morph of two familiar individuals especially unattractive (reducing liking for the blend below the level of the contributing individuals). This prediction follows from theories of ambiguity aversion and cognitive conflict, given that the morph of well-known individuals falls in-between two established categories (Arnal & Giraud, 2012; Dreisbach & Fischer, 2015; Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005). Importantly, these additive and mismatch predictions differ from our familiarity-based predictions not only in mechanism but also in the actual data pattern: Unlike either of these frameworks, our account predicts that blends of highly learned individuals will generate familiarity and preference values in-between actually exposed individuals and novel individuals (Jones & Jacoby, 2001; Kelley & Wixted, 2001).

Finally, our account is also supported from related research using blends of real faces from foreign and local celebrities (e.g., Halberstadt, Pecher, Zeelenberg, Wai, & Winkielman, 2013). This study found that morphs of two celebrity faces were more attractive than the individual celebrities used to generate them (a standard BiA effect). However, this only occurred when those "celebrity" individuals were unknown in the participants' home country (i.e., they were only famous in another country). When local celebrities were blended, participants rated the morph as less attractive than the individual faces (a pattern indicative of a UiA effect). Although this study is consistent with our hypothesis, it fails to answer four essential questions. First, it did not offer or explore any mechanisms for how exposure and blending interact in driving attractiveness judgments, as we propose here with our memory-based framework. Second, since Halberstadt et al. (2013) did not systematically manipulate exposure, those studies cannot provide any evidence for boundary conditions (e.g., perhaps the effects require massive experience with the individuals, over many years and exposures). Third, the study lacked control conditions to address whether blends of well-known faces are actually disliked (below novel faces) or just less liked than individual faces of well-known individuals. Lastly, celebrity morphs do not provide an effective substitute for learning tasks or exposure manipulations, given other confounds. For instance, participants may simply dislike distorted images of media celebrities (i.e., "Don't mess with the face of my sports hero!") or dislike blends of individuals that represent divergent social views (i.e., "Don't mix liberals and conservatives!", in the case of the famous "Bushama" [Bush-Obama] or "Clump" [Clinton-Trump] blends).

## Current Research

The current research offers the first systematic investigation of the idea that the attractiveness of facial blends varies as a relative function of their prior exposure. We used five studies and computational memory modeling to examine how the attractiveness of individual and morph faces changes with learning. The key prediction was that blends of highly familiar faces (with exposure experimentally manipulated) would be *less* attractive than their constituents (UiA effect), but this would not apply to blends of novel or weakly familiar faces (no effect or BiA effect).

To preview the results, Experiment 1 established the traditional BiA effect under standard conditions when all the stimuli were initially unknown and exemplars were only weakly learned. Furthermore, this experiment showed that increased attractiveness for morphs was mediated by their perceived familiarity. Experiments 2A and 2B tested for the UiA effect under empirical conditions that directly compared preferences for blends of learned and unlearned individual faces. Participants were "trained" on a subset of faces (either using a free-recall task with face-name pairs [Experiment 2A] or a perceptual-tracking task with colored squares presented on the faces [Experiment 2B]), where they were repeatedly exposed to one set of individual faces but not the other, thus creating a stimulus set of trained and untrained individuals. Both experiments showed a UiA effect for trained faces, where morphs of trained individuals were rated as *less* attractive than the trained

individuals themselves. In Experiment 3, we restructured the stimulus set to examine whether the UiA effect for trained morphs was driven by cognitive conflict (mismatch account) or a relative reduction in similarity (familiarity account). We found strong support for our familiarity-based hypothesis, where the UiA effect was still generated for morphs that did not have competing individual components (i.e., morphs composed of one trained face and one untrained face). With Experiment 4, using a within-subjects design, we varied the number of exposures for individuals across four different levels, and participants also completed speeded "old/new" recognition judgments on all face stimuli after giving their attractiveness and familiarity ratings. The results supported our memory-based predictions, where a traditional BiA effect emerged with weak learning on individual exemplars, but this reversed into a UiA effect with strong learning. Finally, simulations of memory judgments using the Retrieving Effectively from Memory (REM) model (Shiffrin & Steyvers, 1997) produced the same crossover interaction we observed in Experiment 4, with a BiA effect for weak learning and a UiA effect for strong learning.

## Experiment 1

In Experiment 1, we tested whether our stimulus set generated a standard BiA effect using a design with minimal exemplar learning. We expected that when many individual exemplars are presented without strong learning of any of the specific exemplars, the morphs of those exemplars would be rated as more attractive and familiar. Furthermore, the latter effect (familiarity) should explain the former effect (attractiveness). This prediction follows from previous research showing that incidental exposure to several exemplars, leading to limited item-specific memory, generates familiarity for a prototypical representation (de Fockert & Wolfenstein, 2009; Posner & Keele, 1968; Winkielman et al., 2006).

### Method

**Participants.** One hundred fifty-one University of California, San Diego (UCSD) undergraduates participated for course-credit, and all participants signed consent forms approved by the UCSD Human Research Protection Program (HRPP).

**Materials.** Our stimulus set included 56 individual face images of Dutch and New Zealand (NZ) people (28 each), along with 28 50/50 morphs of those faces (14 Dutch-Dutch and 14 NZ-NZ morphs), for a total of 84 unique stimuli (adapted from a previous study; see Halberstadt et al., 2013). Each individual was only used in one of the morphs, and each morph contained two individuals.

**Design and procedure.** We conducted this as an online study, where all participants were told that they would be rating 84 faces on attractiveness and familiarity. Participants were presented with all 84 faces from our stimulus set (56 individuals and 28 morphs) one-at-a-time, in a randomized order. Note that one feature of this standard design is that morphs will sometimes be preceded by their constituting exemplars (making the morphs somewhat familiar). For each face, participants were asked to rate each image separately on attractiveness and familiarity, using 1 (*not at all* attractive/familiar) to 9 (*very* attractive/familiar) scales.

### Results and Discussion

**Analysis strategy.** To analyze ratings in Experiment 1, we used mixed-effects modeling via restricted maximum likelihood. This method offers numerous analytical advantages over more traditional methods like repeated-measures ANOVA, which were important for our purposes (see the following for more details: Bagiella, Sloan, & Heitjan, 2000; West, Welch, & Galecki, 2014). First, they handle unbalanced designs, unequal sample sizes, and missing observations more efficiently, thus leading to more reliable outcomes. Second, mixed-effects models also involve a model for the error variance, resulting in more powerful and efficient estimates. Further, they are more flexible in allowing one to model the dependence of outcomes on both fixed and random-effect predictors.

All models were built using the *lme4* (Bates, Maechler, Bolker, & Walker, 2014) and *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2014) packages in R. To obtain $p$ value estimates for fixed-effects, we used Type III Satterthwaite approximations (Luke, 2016). Note that this process can result in decimal degrees of freedom (based on the number of observations), and degrees of freedom are often greater with mixed-effects models since the analyses are done on trial-level data (see footnote for more details on mixed-effects modeling strategy).[1] While we report the results from mixed-effects models in the main text for all experiments, alternative analyses using traditional repeated-measures ANOVAs are also reported in the supplementary materials (which corroborate all the results in the main text).

**Attractiveness and familiarity.** For Experiment 1, we used Target Type (2 [within]: individual, morph) as the only fixed-effect to predict attractiveness and familiarity ratings. As predicted, participants rated morphs as more attractive ($M = 4.32$, $SD = 1.17$) than individuals ($M = 4.20$, $SD = 1.15$), $F(1, 150.00) = 26.42$, $p < .001$ (see Figure 1a). This confirms that our stimulus set yields a traditional BiA effect in the standard paradigm, when only weak exemplar learning occurs.

Consistently, morphs were also rated as more familiar ($M = 2.46$, $SD = 1.44$) than the individuals ($M = 2.36$, $SD = 1.37$), $F(1, 150.00) = 6.63$, $p = .01$ (see Figure 1a). Note that the familiarity values are rather low, toward the "not at all" end of the 1–9 familiarity scale. This also confirms that the standard procedure used by most BiA studies yields only minimal learning of exemplars and generates only slightly greater familiarity for the morph.

**Multilevel mediation.** To gauge the relative impact of participants' familiarity ratings on the relationship between morphing and attractiveness ratings, we applied multilevel mediation analyses to each participant's data, via the *mediation* package in R (R Core Team, 2015; Tingley, Yamamoto, Hirose, Keele, & Imai,

[1] Final mixed-effects models were selected based on top-down model building. Maximal random intercept and random slope models were created using all by-participant effects. Next, the two model fits were tested against one another via $\chi^2$ likelihood ratio tests. If there was no significant difference in model fit, the model with fewer random-effect parameters (i.e., only random intercepts) was set as the final model; if there was a significant difference in model fit, the model with more random-effect parameters (i.e., random intercepts and random slopes) was set as the final model. This final model was then used for fixed-effects testing, which employed the *lmerTest* package in R (Kuznetsova, Brockhoff, & Christensen, 2014).

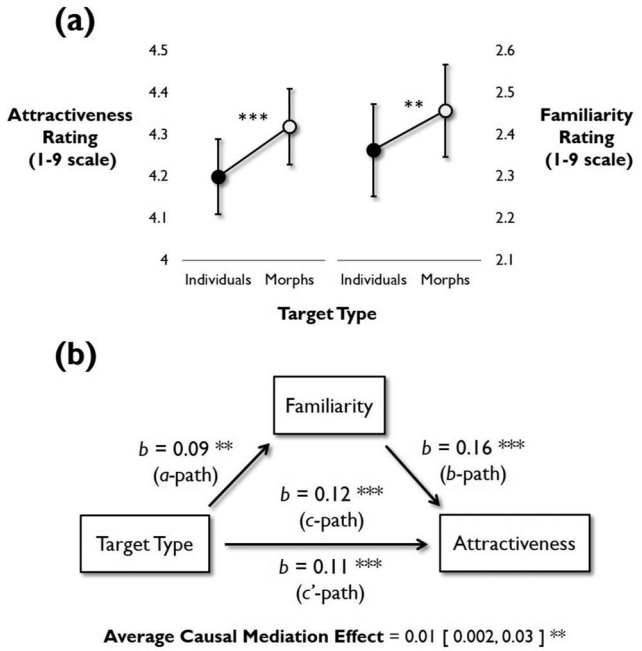


*Figure 1.* Experiment 1 results for attractiveness and familiarity (a), along with multilevel mediation (b). We demonstrated that when weak exemplar learning occurs, our stimulus set generates a standard beauty-in-averageness (BiA) effect, where morphs were rated as more attractive than individuals. Morphs were also rated as more familiar than individuals, and this familiarity mediated the relationship between target type (individuals vs. morphs) and attractiveness ratings (b). Error bars represent ±1 *SEM*. ** $p \leq .01$, *** $p \leq .001$.

2014). Such a strategy is appropriate for repeated-measures designs to account for observations nested within participants, and they allow for model-based estimation of the average total, direct, and indirect mediation effects using hierarchical data structures (Bauer, Preacher, & Gil, 2006). Mixed-effects models were constructed for each of the mediation paths, using by-participant random effects parameters. All simulations from the *mediation* package in R were based on 1,000 samples per estimate, after which quasi-Bayesian confidence intervals were calculated around the average total, direct, and causal mediation effects. Our main predictor was target type (coded as either 0 [individual] or 1 [morph]), our main DV was attractiveness ratings, and our mediator was familiarity ratings.

Figure 1b displays the mediation results. We observed clear evidence for mediation. The total effect ($b = 0.12$, $CI_{95\%}$ [0.07 0.16], $p < .01$) and average direct effect ($b = 0.11$, $CI_{95\%}$ [0.06 0.15], $p < .01$) on attractiveness ratings were both significant. Target type was a significant predictor of familiarity (*a*-path: $b = 0.09$, $t(150.00) = 2.57$, $p = .01$), and familiarity was a significant predictor of attractiveness (*b*-path: $b = 0.16$, $t(282.65) = 3.73$, $p < .001$). When controlling for familiarity (*c'*-path), the original *t*-value estimate of target type on attractiveness (*c*-path: $b = 0.12$, $t(150.01) = 5.14$, $p < .001$) was reduced but still significant ($b = 0.11$, $t(152.29) = 4.68$, $p < .001$), while familiarity was also significant ($b = 0.13$, $t(276.30) = 3.13$, $p = .002$). And critically, the average causal mediation effect was also significant ($b = 0.01$, $CI_{95\%}$ [0.002 0.03], $p = .01$), confirming familiarity as a mediator.

## Experiment 2A

Experiment 1 demonstrated that with weak exemplar learning, morphs were judged as more attractive and familiar than individuals (a traditional BiA effect). These results fit with the memory literature, where in the absence of any strong individual memory traces, the blend has high global familiarity.

We designed Experiment 2A to address our main question. Namely, we wanted to test the idea that an ugliness-in-averageness (UiA) effect could be generated when participants undergo strong learning on the individual exemplars, before rating morphs. Recall that when the memory traces for individual exemplars are strengthened by repeated exposure, they should now be highly familiar and differentiated. Therefore, when a blend of such strongly learned individuals is presented, the blend will be *less* familiar than the exposed individuals, leading to a UiA effect. It is also important to note that when individual exemplar memory is increased, all individuals may appear overall more familiar (even unexposed individuals), given that mastering individual exemplars from a particular face set may give participants a greater sense of familiarity for that specific "face space."

To test our predictions in Experiment 2A, we "trained" participants on a subset of faces (set A vs. set B), using a free-recall task that required pairing names with individuals. Over the course of this task, participants were repeatedly exposed to one set of individual faces but not the other, creating a stimulus set of trained and untrained individuals and morphs. After training, participants rated the attractiveness and familiarity of all morphs and individuals.

### Method

**Participants and equipment.** Seventy-four UCSD undergraduates participated for course-credit, and all participants signed consent forms approved by the UCSD HRPP. During the main task, all stimuli were presented on 17-inch Dell flat screens from PCs running Windows XP and E-Prime 2.0.

We planned our sample size in Experiment 2A based on the effect size of attractiveness ratings between individuals and morphs in Experiment 1 ($d_Z = 0.42$). We conducted a post hoc power analysis of Experiment 1 with GPower software (version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007), which indicated that we achieved more than 99% power (using a two-tailed test at $\alpha = .05$). Because Experiment 2A required in-lab participants, we instead aimed for 85–90% power. Based on the design and smaller effect size estimate of $f = 0.15$ (nonsphericity correction $\varepsilon = 1$), this forecasted a target *n* of 70–81 participants.

**Materials.** The 56 individuals and 28 morphs from Experiment 1 were used to create two different sets of images (set A and set B) that each contained half the total number of individual faces (28 in each set) and half the total number of morph faces (14 in each set). Using attractiveness ratings from a previous study (Halberstadt et al., 2013), we normed both sets, such that the average attractiveness ratings for individuals and morphs were similar across sets. All morphs were 100% within-set, meaning that morphs could either be 50/50 morphs of two set A individuals (A-A morphs) or 50/50 morphs of two set B individuals (B-B morphs). There were no cross-set (A-B) morphs (see the supple-

mentary materials and Figure S1 for more information on the stimulus sets).

**Design and procedure.** All participants were first told that they would be completing a memory task, where they would have to recall different face-name pairs, followed by ratings on different dimensions. Participants were not told until after training that they would be rating attractiveness and familiarity. For training, participants were randomly assigned to study the 28 individual face stimuli in either set A or B, before progressing through 7 rounds of a free-recall task.

Figure 2a depicts the structure of the paradigm. At the start of each round, the 28 individuals in the participant's assigned training condition were each randomly presented in a study phase. Each image was presented with a four-letter name for 3000 ms each, one-at-a-time. Next, after all 28 individuals were presented, participants were given a test where they had to recall the name that was paired with each face. They would type the name in a response box presented on the screen, and feedback (correct vs. incorrect) was given. During test phases, RTs were measured from stimulus onset to the final submission of the participant's typed response to each face (recorded when they hit the ENTER key to advance to the next face). Participants cycled through all 28 faces during every study and test phase, across all 7 training rounds. The names that were paired with each face stayed the same across all training rounds. To encourage high attention and effort throughout the memory task, participants were told that they would only advance to the next part of the experiment once they hit a satisfactory level of performance (in reality, participants always completed 7 training rounds to keep the level of exposure consistent).

After participants finished the training, they rated each stimulus (56 individuals and 28 morphs) using 9-point scales on attractiveness (1 = *not at all attractive*; 9 = *very attractive*) and familiarity (1 = *not at all familiar*; 9 = *very familiar*). Each participant always rated the stimuli in the following block order: (a) morph attractiveness, (b) individual attractiveness, (c) morph familiarity, and (d) individual familiarity. Morph ratings always came first to ensure that they were not influenced by exposure to untrained individuals, because we predicted that any UiA effect would occur after exposure to trained individuals. On attractiveness ratings, participants were asked "How

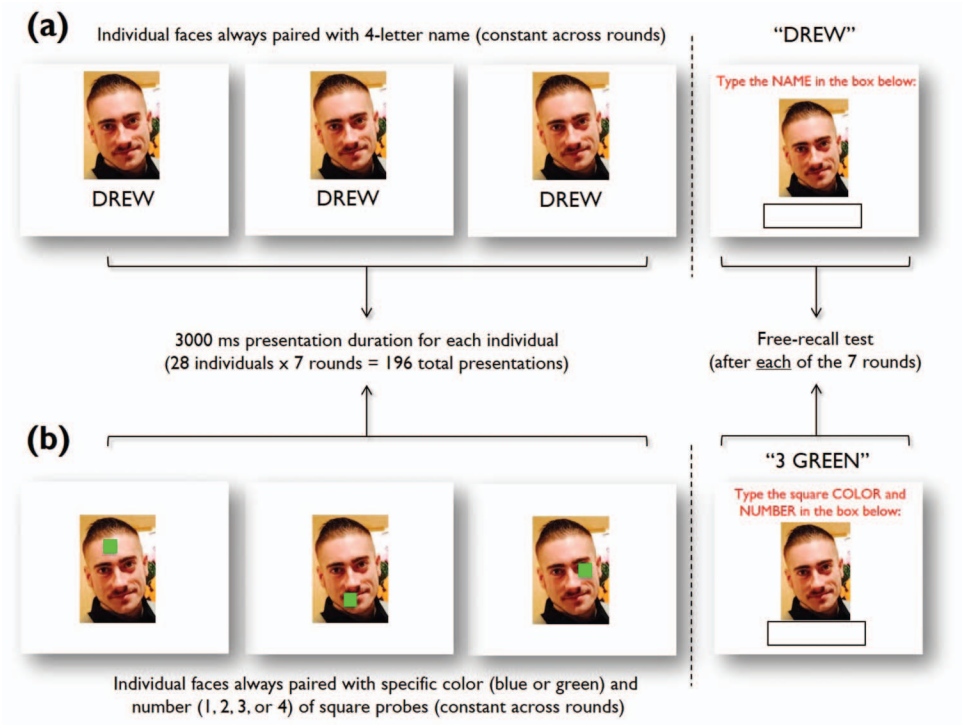


*Figure 2.* Design of the training task for Experiments 2A (a), 2B (b), and 3 (a). Experiments 2A and 3 used a name-learning task, where all 28 individuals in the participant's respective training condition (set A vs. set B) were paired with a four-letter name. Across 7 rounds of study and test phases, participants were instructed to observe each face (presented for 3000 ms with the name) and type the name in a response box when prompted (free-recall test after each study round). Experiment 2B used a similar training task, but it was changed to remove the names, to create training that was perceptually based. Here, participants were instead told that they would see 28 images that would have square probes appear on them, with a random color (blue vs. green) and number of squares (1, 2, 3, or 4). Because the names in Experiments 2A and 3 stayed the same across all rounds, the square probe color/number assigned to each face was also constant across rounds in Experiment 2B. All other timing/exposure parameters for Experiment 2B training were the same as Experiments 2A and 3. See the online article for the color version of this figure.

attractive do you find this individual?" and responded on the 9-point scale described above. On familiarity ratings, participants were asked "How familiar do you find this individual?", and responded on the 9-point scale described above. For the familiarity ratings, participants were only told to rate familiarity based on whether they thought they saw the face at all, before that point in the study session (i.e., they were not explicitly told to reference the training task for giving their familiarity ratings). Within each of the four different rating blocks, stimulus presentation was completely randomized.

## Results and Discussion

**Analysis strategy.** We used the same mixed-effects modeling strategy as Experiment 1.

**Training performance (name-learning task).** We analyzed both accuracy and response times (RTs) using a Training Condition (2 [between]: set A, set B) × Testing Block (7 [within]) fixed-effects structure. To normalize the reaction time (RT) distribution and reduce the impact of outliers, all incorrect RTs were excluded, and the remaining correct RTs were $\log_{10}$-transformed. Confirming the effectiveness of the training, the analysis showed that participants responded progressively faster, $F(6, 388.02) = 111.12$, $p < .001$, and more accurately, $F(6, 124.53) = 352.28$, $p < .001$, across successive test blocks[2] (also see supplementary materials [Figure S2] for more details).

**Attractiveness ratings.** Attractiveness ratings were analyzed using a mixed-effects model with a Training Type (2 [within]: trained, untrained) × Target Type (2 [within]: individual, morph) fixed-effects structure.[3]

Figure 3a displays the attractiveness results. There was strong evidence for a Training Type × Target Type interaction, $F(1, 5995.00) = 25.14$, $p < .001$. Follow-up tests demonstrated that untrained morphs were judged as more attractive than untrained individuals, although this effect was not significant, $b = 0.05$, $t(210.30) = 0.58$, $CI_{95\%}$ [−0.11 0.21], *n*s. This is consistent with the notion that with no exemplar learning, there should be minimal preference for the morph (if any at all). Confirming the key prediction, trained morphs were judged as less attractive than trained individuals, $b = -0.47$, $t(210.30) = -5.84$, $CI_{95\%}$ [−0.63 −0.31], $p < .001$. Thus, we observed robust evidence for the UiA effect (rather than a BiA effect) between trained individuals and morphs. Furthermore, we also found that trained morphs were still judged as more attractive when compared to untrained morphs, $b = 0.27$, $t(560.30) = 3.09$, $CI_{95\%}$ [0.10 0.44], $p = .002$. This aligns with our expectation of a relative decrease in preference for morphs of familiar individuals, rather than an absolute dislike of such morphs. Finally, both main effects were significant. The main effect of Training Type, $F(1, 90.60) = 94.79$, $p < .001$, reflected overall higher ratings for trained targets compared to untrained targets, whereas the main effect of Target Type, $F(1, 73.80) = 11.69$, $p = .001$, demonstrated overall higher ratings for individuals compared to morphs.

**Familiarity ratings.** We analyzed familiarity ratings in the same way as attractiveness ratings, using a mixed-effects model with a Training Type (2 [within]: trained, untrained) × Target Type (2 [within]: individual, morph) fixed-effects structure.[4]

Figure 3b displays the familiarity results. Like attractiveness, we observed strong evidence for all effects. The main effect of Training Type, $F(1, 73.01) = 83.04$, $p < .001$, demonstrated that trained targets were judged as more familiar than untrained targets, and the main effect of Target Type, $F(1, 73.00) = 19.80$, $p < .001$, showed that individuals were judged as more familiar than morphs. Critically though, we also detected a Training Type × Target Type interaction, $F(1, 73.01) = 14.25$, $p < .001$. This interaction revealed a greater difference between trained and untrained individuals, $b = 2.07$, $t(73.00) = 8.55$, $CI_{95\%}$ [1.59 2.55], $p < .001$, compared to trained and untrained morphs, $b = 1.25$, $t(73.00) = 7.09$, $CI_{95\%}$ [0.90 1.60], $p < .001$. Consequently, trained individuals were judged to be more familiar than trained morphs, $b = 1.39$, $t(73.00) = 6.32$, $CI_{95\%}$ [0.95 1.82], $p < .001$. Untrained individuals were also seen as somewhat more familiar than untrained morphs, $b = 0.57$, $t(73.00) = 2.12$, $CI_{95\%}$ [0.04 1.10], $p = .04$, but this difference was smaller than the difference between trained individuals and trained morphs.

Note that the familiarity ratings for Experiment 2A were overall greater than those from Experiment 1 (i.e., Experiment 1 familiarity ratings fell mostly between 2 and 3, whereas Experiment 2A familiarity ratings were mostly between 5 and 9). Since strong learning only occurred in Experiment 2A (not Experiment 1), there are a couple of factors to consider. First, since individual exemplars have much stronger memory traces after training, this would substantially boost familiarity for trained individuals and their morphs (as described previously). Second, in Experiment 2A, familiarity was measured after all attractiveness ratings, in order to limit participants' exposure to untrained exemplars before they rated attractiveness. This would explain why participants rated "novel" untrained individuals and morphs as generally more familiar in Experiment 2A, because they did see those individuals once when rating attractiveness in earlier blocks. Finally, in Experiment 2A, we also observed that untrained individuals were rated as slightly more familiar than untrained morphs. This is likely attributable to the fact that learning on the individual exem-

[2] In Experiment 2A, on RTs, the maximal random slope model did not converge, so the random intercept model was set for fixed-effects testing (AIC = −7413.58, BIC = −7288.47). We observed a main effect of Testing Block on RTs, $F(6, 388.02) = 111.12$, $p < .001$, such that both set A and set B participants logged faster RTs over successive rounds of the free-recall task (with performance beginning to level out around block 5). Here, we did not detect a main effect of Training Condition, $F(1, 72.53) = .51$, *n*s, nor a Training Condition × Testing Block interaction, $F(6, 388.02) = 1.67$, *n*s. On recall accuracy, the maximal random slope model was a significantly better fit than the maximal random intercept model, $\chi^2(26) = 378.10$, $p < .001$, so the random slope model was set for fixed-effects testing (AIC = 5985.68, BIC = 6310.88). We again found a main effect of Testing Block, $F(6, 124.53) = 352.28$, $p < .001$, where both set A and set B participants improved their performance over successive rounds of the free-recall task. Specifically, participants started at approximately 33% correct responses in block 1, but improved to about 98% by block 7 (and similar to RTs, performance began to plateau around block 5). We did not detect a main effect of Training Condition, $F(1, 72.01) = 1.22$, *n*s, nor any evidence for a Training Condition × Testing Block interaction, $F(6, 124.53) = .40$, *n*s.

[3] On attractiveness ratings for Experiment 2A, there was no significant difference in model fit between the maximal random intercept and slope models, $\chi^2(6) = 3.50$, *n*s, so the random intercept model was set for fixed-effects testing (AIC = 26093.67, BIC = 26154.28).

[4] With familiarity ratings in Experiment 2A, the maximal random slope model was a significantly better fit than the maximal random intercept model, $\chi^2(6) = 26.14$, $p < .001$, so the random slope model was set for fixed-effects testing (AIC = 24869.42, BIC = 24970.44).

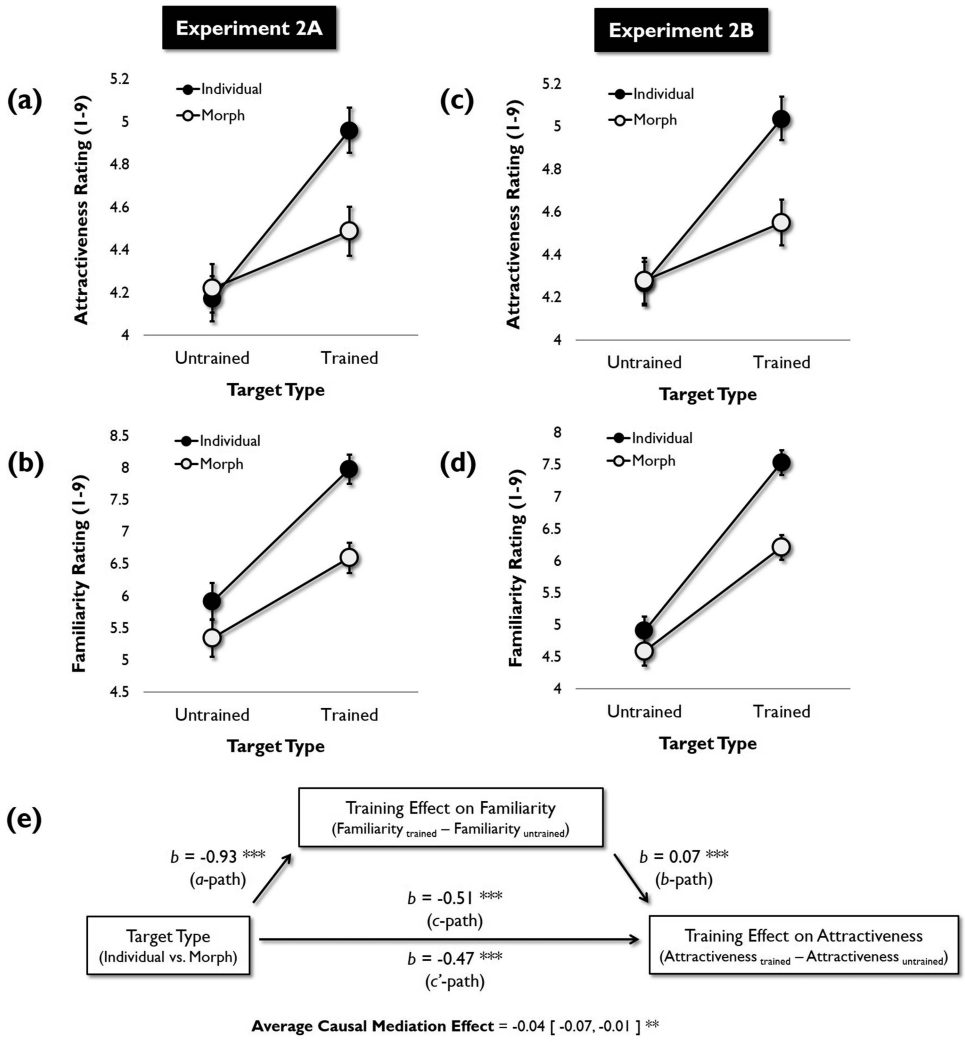


*Figure 3.* Results for attractiveness ratings (a and c), familiarity ratings (b and d), and multilevel mediation (e) across Experiments 2A and 2B. We observed an ugliness-in-averageness (UiA) effect after training in Experiment 2A, such that trained morphs were judged as less attractive than trained individuals (a). For familiarity ratings in Experiment 2A, all effects were significant, and the interaction was driven by the fact that there was a greater increase in familiarity for individuals after training, compared to morphs (b). Experiment 2B replicated the pattern of attractiveness ratings from Experiment 2A, where trained morphs were judged as less attractive than trained individuals (c). All familiarity effects were again significant in Experiment 2B, where the interaction was driven by a greater familiarity increase for individuals after training, compared to morphs (d). Multilevel mediation across both Experiments 2A and 2B demonstrated that the relationship between Target Type (individual vs. morph) and the training effect on attractiveness ratings ($\text{attractiveness}_{\text{trained}} - \text{attractiveness}_{\text{untrained}}$) was mediated by the training effect on familiarity ratings ($\text{familiarity}_{\text{trained}} - \text{familiarity}_{\text{untrained}}$) (e). Error bars represent $\pm 1$ *SEM*. $^{**}$ $p \leq .01$, $^{***}$ $p \leq .001$.

plars gave participants a greater sense of familiarity for that specific "face space" (compared to the other novel morph face set).

In sum, participants judged both trained individuals and trained morphs as more familiar than their untrained counterparts, but this effect was especially amplified for the individuals.

## Experiment 2B

To review, Experiment 1 demonstrated that a traditional BiA effect occurs with weak learning of exemplars in the context of many new face stimuli. Experiment 2A revealed that brief periods of training using a name-learning task generates a mere exposure effect for those trained individuals. This training also elicits a UiA effect, where trained morphs are judged as *less* attractive than trained individuals.

In Experiment 2B, we investigated a different type of training. According to memory frameworks, the mechanisms for eliciting the UiA effect should involve generic stimulus familiarity, as would be the case with low-level visual cues. Indeed, much pre-

vious work in face memory has focused on its sensory aspects, particularly on lower-level changes in visual responses to familiar and unfamiliar faces (Bobes et al., 2013; Buttle & Raymond, 2003; Davies-Thompson, Newling, & Andrews, 2013; Natu & O'Toole, 2011; Visconti di Oleggio Castello & Gobbini, 2015; Yovel & Belin, 2013). Thus, on this view, the UiA effect should occur even if learning is kept only to its "pure" perceptual aspects (without any name information), as was the case in Experiment 2A. Although face–name pairs are frequently used to examine identity-specific memory (e.g., Guo, Voss, & Paller, 2005; Schweinberger, Pickering, Burton, & Kaufmann, 2002; Verosky, Todorov, & Turk-Browne, 2013; Zeineh, Engel, Thompson, & Bookheimer, 2003), we wanted to replicate the effects from Experiment 2A using a pure perceptual training task. This would ensure that the UiA effect is not restricted to the face-name learning task, which may involve more emphasis on identity-level information.

We addressed this in Experiment 2B by changing the training to a perceptual-tracking task without names. Participants were exposed to the same faces from Experiment 2A (in either set A or set B) over similar durations, but they instead had to detect and recall blue and green square probes that randomly appeared on each image. If the UiA effect requires any name-based familiarity on the social identity for trained individuals, then the effect should dissipate in Experiment 2B (because the perceptual-tracking task would not pair names with trained faces). If the UiA effect instead only requires visual familiarity for trained individuals, we should observe similar effects on attractiveness in Experiment 2B (since participants are still receiving the same amount of exposure to each of those faces during training).

## Method

**Participants.** One hundred twenty-eight UCSD undergraduates participated for course-credit, and all participants signed consent forms approved by the UCSD HRPP. To plan our sample size in Experiment 2B, we conducted an a priori power analysis partly based on the effect sizes from Experiments 1 and 2A (we used a slightly lower effect size estimate of $f = 0.12$, given the changes to the training task). When implementing this analysis according to the design of Experiment 2B in GPower (version 3.1.9.2; Faul et al., 2007), to achieve 85–90% power, this forecasted a target range for $n$ at 119–137 participants (two-tailed test at $\alpha = .05$ and nonsphericity correction $\varepsilon = 1$).

**Materials.** All stimuli and materials were the same as Experiment 2A.

**Design and procedure.** Our main changes focused on the training task we used. Figure 2b shows the main revisions to this task in Experiment 2B. Participants still had to progress through 7 rounds of the free-recall task on the 28 individuals in their randomly assigned training set (set A or B). However, the type of recall they performed at the test phase during each round was different. Instead of recalling names, participants were instructed that they would have to recall "both the color and number of either blue or green square probes that would randomly appear on the different images" (no names were presented with the faces). During each study phase presentation (3000 ms for each image), 200 ms blue or green square probes would then appear at random intervals, and participants were tasked with remembering both the color and number of squares that appeared on the face. Each face was assigned to a constant color (either blue or green) and number (between 1 and 4) of square probes, and this color-number assignment did not change across successive rounds of training (similar to the names used in Experiment 2A). All attractiveness and familiarity ratings after the training task were the same as Experiment 2A.

## Results and Discussion

**Analysis strategy.** Our analysis strategy was the same as Experiment 2A.

**Training performance (perceptual-tracking task).** Similar to Experiment 2A, we gauged participants' accuracy and RT performance over all 7 testing blocks during training. We structured this analysis according to a Training Condition (2 [between]: set A, set B) × Testing Block (7 [within]) fixed-effects design, on both accuracy and RTs. As before, all RTs were $\log_{10}$-transformed, after excluding error trials. We also analyzed accuracy and RT performance separately for both the color (blue vs. green) and number (between 1 and 4) of square probes that were assigned to each trained individual.

Once again, our training task was effective, since participants became progressively faster, $F$s $\geq$ 82.37, $p < .001$, and more accurate, $F$s $\geq$ 159.43, $p < .001$, over successive training rounds. Note that there were some less theoretically important effects between performance on color versus number recall, which we do not discuss here[5] (also see supplementary materials [Figure S3] for more details).

**Attractiveness ratings.** We analyzed participants' attractiveness ratings using a mixed-effects model with a Training Type (2 [within]: trained, untrained) × Target Type (2 [within]: individual, morph) fixed-effects structure.[6]

Figure 3c displays the attractiveness results. Most importantly, we found a Training Type × Target Type interaction, $F(1, 10370.80) = 39.54$, $p < .001$. Follow-up tests on this interaction revealed a similar UiA effect as Experiment 2A, with trained morphs judged as less attractive than trained individuals, $b = -0.48$, $t(222.10) = -6.05$, $CI_{95\%}$ [−0.65 −0.33], $p < .001$. Untrained morphs were rated as more attractive than untrained individuals, but not significantly so, $b = 0.01$, $t(222.10) = 0.15$, $CI_{95\%}$ [−0.15 0.17], *n*s. Also similar to Experiment 2A, trained morphs were still judged as more attractive when compared to untrained morphs, $b = 0.27$, $t(768.50) = 3.98$, $CI_{95\%}$ [0.14 0.41], $p < .001$. We also observed a mere exposure effect because trained individuals were judged more attractive than untrained individuals, $b = 0.77$, $t(239.10) = 15.19$, $CI_{95\%}$ [0.67 0.87], $p < .001$. Main effects of both Training Type, $F(1, 151.80) = 132.67$, $p < .001$,

[5] In Experiment 2B, on RTs, maximal random slope models were a significantly better fit than maximal random intercept models for both color, $\chi^2(26) = 168.21$, $p < .001$ (AIC = −14732.81, BIC = −14398.80), and number, $\chi^2(26) = 132.44$, $p < .001$ (AIC = −335.35, BIC = −12.91), so they were set for fixed-effects testing. On accuracy, maximal random slope models would not converge, so maximal random intercept models were set for fixed-effects testing on both color (AIC = 30258.18, BIC = 30396.39) and number (AIC = 31539.63, BIC = 31677.85).

[6] On attractiveness ratings for Experiment 2B, there was no significant difference in model fit between the maximal random intercept and slope models, $\chi^2(6) = 11.33$, *n*s, so the random intercept model was set for fixed-effects testing (AIC = 45412.52, BIC = 45478.07).

and Target Type, $F(1, 127.40) = 11.48$, $p < .001$, showed that trained targets were judged as more attractive overall (compared to untrained targets), and individuals were judged as more attractive overall (compared to morphs).

**Familiarity ratings.** We tested familiarity ratings with a similar method to the attractiveness ratings, using a mixed-effects model with a Training Type (2 [within]: trained, untrained) × Target Type (2 [within]: individual, morph) fixed-effects structure.[7]

Figure 3d displays the familiarity results. All effects were significant. First, we observed strong evidence for a Training Type × Target Type interaction, $F(1, 127.01) = 36.55$, $p < .001$. This interaction revealed the expected effect that trained individuals were rated the most familiar, compared to untrained individuals, $b = 2.63$, $t(127.00) = 13.74$, $CI_{95\%}$ [2.25 3.00], $p < .001$, trained morphs, $b = 1.33$, $t(127.00) = 10.64$, $CI_{95\%}$ [1.08 1.57], $p < .001$, and untrained morphs, $b = 2.95$, $t(127.00) = 14.25$, $CI_{95\%}$ [2.54 3.36], $p < .001$. Crucially though, this interaction yielded a similar pattern to Experiment 2A, where the difference in familiarity ratings between trained individuals and trained morphs was more amplified, compared with the smaller difference between untrained individuals and untrained morphs, $b = 0.33$, $t(127.00) = 2.05$, $CI_{95\%}$ [0.01 0.64], $p = .04$. Main effects for both Training Type, $F(1, 127.00) = 195.80$, $p < .001$, and Target Type, $F(1, 127.00) = 50.32$, $p < .001$, also demonstrated that trained targets were rated as more familiar overall, and individuals were rated as more familiar than morphs.

Generally, these results replicated the familiarity findings from Experiment 2A. Also similar to Experiment 2A, familiarity ratings in Experiment 2B fell mostly between 5 and 9, and untrained individuals were still judged as more familiar than untrained morphs. This is presumably because learning on the individual exemplars gave participants a greater sense of familiarity for that specific "face space" (rather than the novel morph face set).

**Multilevel mediation across Experiments 2A and 2B.** We used the same multilevel mediation procedure as Experiment 1, but with some important changes (because of updates in the data and experiment structure in Experiments 2A and 2B). First, we included data from both Experiments 2A and 2B in one multilevel mediation model, given that these two experiments were very similar and analyzing both data sets in one model allowed for more powerful effect estimates (but we report statistics from mediation analyses on the individual experiments later in the main text and footnotes of this section). Second, note that our main prediction in Experiments 2A and 2B is that the relationship between target type (individual vs. morph) and the training effect on attractiveness ratings is mediated by the training effect on familiarity ratings. In other words, the way in which training impacts attractiveness ratings (for individuals vs. morphs) should be driven by how that training impacts familiarity ratings (for individuals vs. morphs). To address this, we created a new multilevel mediation model where our main predictor was target type (individual vs. morph), our main DV was the training effect on attractiveness (i.e., difference score between attractiveness of trained targets and attractiveness of untrained targets), and our mediator was the training effect on familiarity (i.e., difference score between familiarity of trained targets and familiarity of untrained targets). As before, mixed-effects models were constructed for each of the mediation paths, using by-participant random effects parameters. All simulations from the *mediation* package in R were based on 1,000 samples per estimate, after which quasi-Bayesian confidence intervals were calculated around the average total, direct, and causal mediation effects.

Figure 3e shows the mediation results. We observed convincing evidence for mediation. Target type was a significant predictor of the training effect on familiarity ratings (*a*-path: $b = -0.93$, $t(201.00) = -7.11$, $p < .001$), and this familiarity training effect was a significant predictor of the attractiveness training effect (*b*-path: $b = 0.07$, $t(338.70) = 4.73$, $p < .001$). When controlling for the familiarity training effect (*c'*-path), the original *t*-value estimate of target type on the attractiveness training effect (*c*-path: $b = -0.51$, $t(201.00) = -10.20$, $p < .001$) was reduced but still significant (*c'*-path: $b = -0.47$, $t(221.30) = -9.07$, $p < .001$), while familiarity was also significant ($b = 0.04$, $t(350.00) = 2.70$, $p = .007$). Finally, this further demonstrated that the average causal mediation effect was also significant ($b = -0.04$, $CI_{95\%}$ [−0.07 −0.01], $p \leq .01$).

Note that when we conducted additional analyses using similar multilevel mediation models for each individual experiment, we observed similar results.[8] Specifically, the parallel average causal mediation effect was significant in Experiment 2A ($b = -0.05$, $CI_{95\%}$ [−0.10 −0.01], $p \leq .01$) and marginal in Experiment 2B ($b = -0.03$, $CI_{95\%}$ [−0.07 0.005], $p = .09$).

In sum, the multilevel mediation analysis on Experiments 2A and 2B showed clear evidence that the relationship between target type (individual vs. morph) and the attractiveness training effect was mediated by the familiarity training effect.

## Experiment 3

Experiments 2A and 2B established that repetition of individual faces generates a standard mere exposure effect, while also generating an ugliness-in-averageness (UiA) effect for morphs of trained faces. We observed similar findings using both a name-learning task (Experiment 2A) and perceptual-tracking task (Experiment 2B). These results not only offer a major qualification to the classic beauty-in-averageness (BiA) effect, but they also suggest that generic familiarity is sufficient for eliciting a UiA effect (as would be the case with low-level visual cues; Natu & O'Toole,

[7] With familiarity ratings in Experiment 2B, the maximal random slope model was a significantly better fit than the maximal random intercept model, $\chi^2(6) = 33.37$, $p < .001$, so the random slope model was set for fixed-effects testing (AIC = 44133.39, BIC = 44242.63).

[8] On the multilevel mediation for Experiment 2A, target type was a significant predictor of the training effect on familiarity ratings (*a*-path: $b = -0.81$, $t(73.00) = -3.78$, $p < .001$), and the familiarity training effect was a significant predictor of the attractiveness training effect (*b*-path: $b = 0.09$, $t(146.00) = 3.90$, $p < .001$). When controlling for the familiarity training effect (*c'*-path), the original *t*-value estimate of target type on the attractiveness training effect (*c*-path: $b = -0.52$, $t(73.00) = -7.14$, $p < .001$) was reduced but still significant (*c'*-path: $b = -0.47$, $t(78.62) = -6.22$, $p < .001$), while familiarity was also significant ($b = 0.06$, $t(131.25) = 2.69$, $p = .008$). On Experiment 2B, similarly, target type was a significant predictor of the training effect on familiarity ratings (*a*-path: $b = -1.00$, $t(127.00) = -6.05$, $p < .001$), and the familiarity training effect was a significant predictor of the attractiveness training effect (*b*-path: $b = 0.06$, $t(214.93) = 3.20$, $p = .002$). When controlling for the familiarity training effect (*c'*-path), the original *t*-value estimate of target type on the attractiveness training effect (*c*-path: $b = -0.50$, $t(127.00) = -7.52$, $p < .001$) was reduced but still significant (*c'*-path: $b = -0.47$, $t(141.45) = -6.75$, $p < .001$), while familiarity was trending but not quite significant ($b = 0.03$, $t(217.57) = 1.56$, *ns*).

2011). Importantly, the decline in attractiveness for morphs of familiar individuals was relative—they were still more attractive than untrained individuals. Note that these effects were obtained with relatively minor amounts of exposure, demonstrating that the UiA effect does not require extensive expertise. Theoretically, these results are consistent with predictions from modern memory frameworks, which emphasize the critical role of the amount of learning in familiarity (and thus, preference) for exemplars and their blends.

With Experiment 3, we wanted to further examine the underlying mechanism driving the UiA effect. Recall that in the Introduction, we outlined three alternative patterns for possible results after exemplar training. First, the *additive* prediction would posit that preferences from mere exposure and blending should combine in a positive fashion, making morphs of familiar individuals especially attractive. This prediction seems most intuitive when assuming these two manipulations enhance liking via separate and independent mechanisms. However, both Experiments 2A and 2B offer clear evidence against this idea, because morphs of trained individuals were judged as less attractive than trained individuals themselves (UiA effect). This leaves two other possibilities. First, a *mismatch* account suggests that encountering a blend of two familiar individuals causes a cognitive conflict (Arnal & Giraud, 2012; Dreisbach & Fischer, 2015; Hsu et al., 2005), perhaps not unlike conflict triggered by bistable figures (Kornmeier & Bach, 2012; Topolinski, Erle, & Reber, 2015). The negative affect generated from this conflict is then misattributed to subsequent ratings, causing the relative unattractiveness of trained morphs. Second, per our *familiarity* account, the UiA effect is driven by a relative reduction in familiarity for morphs of trained exemplars.

Experiments 2A and 2B offer some preliminary evidence in favor of our memory-based familiarity account. If cognitive mismatch played a primary role in the UiA effect, we would expect that trained morphs would be judged as not only less attractive than trained individuals, but also less attractive than untrained morphs, but this is not what we observed. Instead, blends of well-learned individuals generated familiarity and preference values in-between actually exposed individuals and novel individuals, which would be predicted by memory frameworks (Jones & Jacoby, 2001; Kelley & Wixted, 2001). Also, the multilevel mediation analysis on Experiments 2A and 2B showed that the training effect on familiarity ratings mediated the relationship between target type (individual vs. morph) and the training effect on attractiveness ratings.

However, Experiments 2A and 2B do not offer a definitive test between the mismatch and familiarity accounts. In these studies, the trained morphs could potentially generate both high conflict and high familiarity, given that they blend two highly familiar individual exemplars. We used Experiment 3 to address this issue with a simple change to Experiments 2A and 2B. In Experiments 2A and 2B, recall that all morphs were 100% within-set, meaning that morphs would either be A-A or B-B, but never A-B (see Figure S1 in supplementary materials). In Experiment 3, we created new versions of the sets (once again based on attractiveness ratings from a previous study; Halberstadt et al., 2013) that instead used cross-set A-B morphs, so the two individuals composing each morph were always in different sets. With this setup, the cross-set morphs should yield familiarity values in-between that of trained and untrained individuals (since they are composed of one trained and untrained individual). Note that this would be similar to the within-set morphs from Experiments 2A and 2B (which showed familiarity ratings in-between untrained individuals and trained individuals; see Figures 3b and 3d), but if anything, within-set morphs should yield somewhat greater familiarity than cross-set morphs (yet still in-between trained and untrained individuals, since they are instead composed of two trained individuals).

Critically though, with the cross-set morphs, any conflict from blending two highly familiar individual exemplars would be reduced or eliminated. Therefore, if conflict is the driving mechanism for the UiA effect, cross-set morphs should now appear *more* attractive than trained individuals (thus, a standard BiA effect). Essentially, this cross-set morph design directly pits the two remaining theories against one another:

> $H_1$: This assumes that the UiA effect for trained morphs is driven by a *mismatch* (conflict) between two strongly learned individuals. This conflict generates negative affect, which leads to lower attractiveness ratings for those morphs. If so, cross-set morphs should be rated as *more* attractive than trained individuals (thus, a standard BiA effect). Since the cross-set morphs contain one trained and one untrained identity, any such conflict that would emerge from blending two known individuals would be substantially reduced (and any UiA effect should dissipate). Further, without such conflict, cross-set morphs would presumably be judged more attractive than trained individuals from the usual benefits of blending faces.

> $H_2$: This assumes that the UiA effect for trained morphs is driven by a relative decrease in *familiarity* of two strongly learned individuals. The specific trained individuals receive increased attractiveness ratings than morphs because they are exact replicates of items from training, whereas the morph is less similar to the trained set. If so, cross-set morphs should be rated as *less* attractive than trained individuals (thus, a UiA effect). Because the cross-set morphs contain one trained and one untrained identity, they should still be judged as relatively less familiar (and less attractive) than the trained individuals.

## Method

**Participants.** One hundred fifty-one UCSD undergraduates participated for course-credit, and all participants signed consent forms approved by the UCSD HRPP. To plan our sample size in Experiment 3, we conducted an a priori power analysis partly based on the effect sizes from Experiments 1, 2A, and 2B (once again using an effect size estimate of $f = 0.12$). We applied this analysis in GPower (version 3.1.9.2; Faul et al., 2007), according to the design of Experiment 3. To achieve 85–90% power, this forecasted a target range for $n$ at 141–163 participants (two-tailed test at $\alpha = .05$ and nonsphericity correction $\varepsilon = 1$).

**Materials.** We made only one change to the stimulus setup from Experiments 2A and 2B. Although both sets A and B each still contained 28 individuals and 14 morphs each, they were reorganized (once again based on attractiveness ratings from a previous study; Halberstadt et al., 2013) so that the two individuals composing each morph were always in different sets. Therefore, in Experiment 3, all morphs were cross-set A-B morphs that blended one trained and one untrained individual (rather than the within-set

A-A and B-B morphs used in Experiments 2A and 2B; see Figure S1 in supplementary materials).

**Design and procedure.** We used the same name-learning task as Experiment 2A (see Figure 2a).

## Results and Discussion

**Analysis strategy.** Our analysis strategy was the same as Experiments 2A and 2B.

**Training performance (name-learning task).** As before, we examined participants' accuracy and RT performance over all 7 testing blocks during training. This analysis was structured according to a Training Condition (2 [between]: set A, set B) × Testing Block (7 [within]) fixed-effects design, on both accuracy and RTs. All RTs were $\log_{10}$-transformed after excluding error trials.

Once again, our training task was effective, since participants became progressively faster, $F(6, 855.43) = 247.36$, $p < .001$, and more accurate, $F(6, 307.55) = 519.55$, $p < .001$, over successive rounds of the free-recall task[9] (also see supplementary materials [Figure S4] for more details).

**Attractiveness ratings.** Given that the cross-set morphs used in Experiment 3 were neither 100% trained nor untrained, we analyzed attractiveness ratings using a mixed-effects model with Target Type (3 [within]: morph, trained individual, untrained individual) as the only fixed-effects factor.[10]

Figure 4a displays the attractiveness results. We detected a strong main effect of Target Type, $F(2, 150.00) = 111.13$, $p < .001$. Critically, a UiA effect still emerged, such that morphs were rated as less attractive than trained individuals, $b = -0.55$, $t(150.00) = -9.07$, $CI_{95\%}$ [−0.68 −0.43], $p < .001$. Interestingly, even though participants did not rate the morphs as more familiar than the untrained individuals (see next section), they still rated the morphs as relatively more attractive, $b = 0.13$, $t(150.00) = 2.22$, $CI_{95\%}$ [0.01 0.24], $p = .03$. And as expected, we replicated the mere exposure effect, where trained individuals were judged as more attractive than untrained individuals, $b = 0.68$, $t(150.00) = 14.69$, $CI_{95\%}$ [0.59 0.77], $p < .001$.

**Familiarity ratings.** We analyzed familiarity ratings in the same way as attractiveness, using a mixed-effects model with Target Type (3 [within]: morph, trained individual, untrained individual) as the only fixed-effects factor.[11]

Figure 4b displays the familiarity results. We observed a clear main effect of Target Type, $F(2, 150.03) = 132.79$, $p < .001$. Trained individuals were judged as more familiar than both untrained individuals, $b = 2.17$, $t(150.00) = 12.18$, $CI_{95\%}$ [1.82 2.52], $p < .001$, and morphs, $b = 2.27$, $t(150.00) = 15.40$, $CI_{95\%}$ [1.98 2.56], $p < .001$. Note that there was also no difference when comparing mean familiarity ratings between morphs and untrained individuals, $b = -0.10$, $t(150.00) = -0.60$, $CI_{95\%}$ [−0.43 0.23], *n*s, though as discussed below, familiarity still played a role in the attractiveness ratings of those targets.

**Multilevel mediation.** We built multilevel mediation models in Experiment 3 using a similar procedure as the previous studies, but with one important change. Because we used cross-set morphs in Experiment 3 that were neither 100% trained nor untrained, this was collapsed into one three-level factor for Training Target Type (3 [within]: morph, trained individual, untrained individual). Note that treatment variables with more than two levels need to be handled differently than binary treatment variables in multilevel mediation (Imai, Keele, & Tingley, 2010). This can be done by creating separate mediation models with different treatment values, compared across the same control value. Therefore, for Experiment 3, we created two separate multilevel mediation models. The first model compared trained individuals with untrained individuals, and the second model compared trained individuals to morphs. With both models, our main predictor was training (trained individuals vs. untrained individuals in Model 1 [$M_1$]; trained individuals vs. morphs in Model 2 [$M_2$]), our main DV was attractiveness ratings, and our mediator was familiarity ratings.

Figure 4c shows a summary of the mediation results.[12] After testing the total, indirect, and average causal mediation effects, we detected evidence for mediation in both models. For $M_1$ (comparing trained individuals vs. untrained individuals), the total effect ($b = 0.68$, $CI_{95\%}$ [0.59 0.78], $p < .01$), average direct effect ($b = 0.53$, $CI_{95\%}$ [0.42 0.66], $p < .01$), and average causal mediation effect ($b = 0.15$, $CI_{95\%}$ [0.06 0.23], $p < .01$) were all highly significant. We saw similar results with $M_2$ (comparing trained individuals vs. morphs), with a significant total effect ($b = 0.55$, $CI_{95\%}$ [0.43 0.67], $p < .01$), average direct effect ($b = 0.39$, $CI_{95\%}$ [0.21 0.56], $p < .01$), and average causal mediation effect ($b = 0.16$, $CI_{95\%}$ [0.03 0.29], $p = .01$). In short, familiarity mediated the relationship between training and attractiveness (both for trained

[9] In Experiment 3, on RTs, maximal random slope models would not converge, so maximal random intercept models were set for fixed-effects testing (AIC = −18806.29, BIC = −18668.33). A main effect of Testing Block, $F(6, 855.43) = 247.36$, $p < .001$, showed that participants got progressively faster over successive training rounds. We observed no evidence for a main effect of Training Condition, $F(6, 149.89) = 1.19$, *n*s, nor a Training Condition × Testing Block interaction, $F(6, 855.43) = 0.43$, *n*s. On accuracy, the maximal random slope model was a significantly better fit than the maximal random intercept model, $\chi^2(26) = 861.33$, $p < .001$, so the random slope model was set for fixed-effects testing (AIC = 10404.80, BIC = 10761.50). We observed the expected main effect of Testing Block, $F(6, 307.55) = 519.55$, $p < .001$, where participants improved their recall throughout the task (starting at approximately 36% correct in block 1 and improving to about 98% correct by block 7, with performance beginning to level out at block 5). We also observed a marginal main effect of Training Condition, $F(1, 149.08) = 3.35$, $p = .07$, such that set A participants ($M_{acc} = 84.71\%$, $SD_{acc} = 7.79\%$) performed better than set B participants ($M_{acc} = 82.27\%$, $SD_{acc} = 8.60\%$) throughout the entirety of the memory task. The Training Condition × Testing Block interaction was not significant, $F(6, 307.55) = 0.82$, *n*s.

[10] On attractiveness ratings for Experiment 3, the maximal random slope model was a significantly better fit than the maximal random intercept model, $\chi^2(4) = 18.61$, $p < .001$, so the random slope model was set for fixed-effects testing (AIC = 52268.69, BIC = 52343.17).

[11] With familiarity ratings in Experiment 3, the maximal random slope model was a significantly better fit than the maximal random intercept model, $\chi^2(4) = 33.05$, $p < .001$, so the random slope model was set for fixed-effects testing (AIC = 51837.35, BIC = 51911.83).

[12] Training target type was a significant predictor of familiarity (*a*-path [$M_1$]: $b = 2.17$, $t(150.00) = 12.18$, $p < .001$; *a*-path [$M_2$]: $b = 2.27$, $t(150.00) = 15.40$, $p < .001$), and familiarity was a significant predictor of attractiveness (*b*-path [$M_1$]: $b = 0.18$, $t(193.46) = 10.49$, $p < .001$; *b*-path [$M_2$]: $b = 0.16$, $t(205.08) = 7.85$, $p < .001$). When controlling for familiarity, the original *t*-value estimate of training on attractiveness (*c*-path [$M_1$]: $b = 0.68$, $t(150.00) = 14.69$, $p < .001$; *c*-path [$M_2$]: $b = 0.55$, $t(150.00) = 9.07$, $p < .001$) was reduced but still significant (*c'*-path [$M_1$]: $b = 0.53$, $t(171.81) = 8.65$, $p < .001$; *c'*-path [$M_2$]: $b = 0.39$, $t(200.57) = 4.38$, $p < .001$), while familiarity was also significant (*c'*-path [$M_1$]: $b = 0.07$, $t(203.77) = 3.56$, $p < .001$; *c'*-path [$M_2$]: $b = 0.07$, $t(254.25) = 2.52$, $p = .01$).

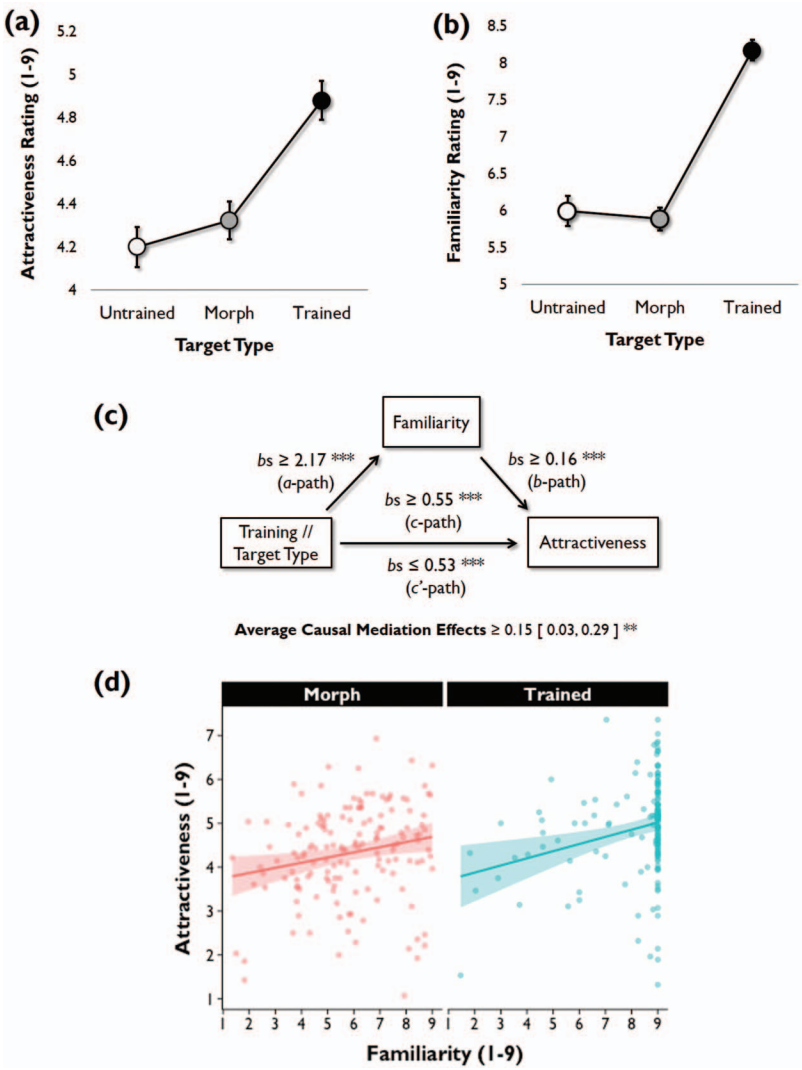


*Figure 4.* Attractiveness ratings (a), familiarity ratings (b), multilevel mediation results (c), and correlation analyses (d) in Experiment 3. We still observed an ugliness-in-averageness (UiA) effect after training using cross-set morphs (rather than the within-set morphs from Experiments 2A and 2B), such that morphs were judged as less attractive than trained individuals (a). Trained individuals were judged as more familiar than both untrained individuals and cross-set morphs (b). Multilevel mediation demonstrated that the relationship between training target type (trained individuals vs. cross-set morphs & trained individuals vs. untrained individuals) and attractiveness ratings were significantly mediated by familiarity (c). Separate correlation analyses within morphs (left plot in panel d) and trained individuals (right plot in panel d) showed significant positive correlations between familiarity and attractiveness. Linear fits are shown in each plot in panel d, along with 95% confidence interval bands. Error bars represent ±1 *SEM*. ** $p \leq .01$, *** $p \leq .001$. See the online article for the color version of this figure.

individuals vs. untrained individuals and trained individuals vs. morphs).

**Correlations by target type.** Finally, we also wanted to assess the relationship between attractiveness and familiarity *within* trained individuals and morphs (rather than comparing across them). In other words, are morphs that appear more familiar rated higher on attractiveness, compared to other morphs that appear relatively unknown? We investigated this by simply aggregating participants' mean attractiveness and familiarity ratings for morphs and trained individuals, then running separate Pearson ($r$) product-moment correlation tests within each target type.

Figure 4d shows the results of this analysis. Attractiveness and familiarity were positively correlated for both morphs, $r(149) = .20$, $CI_{95\%}$ [.05 .35], $p = .01$, and trained individuals, $r(149) = .24$, $CI_{95\%}$ [.09 .39], $p = .002$. This demonstrates that familiarity not only impacted attractiveness ratings across target types (i.e., morphs vs. trained individuals), but it also affected attractiveness within target types as well (i.e., more familiar morphs were more attractive than less familiar morphs).

Overall, Experiment 3 favored $H_2$ (*familiarity* account) over $H_1$ (*mismatch* account). Results from attractiveness and familiarity ratings, multilevel mediation, and correlational analyses all suggest that the UiA effect depends on the similarity of the morph to the exemplars. This idea assumes that increased exemplar learning leads to greater familiarity for those trained individuals. In turn, the "dip" in attractiveness ratings for trained morphs is actually due to the relative reduction of those familiarity cues (where trained individuals feel more familiar than trained morphs, since they are "pure" replicates of what was shown during the memory task).

## Experiment 4

To recap, Experiment 1 showed that a traditional BiA effect occurs with weak learning of individual exemplars. We also demonstrated that brief periods of training using both a name-learning task (Experiments 2A and 3) and perceptual-tracking task (Experiment 2B) generates a mere exposure effect for those trained individuals. Importantly, these memory tasks also produce a UiA effect, where trained morphs are judged as less attractive than trained individuals. Finally, we extended these findings in Experiment 3 using cross-set morphs, which showed that these results are driven by a relative reduction in familiarity cues between trained individuals and morphs—thus supporting the familiarity-driven (memory-based) framework for the UiA effect (over the additive and mismatch frameworks).

We used Experiment 4 to address two unanswered questions. One issue is a potential role of differences in task goals across the previous experiments. Recall that in Experiments 2A, 2B, and 3, we instructed participants to pair and memorize name or square information with different faces, whereas in Experiment 1, participants merely proceeded through all the faces to give ratings (no memorization required). It might be the case that these different tasks induced different goals while encoding the faces. The weak learning context in Experiment 1 may have biased participants toward a more global encoding strategy, because they did not have to actively engage with the stimuli (thus leading morphs to appear more familiar and attractive). In contrast, the strong learning contexts in Experiments 2A, 2B, and 3 may have encouraged a more specific encoding strategy, since the task requires more detailed memory on the individual exemplars (thus leading trained individuals to appear more familiar and attractive). On this account, our effects are not driven by the amount of exposure per se, but rather individuation of different face stimuli depending on the task at-hand (which is believed to increase differentiation by changing the structure of the stimulus space; e.g., McGugin, Tanaka, Lebrecht, Tarr, & Gauthier, 2011). A second issue is that in our previous experiments, we did not have any measures of more objective memory strength—only ratings of subjective familiarity. If our effects are indeed driven by memory processes, then differences in attractiveness between individuals and morphs should also be reasonably linked to objective performance in recognition of the face stimuli (i.e., "old/new" judgments), though we will return to the difference between familiarity and recognition judgments later.

To address these concerns, we made three main changes to the design from Experiment 2B (which used the perceptual-tracking task on trained and untrained faces). First, instead of dividing the faces into study sets, we varied the number of exposures for individual faces (i.e., 14 individual faces each at 0, 1, 3, or 7 exposures), with all participants receiving all levels of prior exposure as a within-subjects manipulation of training. Second, we changed the nature of the perceptual-tracking task such that no consistent information was paired with the faces—participants only had to remember general spatial locations for where blue/green squares were presented. With this version of the task, the exposure is completely passive and does not require any individuating information to be paired with the trained exemplars. Obviously, this also means that even at high levels of exposure in Experiment 4 (i.e., 7 exposures), the individual exemplars are going to be less strongly encoded than the more "active" exposures in previous experiments, and on our account, this should slightly decrease subjective familiarity for those individuals. Finally, we also had participants make speeded "old/new" judgments on all face stimuli, after they gave all their attractiveness and familiarity ratings. This allowed us to calculate objective measures of memory strength (i.e., proportion "old" judgments and response times).

### Method

**Participants.** One hundred UCSD undergraduates participated for course-credit, and all participants signed consent forms approved by the UCSD HRPP. As before, to plan our sample size in Experiment 4, we conducted an a priori power analysis based on the effect sizes from Experiments 1, 2A, 2B, and 3 (once again using an effect size estimate of $f = 0.12$). We generated this analysis in GPower (version 3.1.9.2; Faul et al., 2007), according to the updated design of Experiment 4. To achieve 85–90% power, this forecasted a target range for $n$ at 92–105 participants (two-tailed test at $\alpha = .05$ and nonsphericity correction $\varepsilon = 1$).

**Materials.** We used the same stimuli as the previous experiments, but study sets (A or B) were not used in Experiment 4. Instead, for each participant, individual faces were randomly assigned to one of four exposure levels during training as a within-subjects manipulation (see next).

**Design and procedure.** As with previous experiments, participants went through a training task and subsequently provided attractiveness and familiarity ratings for individual and morph faces. However, we made three main changes for the design in Experiment 4.

First, individual face stimuli were not divided into study sets (A or B). Instead, individual faces were randomly assigned to one of four exposure levels during the training task (i.e., 0, 1, 3, or 7 exposures), with 14 individual faces at each level. In turn, there were 154 exposures during the training task, which was divided into 7 blocks of 22 trials (see Figure 5).

Second, we also modified the perceptual-tracking task from Experiment 2B. Figure 5 displays the training task used in Experiment 4. During each trial, an individual face stimulus was presented for 3000 ms, along with random 200 ms blue or green square probes that would briefly appear at different locations on the images. At the end of each block, participants were asked to indicate whether they thought there were more blue/green squares on the left/right or upper/lower part of the images for that block (where the stems for square color and location were randomly selected across blocks). Importantly, the color, number, and location of square presentations was randomized across trials and stimuli (i.e., individual faces were *not* paired with a specific

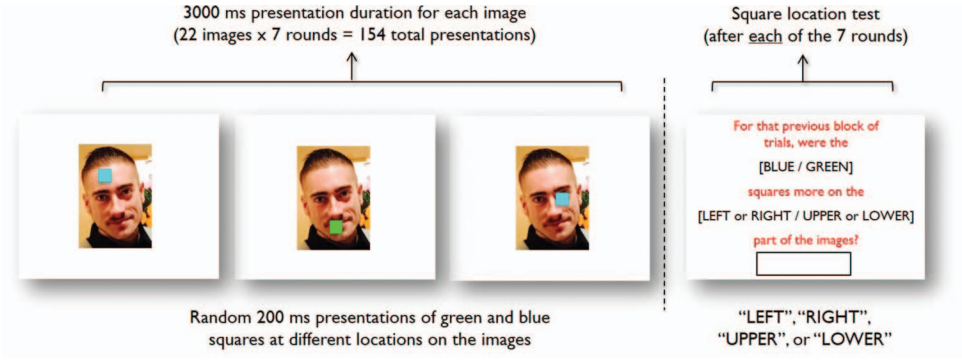


*Figure 5.* Design of the training task for Experiment 4, modified from Experiment 2B (see Figure 2b). Individual faces were randomly assigned to be exposed 0, 1, 3, or 7 times during training, which totaled 154 exposures ([14 individuals × 0 exposures] + [14 individuals × 1 exposure] + [14 individuals × 3 exposures] + [14 individuals × 7 exposures]). These 154 exposures were divided into 7 blocks with 22 trials each, with equal probabilities for each of the exposures being assigned to one of the blocks. During the task, individual faces would be presented for 3000 ms each, along with random 200 ms blue or green square probes that would briefly appear at different locations on the images. After each block of images, participants were asked to report the general spatial location (i.e., left/right or upper/lower, depending on the testing block) for a specific square color (i.e., blue or green, depending on the testing block). In the above figure, the bracketed text on the question screen indicates fields that would vary by testing block. See the online article for the color version of this figure.

color/number of squares, in contrast to Experiment 2B). This was done to ensure that exposure to the individuals was completely passive and to avoid having participants attach any individuating information to the trained faces.

Lastly, at the end of the experiment, we had participants make speeded "old/new" judgments on all face stimuli, in order to obtain measures of memory strength. More specifically, after the attractiveness and familiarity ratings in Experiment 4, participants progressed through all 84 face stimuli (56 individuals and 28 morphs; trial order randomized). They were instructed to judge, as quickly and accurately as possible, whether each face was "old or new" (using the A and L keys on the keyboard), and we specified that they should make their judgments according to what they saw during the blue/green square training task.

## Results and Discussion

**Analysis strategy.** We used the same mixed-effects modeling strategy as the previous experiments.

**Training performance (perceptual-tracking task).** Like the previous studies, we examined participants' accuracy and RT performance over all 7 testing blocks during training. We did this with Testing Block (7 [within]) as the only fixed-effect, on both accuracy and RTs.[13]

Once again, overall, our training task was effective. Participants responded progressively faster across successive rounds in the training task, as evident from a main effect of Testing Block on log10-transformed RTs, $F(6, 594.00) = 51.26$, $p < .001$. We did not observe any main effect of Testing Block on accuracy, $F(6, 637.00) = 0.30$, *n*s, but when collapsing across all 7 blocks, participants showed stable above-chance performance ($M_{acc} = 57.66\%$, $SD_{acc} = 18.57\%$), $t(99) = 4.13$, $CI_{95\%}$ [0.54 0.61], $p < .001$ (see supplementary materials [Figure S5] for more details).

**Attractiveness ratings.** To analyze attractiveness ratings in Experiment 4, we created a mixed-effects model with an Exposure Level (4 [within]: 0, 1, 3, 7 exposures) × Target Type (2 [within]: individual, morph) fixed-effects structure.[14]

Figure 6a displays the results for attractiveness. Critically, we observed an Exposure Level × Target Type interaction, $F(3, 7897.70) = 3.45$, $p = .016$. This showed that the attractiveness advantage for morphs (traditional BiA effect) emerged at low levels of exposure, but eventually dissipated and reversed with increasing exposure to the constituent faces. Specifically, morphs were judged as more attractive than their constituent individuals at the lower exposure levels, including no exposure (level 0), $b = 0.23$, $t(409.60) = 2.34$, $CI_{95\%}$ [0.04 0.42], $p = .02$, and weak exposure (level 1), $b = 0.23$, $t(409.60) = 2.34$, $CI_{95\%}$ [0.04 0.42], $p = .02$. However, this changed at higher exposure levels. At medium exposure (level 3), morphs were still judged as more attractive, but this difference did not reach significance, $b = 0.09$, $t(409.60) = 0.89$, $CI_{95\%}$ [−0.11 0.28], *n*s. And importantly, with high exposure (level 7), morphs were actually judged as less attractive than their constituent individuals (albeit this comparison did not reach significance), $b = -0.09$, $t(409.60) = 0.89$, $CI_{95\%}$ [−0.28 0.11], *n*s. This suggests the transition of a traditional BiA effect at lower exposure levels to a UiA effect at higher exposure levels. Note that we also observed a main effect of Exposure Level, $F(3, 336.80) = 8.63$, $p < .001$, which just showed that targets were judged as overall more attractive with greater exposure. The main effect of Target Type was marginal, $F(1, 99.30) = 2.79$,

[13] Note that participants were required to answer fewer questions during test in Experiment 4 compared to the previous studies (i.e., 7 questions in Experiment 4 vs. 196 questions in Experiments 2A, 2B, and 3). Thus, RTs were still log10-transformed before analysis, but RTs on both correct and incorrect trials were included.

[14] On attractiveness ratings for Experiment 4, the maximal random slope model would not converge, so the random intercept model was set for fixed-effects testing (AIC = 33968.03, BIC = 34059.50).

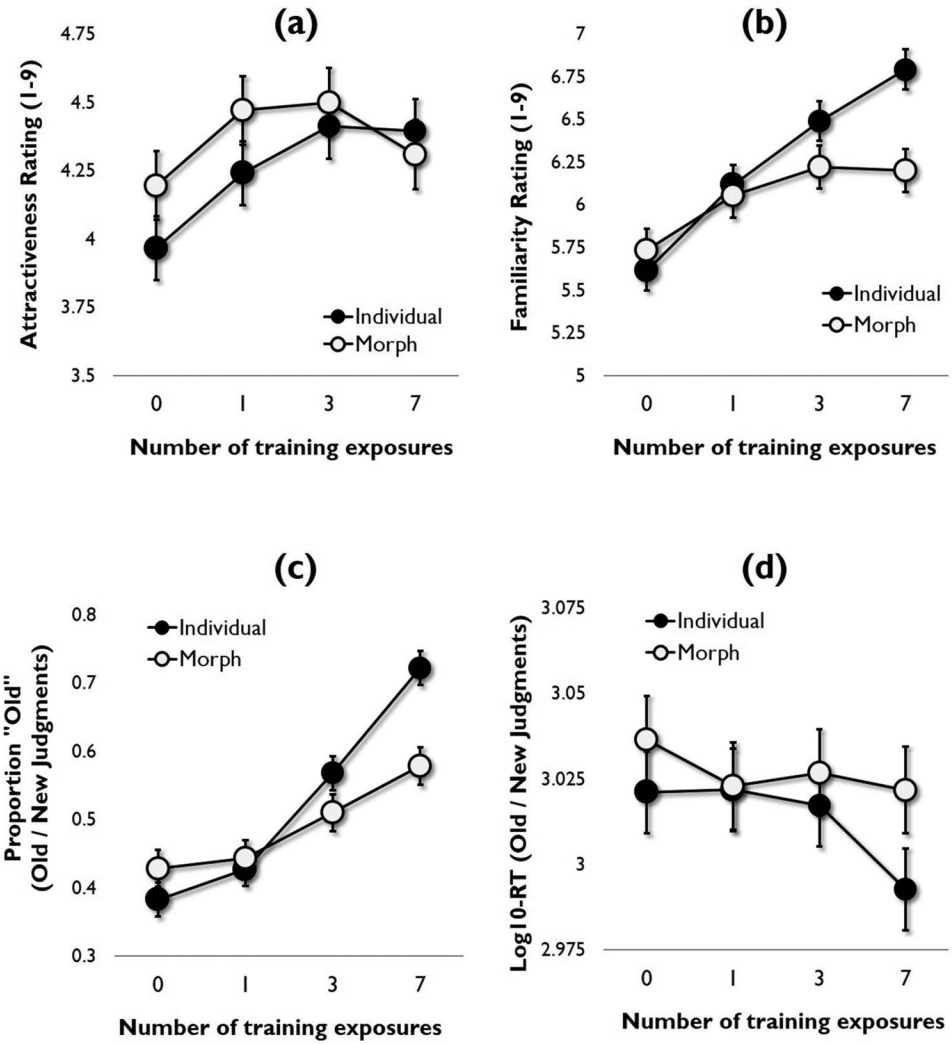


*Figure 6.* Attractiveness ratings (a), familiarity ratings (b), proportion "old" on old/new judgments (c), and $\log_{10}$-transformed RTs on old/new judgments (panel *d*) in Experiment 4. Morphs were judged as more attractive with no/weak exposure (levels 0 and 1) showing a traditional BiA effect, but individuals were judged more attractive with high exposure (level 7), trending toward a UiA effect (a). We did not observe any clear differences in familiarity between individuals and morphs with no/weak exposure (levels 0 and 1), but as the number of exposures increased (levels 3 and 7), individuals were judged to be more familiar (b). Proportion "old" judgments mirrored the subjective familiarity ratings, with better performance for individuals (relative to morphs) as the number of exposures increased from low (levels 0 and 1) to high (levels 3 and 7) (panel c). Participants showed faster old/new RTs to individuals, but similar to familiarity and proportion "old" judgments, this difference grew larger with more exposures (d). Error bars represent ±1 *SEM*.



$p = .098$, revealing that morphs appeared marginally more attractive than individuals when collapsing across exposure levels.

**Familiarity ratings.** We analyzed familiarity in Experiment 4 using similar mixed-modeling methods as attractiveness, according to an Exposure Level (4 [within]: 0, 1, 3, 7 exposures) × Target Type (2 [within]: individual, morph) fixed-effects structure.[15]

Figure 6b displays the results for familiarity. We observed the predicted Exposure Level × Target Type interaction, $F(3, 7896.70) = 11.97$, $p < .001$. Although there were no clear familiarity differences between individuals and morphs with no exposure (level 0), $b = 0.12$, $t(175.90) = 0.78$, $CI_{95\%}$ [−0.18 0.42], *n*s, or weak exposure (level 1), $b = 0.06$, $t(175.90) = 0.42$, $CI_{95\%}$ [−0.24 0.36], *n*s, individuals were judged as marginally more familiar with medium exposure (level 3), $b = 0.27$, $t(175.90) = 1.76$, $CI_{95\%}$ [−0.03 0.57], $p = .08$, and significantly more familiar

[15] With familiarity ratings in Experiment 4, the maximal random slope model would not converge, so the maximal random intercept model was set for fixed-effects testing (AIC = 35547.64, BIC = 35639.11).

with high exposure (level 7), $b = 0.59$, $t(175.90) = 3.88$, $CI_{95\%}$ [0.29 0.89], $p < .001$. Note, however, that at high exposure in Experiment 4 (level 7), the maximum level of familiarity (*M*s between 6 and 7 on 9-point scale) was lower than in Experiments 2A, 2B, and 3 (*M*s between 7.5–8.5 on 9-point scale). This is important because it demonstrates that the presence of the UiA effect (and its transition from the BiA effect) was less dramatic in Experiment 4 compared to previous studies. This is likely attributable to the modified passive exposure task in Experiment 4, which led to lower overall ratings of subjective familiarity across exposure levels (we will return to this issue in the General Discussion). We also observed a main effect of Exposure Level, $F(3, 343.60) = 43.09$, $p < .001$, which showed, unsurprisingly, that both morphs and individuals were rated as more familiar at higher exposure levels. The main effect of Target Type was not significant, $F(1, 99.00) = 2.32$, *n*s.

**Comparative analysis between attractiveness and familiarity.** One important consideration is the extent to which the attractiveness and familiarity ratings were similar for individuals and morphs, across different levels of exposure. To examine this, we *z*-scored participants' attractiveness and familiarity ratings, then combined them into one dataset (to put all ratings on the same scale). Next, we created a new mixed-effects model that predicted the *z*-scored ratings, according to an Exposure Level (4 [within]: 0, 1, 3, 7 exposures) × Target Type (2 [within]: individual, morph) × Rating Type (2 [within]: attractiveness, familiarity) fixed-effects structure.[16]

Crucially, we did *not* observe a three-way interaction between Exposure Level, Target Type, and Rating Type, $F(3, 297.02) = 0.34$, *n*s. This suggests that the rating curves across exposure levels for individuals and morphs did not significantly differ by the type of rating (attractiveness vs. familiarity). We also observed some less theoretically important effects, which we do not discuss here.[17]

**Multilevel mediation.** We once again built multilevel mediation models using a similar procedure as the previous studies. However, given that we had four different exposure levels in Experiment 4, we needed to create three separate mediation models with different treatment values (i.e., exposure levels 1, 3, and 7) compared with the same control value (i.e., exposure level 0) for each target type (i.e., individuals vs. morphs). Recall that treatment variables with more than two levels need to be handled differently than binary treatment variables in multilevel mediation (Imai, Keele, & Tingley, 2010). Therefore, in Experiment 4, we generated six separate multilevel mediation models—where two models compared exposure levels 0 versus 1 (i.e., $M_{1(individual)}$ and $M_{1(morph)}$), two models compared exposure levels 0 versus 3 (i.e., $M_{3(individual)}$ and $M_{3(morph)}$), and two models compared exposure levels 0 versus 7 (i.e., $M_{7(individual)}$ and $M_{7(morph)}$). Across all models, our main predictor was amount of exposure, our main DV was attractiveness ratings, and our mediator was familiarity ratings.

Table 1 displays complete results for all multilevel mediation models in Experiment 4.[18] On individuals, familiarity mediated the relationship between exposure and attractiveness in all models (i.e., $M_{1(individual)}$, $M_{3(individual)}$, and $M_{7(individual)}$). Critically, this average causal mediation effect (ACME) from familiarity became steadily stronger as the amount of exposure increased from $M_{1(individual)}$ to $M_{3(individual)}$ to $M_{7(individual)}$. This pattern suggests that the mediating effect of familiarity between exposure and attractiveness for individuals becomes especially robust in strong learning contexts (i.e., level 7). Interestingly, for morphs, the reverse effect seemed to emerge. As exposure increased from $M_{1(morph)}$ to $M_{7(morph)}$, the mediating effects of familiarity generally dissipated with increasing exposure (i.e., the ACME was marginal in $M_{1(morph)}$ and significant in $M_{3(morph)}$ but not significant in $M_{7(morph)}$). This demonstrates that familiarity drives the relationship between exposure and attractiveness for morphs, but more so in conditions of weaker learning (i.e., levels 1 and 3).

Note that this aligns with predictions from our memory-based framework. Under conditions of weak learning, increased familiarity should drive attractiveness for morphs (and thus, a traditional BiA effect). Under conditions of strong learning, there is a degradation of familiarity cues for morphs relative to individuals (and thus, a UiA effect occurs).

**Old/new judgments.** In Experiment 4, after all attractiveness and familiarity ratings, we also had participants do speeded trials of "old or new" judgments on all individual and morph stimuli, based on the memory for the faces that they were exposed to during the training task. This allowed us to obtain two measures

[16] For the comparative analysis between attractiveness and familiarity in Experiment 4, there was no significant difference between the maximal random slope and intercept models, $\chi^2(14) = 0$, *n*s, so the random intercept model was set for fixed-effects testing (AIC = 2620.42; BIC = 2749.49).

[17] On the comparative analysis between attractiveness and familiarity in Experiment 4, we did observe a main effect of Exposure Level, $F(3, 297.20) = 26.81$, $p < .001$, which only showed that the *z*-scored ratings varied with different levels of exposure. An Exposure Level × Target Type interaction, $F(3, 297.03) = 14.64$, $p < .001$, indicated that across both attractiveness and familiarity, individuals received steadily increasing ratings from low exposure (level 0) to high exposure (level 7). For morphs, there was a similar rating increase from no exposure (level 0) to medium exposure (level 3), but this decreased at high exposure (level 7). A marginal Exposure Level × Rating Type interaction, $F(3, 297.16) = 2.34$, $p = .07$, demonstrated that for both individuals and morphs, attractiveness gradually increased from no exposure (level 0) to medium exposure (level 3) but then dropped off during high exposure (level 7). For familiarity, there was a more linear increase in ratings from no exposure (level 0) to high exposure (level 7). Finally, a Target Type × Rating Type interaction, $F(1, 99.00) = 6.00$, $p = .016$, showed that morphs were rated as more attractive than individuals, but individuals were rated as more familiar than morphs.

[18] In Experiment 4, all *a*-path models for individuals and morphs were significant ($M_{1(individual)}$: $b = 0.50$, $t(99.00) = 6.68$, $p < .001$; $M_{3(individual)}$: $b = 0.88$, $t(99.00) = 9.76$, $p < .001$; $M_{7(individual)}$: $b = 1.18$, $t(99.00) = 9.83$, $p < .001$; $M_{1(morph)}$: $b = 0.32$, $t(99.00) = 3.13$, $p = .002$; $M_{3(morph)}$: $b = 0.49$, $t(99.00) = 4.93$, $p < .001$; $M_{7(morph)}$: $b = 0.47$, $t(99.00) = 4.59$, $p < .001$), and all *b*-path models for individuals were significant ($M_{1(individual)}$: $b = 0.16$, $t(154.06) = 4.04$, $p < .001$; $M_{3(individual)}$: $b = 0.18$, $t(170.52) = 4.89$, $p < .001$; $M_{7(individual)}$: $b = 0.17$, $t(197.28) = 5.06$, $p < .001$). For morphs, only the *b*-path for $M_{3(morph)}$ was significant ($M_{1(morph)}$: $b = 0.07$, $t(132.43) = 1.88$, $p = .06$; $M_{3(morph)}$: $b = 0.12$, $t(141.93) = 2.84$, $p = .005$; $M_{7(morph)}$: $b = 0.06$, $t(151.16) = 1.56$, *n*s). All *c*-path models were also significant except for $M_{7(morph)}$ ($M_{1(individual)}$: $b = 0.28$, $t(99.00) = 3.86$, $p < .001$; $M_{3(individual)}$: $b = 0.45$, $t(99.00) = 5.56$, $p < .001$; $M_{7(individual)}$: $b = 0.43$, $t(99.00) = 5.48$, $p < .001$; $M_{1(morph)}$: $b = 0.28$, $t(99.00) = 2.67$, $p = .009$; $M_{3(morph)}$: $b = 0.30$, $t(99.00) = 3.00$, $p = .003$; $M_{7(morph)}$: $b = 0.11$, $t(99.00) = 1.18$, *n*s). On the *c'*-path models, familiarity still significantly predicted attractiveness while reducing the significance of exposure, except for $M_{1(morph)}$ and $M_{7(morph)}$ ($M_{1(individual)}$: $b = 0.21$, $t(112.08) = 2.89$, $p = .005$; $M_{3(individual)}$: $b = 0.34$, $t(127.55) = 3.90$, $p < .001$; $M_{7(individual)}$: $b = 0.30$, $t(135.82) = 3.36$, $p = .001$; $M_{1(morph)}$: $b = 0.26$, $t(101.37) = 2.47$, $p = .015$; $M_{3(morph)}$: $b = 0.26$, $t(105.20) = 2.52$, $p = .01$; $M_{7(morph)}$: $b = 0.09$, $t(105.34) = 0.88$, *n*s).

Table 1
*Multilevel Mediation Results for Experiment 4*

| Target type | Exposure level (IV) | Model index | ACME [$CI_{95\%}$] | ADE [$CI_{95\%}$] | TE [$CI_{95\%}$] |
|---|---|---|---|---|---|
| Individuals | Level 0 vs. Level 1 | $M_{1(individual)}$ | .06 [.02 .11]** | .21 [.06 .37]** | .27 [.12 .42]** |
| | Level 0 vs. Level 3 | $M_{3(individual)}$ | .11 [.04 .18]** | .34 [.17 .50]** | .45 [.29 .60]** |
| | Level 0 vs. Level 7 | $M_{7(individual)}$ | .13 [.04 .23]** | .30 [.13 .46]** | .43 [.28 .59]** |
| Morphs | Level 0 vs. Level 1 | $M_{1(morph)}$ | .02 [−.004 .05]# | .27 [.06 .47]** | .28 [.07 .49]** |
| | Level 0 vs. Level 3 | $M_{3(morph)}$ | .05 [.008 .10]* | .26 [.06 .45]** | .31 [.12 .49]** |
| | Level 0 vs. Level 7 | $M_{7(morph)}$ | .03 [−.01 .08] | .09 [−.09 .29] | .12 [−.06 .32] |

*Note.* In all models, familiarity was our mediator, and attractiveness was our DV. ACME = average causal mediation effect; ADE = average direct effect; IV = independent variable (treatment); TE = total effect.
# $p \leq .10$. * $p \leq .05$. ** $p \leq .01$.



indicative of memory strength—proportion "old" responses and RTs. Importantly, recall that old/new judgments are widely considered in memory literature to be more context-bound than generic "familiarity" judgments, as answering the old/new recognition question requires determining whether the item was on the particular list the experimenter is asking about. Accordingly, global familiarity and recognition judgments can show somewhat different patterns (Whittlesea & Price, 2001; Wixted & Mickes, 2014).

For both DVs, we used similar mixed-effects modeling methods as attractiveness and familiarity, according to an Exposure Level (4 [within]: 0, 1, 3, 7 exposures) × Target Type (2 [within]: individual, morph) fixed-effects structure. RTs were $\log_{10}$-transformed before analysis to reduce the impact of outliers (both correct and incorrect RTs were included here).[19]

**Proportion "old" responses.** Figure 6c displays the results for proportion "old" responses. We observed the predicted Exposure Level × Target Type interaction, $F(3, 329.46) = 16.93$, $p < .001$. With no exposure (level 0), participants logged marginally more "old" responses (i.e., false alarms) to morphs than individuals, $b = 0.05$, $t(303.70) = 1.81$, $CI_{95\%}$ [−0.004 0.10], $p = .07$. This effect was in the same direction with weak exposure (level 1) but did not reach significance, $b = 0.02$, $t(303.70) = 0.62$, $CI_{95\%}$ [−0.03 0.07], *n*s. With greater exposure, the proportion of "old" responses between individuals and morphs started to diverge, where participants logged more "old" judgments for individuals at medium exposure (level 3), $b = 0.06$, $t(303.70) = 2.29$, $CI_{95\%}$ [0.008 0.11], $p = .02$, with this significant difference increasing further at high exposure (level 7), $b = 0.14$, $t(303.70) = 5.69$, $CI_{95\%}$ [0.09 0.19], $p < .001$. Aside from the interaction, we also detected a main effect of Exposure Level, $F(3, 329.48) = 59.70$, $p < .001$, which just revealed that there were greater proportions of "old" responses as exposure increased. The main effect of Target Type was marginal, $F(1, 99.51) = 3.83$, $p = .053$, which just showed that individuals garnered greater "old" proportions than morphs when collapsing across exposure levels.

It is also worth noting that participants' proportion "old" responses closely tracked their familiarity ratings, even though these old/new judgments were given under time pressure in a later, separate phase of the experiment. More specifically, although there were no clear differences between individuals and morphs with low exposure (levels 0–1) for either familiarity or proportion "old," these differences grew larger when moving to medium exposure (level 3) and high exposure (level 7).

**RTs.** Figure 6d displays the RT results. We observed a marginal Exposure Level × Target Type interaction, $F(3, 358.01) = 2.50$, $p = .059$, along with significant main effects for both Exposure Level, $F(3, 359.28) = 5.47$, $p = .001$, and Target Type, $F(1, 97.85) = 10.03$, $p = .002$. Participants showed overall faster RTs when responding to individuals, and not surprisingly, their RTs were faster at higher exposure levels. A breakdown of the interaction revealed that participants made faster "old/new" responses to individuals than morphs with no exposure (level 0), $b = -0.02$, $t(431.40) = -1.99$, $CI_{95\%}$ [−0.03 −0.0002], $p = .047$, and high exposure (level 7), $b = -0.03$, $t(433.50) = -3.74$, $CI_{95\%}$ [−0.04 −0.01], $p < .001$. Participants were still faster to respond to individual faces with weak exposure (level 1), $b = -0.001$, $t(432.80) = -0.14$, $CI_{95\%}$ [−0.02 0.01], *n*s, and medium exposure (level 3), $b = -0.01$, $t(431.50) = -1.23$, $CI_{95\%}$ [−0.02 0.006], *n*s, but these differences did not reach significance.

## Computational Memory Modeling

Our experiments suggest the critical role of memory processes underlying subjective familiarity in preferences for individual faces and their blends. One additional way to appreciate the role of such memory processes is to use simple computational memory modeling and examine whether our assumptions can produce the observed empirical patterns—especially the crossover interaction from Experiment 4. Here, we offer a very simple REM model (Shiffrin & Steyvers, 1997) that implements such core assumptions and provides a concrete "existence proof" that the global match memory models actually produce the patterns we observed (again, without trying to fit all aspects of the data).

Before we go into some details of our particular model, let us note a few general issues concerning modeling the BiA and UiA effects using memory models with differentiation (Criss, 2006; Criss, Wheeler, & McClelland, 2013). Although a variety of memory models with differentiation naturally predict a greater UiA effect with increasing prior training, they do not necessarily predict BiA with weak prior training, particularly for a two-face blend. This is because memory models that include differentiation necessarily stipulate that the retrieval strength between a blend and a

[19] For proportion "old" in Experiment 4, the maximal random slope model would not converge, so the maximal random intercept model was set for fixed-effects testing (AIC = 10654.60, BIC = 10746.07). The same was true for old/new RTs (AIC = −6752.61, BIC = −6661.69).

memory trace that halfway matches the blend will be less than half the strength of a perfectly matching memory trace. Thus, in the global match familiarity signal, the two half-matches for the blend add up to a familiarity value that is less than the familiarity value for one whole match.

However, this idea ignores two highly plausible assumptions, which in turn allows these models to produce a BiA effect even for a two-face blend (despite a reduction in familiarity owing to differentiation). First, it is likely that the blend can appear similar to a large number of faces in memory (beyond only the constituent faces that compose it). If so, the collection of partial matches can readily add up to more than one whole match, despite the inclusion of differentiation. Second, it is also safe to assume that attention fluctuates when participants are studying the faces. As a result, some faces are well-encoded even with just one exposure, whereas other faces are missed entirely. By including the well-supported assumption of trial-by-trial encoding variability (e.g., Young & Bellezza, 1982), a BiA effect is produced for a two-face blend.

With trial-by-trial encoding variability, the key question is whether each individual face was or was not encoded during training. To make this more concrete with an extreme example, suppose that there was only a 10% chance that each face was encoded during training, in a situation of a two-face blend between faces that received just one training exposure. If we label the blend A-B, we can consider the separate possible outcomes of training: (a) face A was encoded (10%) but face B was not (90%), with this combination of encoding occurring with a 9% chance (the product of 10% and 90%); (b) face A was not encoded (90%) but face B was (10%), with this combination of encoding occurring with a 9% chance; (c) and finally, both face A (10%) and B (10%) were encoded, with this combination of encoding occurring with a 1% chance. Thus, across these outcomes, there is a 19% chance (9% + 9% + 1%) that at least one of the two individual faces were encoded. This 19% is the chance that the blend will elicit an above-baseline level of familiarity. Next, when you consider the familiarity for a test with one of the individual faces (either face A or face B), the chance of an above-baseline level of familiarity is only 10% (i.e., the chance that face was or was not encoded). In turn, due to trial-by-trial encoding variability, this extreme example produces nearly twice the chance that the A-B blend will produce above-baseline familiarity, compared to an individual face. Keep in mind that with increasing numbers of prior exposures, it becomes certain that both the A and B faces will have been encoded, and once this occurs, the A-B blend and both the A and B individual faces will assuredly elicit above-baseline familiarity. Differentiation thus takes over, and the two half-matches for the A-B blend will add up to a familiarity total that is less than that which occurs for a test of the A or B individual faces.

**REM model.** Many memory models include differentiation, but for our specific implementation, we chose the Bayesian Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997) model as representative of this class of memory models. The REM model is arguably the most successful of this class of memory models, and its differentiation assumption is well-supported (Criss, 2006; Criss, Wheeler, & McClelland, 2013). As mentioned, rather than fitting this model to our data, we present an existence proof that this model produces the observed crossover interaction when comparing familiarity for parents versus morphs as a function of training for the parent faces. Our simulation with the REM model was straightforward, using the "off-the-shelf" original version of the model, except for one simple change. The original model assumed an independent storage probability for each feature of a studied item (e.g., the first time you study an item, you might store 50% of the features, and then the second time you might store 50% of the remaining features [yielding 75% in total], etc.). However, this simplifying assumption ignores the earlier discussed principle of encoding variability, whereby the participant is sometimes in a state of high arousal during encoding, creating strong memories, while other times they completely fail to encode anything into memory (Young & Bellezza, 1982). A more realistic model would involve a mixture of feature-by-feature and trial-by-trial encoding variability. Based on our account of the BiA effect, we suspected that trial-by-trial variability would be the more important factor. Thus, our simulation only used trial-by-trial encoding in which the encoding probability parameter dictated the chance of encoding all the features versus none of the features with each study trial.[20]

As shown in Figure 7, our version of the REM model produces the crossover interaction, with a BiA effect for low training and a UiA effect for high training.

**Additional issues and alternative frameworks.** As we have emphasized throughout, our REM model is intended to provide a simple existence proof that computationally simulating memory mechanisms can generate the empirically observed transitions of familiarity responses from exemplars to blends. We also want to briefly address some questions about this choice and related frameworks. One issue is why we chose a simple memory model (REM) and not a face-space model (see O'Toole, 2011 for a review). This is because we focused on memory processes, collected familiarity ratings, and manipulated face exposure levels, whereas the face-space literature cares more about representing the similarity relation between many faces and their features (an important issue, but not for our purposes here). Another issue is why we did not model additional influences on familiarity and preference for morphs—most critically, the fact that morphs are more likely to be in the center of the face space (especially with large numbers of faces). Indeed, empirically, with a very large number of faces (like in Experiment 4), the morphs are empirically more "familiar" even with 0 exposures to individual exemplars. However, modeling this influence is not central to our main point about the degree of learning, and this would make our REM model more complicated. Finally, one could also argue that it might be more optimal to capture the underlying changes in memory representation by modeling changes in probability distributions associated with each face or its features (Dailey, Cottrell, & Busey, 1999). Repeated exposure to a face essentially makes the variance of the probability for that specific presented face narrower and taller (sharper). Conse-

[20] Without any trial-by-trial variability, the REM model typically produces a UiA effect (i.e., less familiarity for the morph than its parents). This is true even for low levels of training, although the magnitude of the UiA effect increases with training. The exception to this is when feature-by-feature encoding probability is set very low, which then produces a similar curve, even without trial-by-trial variability.

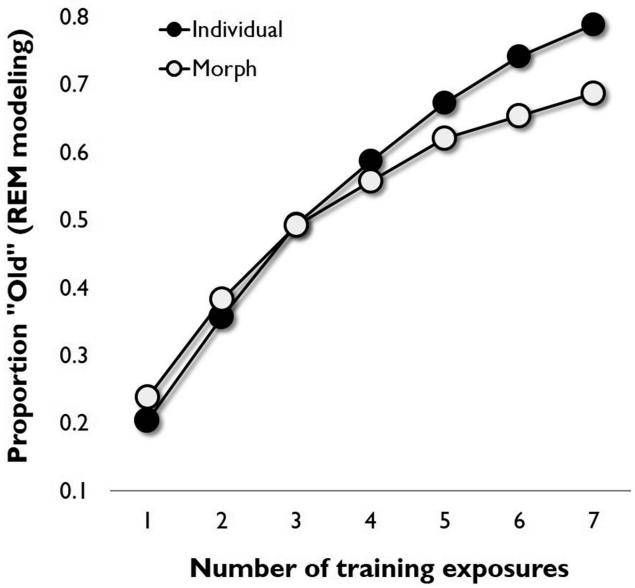


*Figure 7.* Simulations with the Retrieving Effectively from Memory (REM) model of Shiffrin and Steyvers (1997). The simulations assumed study of 28 individual parent faces followed by testing of these same individual faces or testing of morph faces. With 20 features per face, morph faces were created by having 10 features match the features of one parent face and the remaining 10 features match the features of the other parent face. The encoding probability parameter ($u$) was set to 0.2, the geometric distribution parameter ($g$) was set to 0.2, and the correct storage parameter ($c$) was set to 0.7. These are typical values for the REM model, but a wide range of parameter values produce the same results (all parameter values produced a UiA effect with high training and a subset of parameter values also produced a BiA effect with low training). The criterion for "old" responses was set to the default value of log odds equal to 0. The only substantive change made to the model was all-or-none encoding for all the features of a studied face, rather than feature-by-feature encoding.

quently, weakly learned faces have wider and shallower probability distributions, making their blend more probable than each individual face (BiA effect). In contrast, strongly learned faces have narrower and taller probability distributions, making the blend relatively less probable than the specific face (UiA effect).[21] For our purposes in the current studies, however, we decided to focus on a simpler memory model, like REM, which easily produces our observed effects.

## General Discussion

The current research addressed the mechanisms underlying classic social preference effects and tested predictions generated by modern models of memory that include the mechanism of differentiation. With five experiments and computational memory modeling, we found that different amounts of exposure predictably change the absolute and relative preferences for individuals and morphs. Our experiments replicate classic phenomena of mere exposure (all experiments) and the beauty-in-averageness (BiA) effect (Experiments 1 and 4). Critically, they also extensively document an *ugliness*-in-averageness (UiA) effect, where morphs of familiar individuals are judged as *less* attractive than contributing individuals (Experiments 2A, 2B, 3, and 4). The experiments also suggest that the UiA effect is attributable to a relative reduction in familiarity for morphs of trained individuals, where the attractiveness of highly familiar exemplars "trumps" the less familiar morphs. Moreover, consistent with predictions derived from memory theories, the UiA effect does not require a conflict between two well-known individuals, but only requires a decrease of familiarity of a single well-known exemplar (Experiment 3). This suggests a relatively basic, low-level process, especially given that the UiA effects can be generated by both identity-specific familiarity (e.g., names; Experiment 2A) and basic visual familiarity (e.g., perceptual tracking; Experiment 2B). Note that different encoding goals across experiments cannot explain our findings, given that parametrically scaling the number of exposures produces a BiA effect with weak learning and a UiA effect with strong learning (within the same task, using a within-subject design). These attractiveness effects also paralleled subjective familiarity ratings and objective measures of memory strength (Experiment 4). Finally, we replicated the crossover interaction for attractiveness ratings in Experiment 4 using simulations from the well-established Retrieving Effectively from Memory (REM) model (Shiffrin & Steyvers, 1997). Taken together, these studies offer the first systematic and mechanistic demonstration of the UiA effect, which combines two classic determinants of preferences in social psychology—*mere exposure* (i.e., stimulus repetition) and *blending* (i.e., stimulus averaging). Our findings not only highlight the importance of memory processes in understanding social judgments like attractiveness, but the results also represent a major qualification to the classic BiA effect, known since Galton (1879) and confirmed by a multitude of studies using a variety of different paradigms, stimuli, and modalities (e.g., Halberstadt & Rhodes, 2003; Langlois & Roggman, 1990; Rhodes & Tremewan, 1996). As such, our results should extend beyond social judgments of faces, since the interaction between prototypicality (blending) and exposure is evident in a variety of other domains (e.g., understanding market dynamics; Landwehr, Wentzel, & Herrmann, 2010).

We will now review in detail each of the major findings, while highlighting their broader theoretical implications—but first, let us restate some major assumptions of modern memory theories. Recall that on those theories, memories contain traces for individual exemplars (e.g., specific faces that are studied). The familiarity of

[21] The perceptual categorization literature often represents similar processes using so-called "Bayesian mixture models." Basically, such models are learning stimulus features in some multidimensional space, with values represented by the mean and precision (width) of two Gaussian components. In such a model, one can consider the plausibility of the average stimulus, given what the model learned about the two individuals or categories. With weak learning, there is a lot of uncertainty about what the two individuals/categories are like. As a result, much probability gets assigned to middle values, thus making the blend of two individuals plausible. With greater certainty (strong learning), all the probability gets assigned quite precisely to the actual trained features. In turn, the blend stops being plausible. Simply put, with weak learning, the model has learned the values imprecisely (wide distribution and high uncertainty). Consequently, the stimulus in the middle is relatively more probable as a member of the previous category. However, with sufficient learning, the probability density in the middle decreases. That is, the probability that the average (blended) stimulus was in the training set decreases because learning leads to greater precision (separation and narrowing) of two probability clusters (for an example, see Feldman, Griffiths, & Morgan, 2009).

a probe (target) is calculated from the similarity values of the probe with all traces in memory (a so-called "global match" familiarity signal). The similarity between the probe and the memory trace is a function of the overlap between them and the strength of the memory. If the memory trace is weak (because only a few features of the item were stored), the similarity between the probe and the memory trace will be lower than when the memory trace contains many stored features. Thus, familiarity (and preference) will be higher for strong items than for weak items (i.e., mere exposure effect). With weak learning of multiple items, blend probes will partially match several memory traces and the sum of these partial matches can add up to a greater familiarity signal than what occurs for a nonblended face that only matches itself in memory. This situation predicts the BiA effect, as shown with our computational REM model that included parameters for trial-by-trial encoding variability. Crucially though, with strong learning, it is easier to note the differences between the known individual faces and the blend (also called "differentiation"), so the global familiarity signal elicited by the blend will be reduced, predicting the UiA effect. When participants rate morphs made from exemplars without any previous training at all, the memory literature predicts no BiA or UiA effects, assuming the "novel" faces do not activate familiarity signals for exposed faces (but see the next paragraph). Finally, note that our framework is not a simple extension of previous experiments on celebrity blends (e.g., Halberstadt et al., 2013). Aside from obvious challenges of using real local celebrities as stimuli, these previous studies (a) did not systematically manipulate exposure, (b) did not assess whether blends of well-known individuals are actually disliked or simply less liked than novel individual, (c) did not provide any evidence for boundary conditions, and (d) did not explore underlying mechanisms or ground the findings in broader cognitive principles (as we have done here with our memory-based framework).

Moving on to the main results, in Experiment 1, we found that weak training on exemplars generates the standard BiA effect—where morphs are judged as more attractive and familiar than individuals. This finding matches our memory account and fits with previous cognitive explanations of the BiA effect, which posit that blending two faces makes it better match to the "gist" or prototype (Principe & Langlois, 2012). Critically, the relationship between target type (individual vs. morph) and attractiveness was mediated by familiarity (such that morphs appear more familiar, and thereby more attractive). This is consistent with findings that attractiveness of average faces is associated with their implicit familiarity (Peskin & Newell, 2004; Rhodes, Halberstadt, & Brajkovich, 2001). Experiments 2A and 2B investigated the attractiveness for morphs of highly learned exemplars (i.e., when the individual exemplars have strong traces in memory) and morphs made from completely unfamiliar exemplars. In these experiments, no BiA effect emerged for morphs made from completely unfamiliar individuals, whereas the UiA effect emerged for trained morphs in both experiments. Interestingly, however, we did observe a BiA effect in Experiment 4 when using a passive exposure paradigm that parametrically varied the number of exposures within-subject. Here, not only was there a BiA effect when individuals were weakly learned (level 1), but it also occurred when there was no exposure (level 0). This is likely attributable to subjects having a noisier representation of the overall face space obtained during training. Recall that the memory literature would seem to predict no BiA or UiA effects on blends of novel faces, but this assumes that the novel faces do not share any similarity with actually exposed faces. In Experiment 4, the exposures during training encompassed a greater variety of faces (i.e., 196 exposures of 28 different individuals in Experiments 2A and 2B vs. 154 exposures of 42 different individuals in Experiment 4). Consequently, the setup in Experiment 4 would also lead to a greater likelihood that a "novel" morph (i.e., blend of two unknown individuals) would share seemingly similar features with a face at one of the other three exposure levels, thus generating more familiarity (i.e., "false alarms") and attractiveness (i.e., BiA effect) over its constituent individuals. Indeed, this is what we observed in Experiment 4. It is also worth noting that the UiA effect in Experiment 4 was weaker than in the other experiments, presumably because of the change to more passive exposures during training. In short, the 7 passive exposures in Experiment 4 were likely not as strongly encoded as the 7 more "active" exposures during the other studies (Experiments 2A, 2B, and 3), which also explains the relatively lower subjective familiarity ratings in Experiment 4 (see Figure 6b).

The results from Experiment 4 and the memory modeling clearly show that the BiA effect transitions into a UiA effect with greater exposure, which is driven by increased familiarity and memory strength for the learned individuals. Theoretically, this follows from our memory-based predictions, because individual target faces are more similar to strong memory traces than blended faces. Another feature of our data that offers additional support to the familiarity (memory-based) account is that blends of well-learned individuals generated familiarity and preference values in-between actually exposed individuals and novel individuals. This makes sense from a memory-based viewpoint, given that familiarity and liking is reduced with increased dissimilarity of the probe, but there are still positive effects from partial familiarity (Gordon & Holyoak, 1983). These robust confirmations of our memory-based account of familiarity can be contrasted with alternative theoretical predictions (*additive* and *mismatch* accounts), as previously described in the Introduction. Of particular note, in Experiment 3, our data directly supported the *familiarity* (memory-based) account over the *mismatch* (conflict-based) account, because a UiA effect still emerged when using cross-set morphs composed of one trained and one untrained individual (as opposed to the within-set morphs in Experiments 2A and 2B). Keep in mind, however, that our results do not challenge the overall validity of mismatch accounts (conflict-based or prediction-error-based) in the generation of negative affect (Dreisbach & Fischer, 2015; Shackman et al., 2011).

The current research also observed very strong support for a familiarity-positivity link. This connection has long been assumed to be at the core of the mere exposure effect (Titchener, 1915), and it works in a bidirectional manner, with positivity breeding familiarity (Garcia-Marques et al., 2004; Monin, 2003; Phaf & Rotteveel, 2005). Note, however, that this "warm glow" of familiarity can also fluctuate based on contextual factors, like mood, motivation, or goals (de Vries, Holland, Chenier, Starr, & Winkielman, 2010; Freitas, Azizian, Travers, & Berry, 2005; Hertwig, Herzog, Schooler, & Reimer, 2008). It also may depend on the specific judgment in-question, with attractiveness, liking, and desirability ratings sometimes showing different sensitivity to manipulations of mere exposure and prototypicality (DeBruine, 2005; Rhodes,

Halberstadt, & Brajkovich, 2001; Rhodes, Halberstadt, Jeffery, & Palermo, 2005). Thus, an interesting avenue for future research would be to investigate the role of affective, motivational, and judgmental contexts in the UiA effect.

Mechanistically, the familiarity-preference link could arise from underlying changes in perceptual fluency (Winkielman et al., 2003). However, there are also alternative models in which familiarity arises via alternative processes, linked to context-free recognition (e.g., Wagner & Gabrieli, 1998). While the fine-grained distinctions between "pure" fluency and "pure" familiarity are not essential for our main points, future research should disentangle these constructs. For instance, future studies could manipulate both fluency and familiarity to gauge the consequences on responses to individual and blended faces. This would be especially interesting, given that much previous research has shown a tight connection between familiarity- and fluency-based judgments (e.g., Whittlesea, 1993; Whittlesea, Jacoby, & Girard, 1990; Whittlesea & Williams, 2000, 2001a, 2001b). Moreover, the distinctions left open by the current studies could be addressed by neural measures (e.g., event-related potentials or fMRI) that have been shown to separate fluency from familiarity, via differences in activation timing (Wolk et al., 2004) and spatial localization (Nessler, Mecklinger, & Penney, 2005; Voss et al., 2008).

Going forward, the current work prompts many other intriguing questions. As one example, our experiments do not fully address how changes in typicality drive attractiveness ratings (rather than only familiarity). Previous research has shown that both typicality and familiarity are highly correlated with attractiveness, and the strength of these relationships depends on the specific stimulus category (Bartlett, Hurry, & Thorley, 1984; Halberstadt & Rhodes, 2003). It would be interesting for future studies to simultaneously manipulate both typicality and familiarity, to gauge the underlying links to attractiveness for both individual and morphed faces. Moreover, the current studies focused on neutral faces, but did not investigate the role of emotional expressions (e.g., smiling and frowning faces). Not only can valence modify our effects, but with such expressions, social familiarity may become more important. This is likely, given that fMRI studies have found activation of unique brain regions to person-based familiarity (Cloutier, Kelley, & Heatherton, 2011), and more generally, between social and nonsocial stimuli (Gobbini & Haxby, 2007; Haxby, Hoffman, & Gobbini, 2000; Johnson, 2005). Clearly, dimensions with social complexity also need to be considered (e.g., race or gender), as the effects of blending on these dimensions go substantially beyond simple memory processes (Bernstein, Young, & Hugenberg, 2007; Malpass & Kravitz, 1969; Hugenberg & Bodenhausen, 2004). Finally, it would also be interesting to gauge whether our UiA effect extends to modalities beyond vision (e.g., audition, via blended tones or melodies; Bruckert et al., 2010) or even cross-modal blends (Winkielman, Ziembowicz, & Nowak, 2015).

In sum, our studies represent the first systematic investigation of the UiA effect. We demonstrated how mere exposure and blending combine to impact familiarity—and how memory-based processes modify and reverse classic patterns of facial attractiveness. Simply put, the current experiments reveal that when it comes to highly familiar individuals, blends are not always most beautiful.



## References

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences, 16,* 390–398. http://dx.doi.org/10.1016/j.tics.2012.05.003

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology, 37,* 13–20. http://dx.doi.org/10.1111/1469-8986.3710013

Baker, W. E. (1999). When can affective conditioning and mere exposure directly influence brand choice? *Journal of Advertising, 28,* 31–46. http://dx.doi.org/10.1080/00913367.1999.10673594

Balogh, R., & Porter, R. H. (1986). Olfactory preferences resulting from mere exposure in human neonates. *Infant Behavior & Development, 9,* 395–401. http://dx.doi.org/10.1016/0163-6383(86)90013-5

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version, 1*(7).

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11,* 142–163. http://dx.doi.org/10.1037/1082-989X.11.2.142

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The cross-category effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological Science, 18,* 706–712. http://dx.doi.org/10.1111/j.1467-9280.2007.01964.x

Berntson, G. G., & Cacioppo, J. T. (2009). Evaluative processing. In D. Sander & K. Scherer (Eds.), *Oxford companion to emotion and the affective sciences*. New York, NY: Oxford University Press.

Bobes, M. A., Lage Castellanos, A., Quiñones, I., García, L., & Valdes-Sosa, M. (2013). Timing and tuning for familiarity of cortical responses to faces. *PLoS ONE, 8*(10), e76100. http://dx.doi.org/10.1371/journal.pone.0076100

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin, 106,* 265–289.

Bornstein, R. F., & D'Agostino, P. R. (1992). Stimulus recognition and the mere exposure effect. *Journal of Personality and Social Psychology, 63,* 545–552. http://dx.doi.org/10.1037/0022-3514.63.4.545

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., . . . Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology, 20,* 116–120. http://dx.doi.org/10.1016/j.cub.2009.11.034

Butler, L. T., & Berry, D. C. (2004). Understanding the relationship between repetition priming and mere exposure. *British Journal of Psychology, 95,* 467–487. http://dx.doi.org/10.1348/0007126042369776

Buttle, H., & Raymond, J. E. (2003). High familiarity enhances visual change detection for face stimuli. *Perception & Psychophysics, 65,* 1296–1306. http://dx.doi.org/10.3758/BF03194853

Cloutier, J., Kelley, W. M., & Heatherton, T. F. (2011). The influence of perceptual and knowledge-based familiarity on the neural substrates of face perception. *Social Neuroscience, 6,* 63–75. http://dx.doi.org/10.1080/17470911003693622

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55,* 461–478. http://dx.doi.org/10.1016/j.jml.2006.08.003

Criss, A. H., Wheeler, M. E., & McClelland, J. L. (2013). A differentiation account of recognition memory: Evidence from fMRI. *Journal of Cognitive Neuroscience, 25,* 421–435. http://dx.doi.org/10.1162/jocn_a_00292

Dailey, M. N., Cottrell, G. W., & Busey, T. A. (1999). Facial memory is

kernel density estimation (almost). In M. S. Kearns, S. A. Solla, & D. A. Cohen (Eds.), *Advances in neural information processing systems* (pp. 24–30). Cambridge, MA: MIT Press.

Davies-Thompson, J., Newling, K., & Andrews, T. J. (2013). Image-invariant responses in face-selective regions do not explain the perceptual advantage for familiar face recognition. *Cerebral Cortex, 23,* 370–377.

DeBruine, L. M. (2005). Trustworthy but not lust-worthy: Context-specific effects of facial resemblance. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 272,* 919–922. http://dx.doi.org/10.1098/rspb.2004.3003

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 62,* 1716–1722. http://dx.doi.org/10.1080/17470210902811249

de Vries, M., Holland, R. W., Chenier, T., Starr, M. J., & Winkielman, P. (2010). Happiness cools the warm glow of familiarity: Psychophysiological evidence that mood modulates the familiarity-affect link. *Psychological Science, 21,* 321–328. http://dx.doi.org/10.1177/0956797609359878

Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour, 1,* 1–6. http://dx.doi.org/10.1038/s41562-016-0001

Dreisbach, G., & Fischer, R. (2015). Conflicts as aversive signals for control adaptation. *Current Directions in Psychological Science, 24,* 255–260. http://dx.doi.org/10.1177/0963721415569569

Fang, X., Singh, S., & Ahluwalia, R. (2007). An examination of different explanations for the mere exposure effect. *Journal of Consumer Research, 34,* 97–103. http://dx.doi.org/10.1086/513050

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116,* 752–782. http://dx.doi.org/10.1037/a0017196

Freitas, A. L., Azizian, A., Travers, S., & Berry, S. A. (2005). The evaluative connotation of processing fluency: Inherently positive or moderated by motivational context? *Journal of Experimental Social Psychology, 41,* 636–644. http://dx.doi.org/10.1016/j.jesp.2004.10.006

Galton, F. (1879). Composite portraits made by combining those of many different persons into a single figure. *Nature, 18,* 97–100.

Garcia-Marques, T., Mackie, D. M., Claypool, H. M., & Garcia-Marques, L. (2004). Positivity can cue familiarity. *Personality and Social Psychology Bulletin, 30,* 585–593. http://dx.doi.org/10.1177/0146167203262856

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67. http://dx.doi.org/10.1037/0033-295X.91.1.1

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia, 45,* 32–41. http://dx.doi.org/10.1016/j.neuropsychologia.2006.04.015

Gordon, P. C., & Holyoak, K. J. (1983). Implicit learning and generalization of the "mere exposure" effect. *Journal of Personality and Social Psychology, 45,* 492–500. http://dx.doi.org/10.1037/0022-3514.45.3.492

Guo, C., Voss, J. L., & Paller, K. A. (2005). Electrophysiological correlates of forming memories for faces, names, and face-name associations. *Cognitive Brain Research, 22,* 153–164. http://dx.doi.org/10.1016/j.cogbrainres.2004.08.009

Halberstadt, J. (2006). The generality and ultimate origins of the attractiveness of prototypes. *Personality and Social Psychology Review, 10,* 166–183. http://dx.doi.org/10.1207/s15327957pspr1002_5

Halberstadt, J., Pecher, D., Zeelenberg, R., Ip Wai, L., & Winkielman, P. (2013). Two faces of attractiveness: Making beauty in averageness appear and reverse. *Psychological Science, 24,* 2343–2346. http://dx.doi.org/10.1177/0956797613491969

Halberstadt, J., & Rhodes, G. (2003). It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin & Review, 10,* 149–156. http://dx.doi.org/10.3758/BF03196479

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4,* 223–233. http://dx.doi.org/10.1016/S1364-6613(00)01482-0

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1191–1206. http://dx.doi.org/10.1037/a0013025

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428. http://dx.doi.org/10.1037/0033-295X.93.4.411

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science, 310,* 1680–1683. http://dx.doi.org/10.1126/science.1115327

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science, 15,* 342–345. http://dx.doi.org/10.1111/j.0956-7976.2004.00680.x

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15,* 309–334. http://dx.doi.org/10.1037/a0020761

Johnson, M. H. (2005). Subcortical face processing. *Nature Reviews Neuroscience, 6,* 766–774. http://dx.doi.org/10.1038/nrn1766

Jones, T. C., & Jacoby, L. L. (2001). Feature and conjunction errors in recognition memory: Evidence for Dual-Process Theory. *Journal of Memory and Language, 45,* 82–102. http://dx.doi.org/10.1006/jmla.2000.2761

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 701–722. http://dx.doi.org/10.1037/0278-7393.27.3.701

Klinger, M. R., & Greenwald, A. G. (1994). Preferences need no inferences? The cognitive basis for unconscious emotional effects. In P. M. Niedenthal & S. Kitayama (Eds.), *The heart's eye: Emotional influences in perception and attention* (pp. 67–85). Orlando, FL: Academic Press. http://dx.doi.org/10.1016/B978-0-12-410560-7.50010-7

Kornmeier, J., & Bach, M. (2012). Ambiguous figures—What happens in the brain when perception changes but not the stimulus. *Frontiers in Human Neuroscience, 6*.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests in linear mixed effects models. *R package version* 2.0–20.

Landwehr, J. R., Wentzel, D., & Herrmann, A. (2010). The influence of prototypicality and level of exposure on consumers' responses to product designs: Field evidence from German car buyers. *Advances in Consumer Research Association for Consumer Research, 37,* 682–683.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1,* 115–121. http://dx.doi.org/10.1111/j.1467-9280.1990.tb00079.x

Luke, S. G. (2016). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*. Advance online publication. http://dx.doi.org/10.3758/s13428-016-0809-y

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology, 13,* 330–334. http://dx.doi.org/10.1037/h0028434

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271. http://dx.doi.org/10.1037/0033-295X.87.3.252

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760. http://dx.doi.org/10.1037/0033-295X.105.4.734-760

McGugin, R. W., Tanaka, J. W., Lebrecht, S., Tarr, M. J., & Gauthier, I. (2011). Race-specific perceptual discrimination improvement following short individuation training with faces. *Cognitive Science, 35,* 330–347. http://dx.doi.org/10.1111/j.1551-6709.2010.01148.x

Monahan, J. L., Murphy, S. T., & Zajonc, R. B. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science, 11,* 462–466. http://dx.doi.org/10.1111/1467-9280.00289

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology, 85,* 1035–1048. http://dx.doi.org/10.1037/0022-3514.85.6.1035

Moreland, R. L., & Topolinski, S. (2010). The mere exposure phenomenon: A lingering melody by Robert Zajonc. *Emotion Review, 2,* 329–339. http://dx.doi.org/10.1177/1754073910375479

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Natu, V., & O'Toole, A. J. (2011). The neural processing of familiar and unfamiliar faces: A review and synopsis. *British Journal of Psychology, 102,* 726–747. http://dx.doi.org/10.1111/j.2044-8295.2011.02053.x

Nessler, D., Mecklinger, A., & Penney, T. B. (2005). Perceptual fluency, semantic familiarity and recognition-related familiarity: An electrophysiological exploration. *Cognitive Brain Research, 22,* 265–288. http://dx.doi.org/10.1016/j.cogbrainres.2004.03.023

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. http://dx.doi.org/10.1037/0096-3445.115.1.39

Obermiller, C. (1985). Varieties of mere exposure: The effects of processing style and repetition on affective response. *Journal of Consumer Research, 12,* 17–30. http://dx.doi.org/10.1086/209032

O'Toole, A. J. (2011). Cognitive and computational approaches to face recognition. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *The Oxford handbook of face perception* (pp. 15–30). Oxford, UK: Oxford University Press.

Peskin, M., & Newell, F. N. (2004). Familiarity breeds attraction: Effects of exposure on the attractiveness of typical and distinctive faces. *Perception, 33,* 147–157. http://dx.doi.org/10.1068/p5028

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology, 38,* 922–934. http://dx.doi.org/10.1002/ejsp.504

Phaf, R. H., & Rotteveel, M. (2005). Affective modulation of recognition bias. *Emotion, 5,* 309–318. http://dx.doi.org/10.1037/1528-3542.5.3.309

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363. http://dx.doi.org/10.1037/h0025953

Principe, C. P., & Langlois, J. H. (2012). Shifting the prototype: Experience with faces influences affective and attractiveness preferences. *Social Cognition, 30,* 109–120. http://dx.doi.org/10.1521/soco.2012.30.1.109

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178. http://dx.doi.org/10.1037/0278-7393.16.2.163

R Core Team. (2015). R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing, 2013. http://www.r-project.org

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8,* 364–382. http://dx.doi.org/10.1207/s15327957pspr0804_3

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition, 19,* 57–70. http://dx.doi.org/10.1521/soco.19.1.57.18961

Rhodes, G., Halberstadt, J., Jeffery, L., & Palermo, R. (2005). The attractiveness of average faces is not a generalized mere exposure effect. *Social Cognition, 23,* 205–217. http://dx.doi.org/10.1521/soco.2005.23.3.205

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological Science, 7,* 105–110. http://dx.doi.org/10.1111/j.1467-9280.1996.tb00338.x

Rhodes, G., & Zebrowitz, L. A. (Eds.). (2002). *Facial attractiveness: Evolutionary, cognitive, and social perspectives*. New York, NY: Ablex.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. http://dx.doi.org/10.1037/0278-7393.21.4.803

Rubenstein, A. J., Kalakanis, L., & Langlois, J. H. (1999). Infant preferences for attractive faces: A cognitive explanation. *Developmental Psychology, 35,* 848–855. http://dx.doi.org/10.1037/0012-1649.35.3.848

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25,* 638–656. http://dx.doi.org/10.1521/soco.2007.25.5.638

Schweinberger, S. R., Pickering, E. C., Burton, A. M., & Kaufmann, J. M. (2002). Human brain potential correlates of repetition priming in face and name recognition. *Neuropsychologia, 40,* 2057–2073. http://dx.doi.org/10.1016/S0028-3932(02)00050-7

Shackman, A. J., Salomons, T. V., Slagter, H. A., Fox, A. S., Winter, J. J., & Davidson, R. J. (2011). The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nature Reviews Neuroscience, 12,* 154–167. http://dx.doi.org/10.1038/nrn2994

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 267–287. http://dx.doi.org/10.1037/0278-7393.21.2.267

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179–195. http://dx.doi.org/10.1037/0278-7393.16.2.179

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. http://dx.doi.org/10.3758/BF03209391

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70,* 893–912. http://dx.doi.org/10.1037/0022-3514.70.5.893

Smith, P. K., Dijksterhuis, A., & Chaiken, S. (2008). Subliminal exposure to faces and racial attitudes: Exposure to Whites makes Whites like Blacks less. *Journal of Experimental Social Psychology, 44,* 50–64. http://dx.doi.org/10.1016/j.jesp.2007.01.006

Thompson, D. (2017). *Hit makers: The science of popularity in an age of distraction*. New York, NY: Penguin Press.

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry, and parasite resistance. *Human Nature, 4,* 237–269. http://dx.doi.org/10.1007/BF02692201

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis.

Titchener, E. B. (1915). *A beginner's psychology*. New York, NY: Macmillan. http://dx.doi.org/10.1037/11238-000

Topolinski, S., Erle, T. M., & Reber, R. (2015). Necker's smile: Immediate affective consequences of early perceptual processes. *Cognition, 140,* 1–13. http://dx.doi.org/10.1016/j.cognition.2015.03.004

Tremblay, K. L., Inoue, K., McClannahan, K., & Ross, B. (2010). Repeated stimulus exposure alters the way sound is encoded in the human brain. *PLoS ONE, 5,* e10283. http://dx.doi.org/10.1371/journal.pone.0010283

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science, 21,* 779–785. http://dx.doi.org/10.1177/0956797610371965

Verosky, S. C., Todorov, A., & Turk-Browne, N. B. (2013). Representations of individuals in ventral temporal cortex defined by faces and biographies. *Neuropsychologia, 51,* 2100–2108. http://dx.doi.org/10.1016/j.neuropsychologia.2013.07.006

Visconti di Oleggio Castello, M., & Gobbini, M. I. (2015). Familiar face detection in 180 ms. *PLoS ONE, 10,* e0136548. http://dx.doi.org/10.1371/journal.pone.0136548

Voss, J. L., Reber, P. J., Mesulam, M. M., Parrish, T. B., & Paller, K. A. (2008). Familiarity and conceptual priming engage distinct cortical networks. *Cerebral Cortex, 18,* 1712–1719. http://dx.doi.org/10.1093/cercor/bhm200

Wagner, A. D., & Gabrieli, J. D. E. (1998). On the relationship between recognition familiarity and perceptual fluency: Evidence for distinct mnemonic processes. *Acta Psychologica, 98,* 211–230. http://dx.doi.org/10.1016/S0001-6918(97)00043-7

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/b17198

Whittlesea, B. W. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1235–1253. http://dx.doi.org/10.1037/0278-7393.19.6.1235

Whittlesea, B. W. (2002). False memory and the discrepancy-attribution hypothesis: The prototype-familiarity illusion. *Journal of Experimental Psychology: General, 131,* 96–115. http://dx.doi.org/10.1037/0096-3445.131.1.96

Whittlesea, B. W., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language, 29,* 716–732. http://dx.doi.org/10.1016/0749-596X(90)90045-2

Whittlesea, B. W., & Price, J. R. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition, 29,* 234–246. http://dx.doi.org/10.3758/BF03194917

Whittlesea, B. W., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 547–565. http://dx.doi.org/10.1037/0278-7393.26.3.547

Whittlesea, B. W., & Williams, L. D. (2001a). The discrepancy-attribution hypothesis: I. The heuristic basis of feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 3–13. http://dx.doi.org/10.1037/0278-7393.27.1.3

Whittlesea, B. W., & Williams, L. D. (2001b). The discrepancy-attribution hypothesis: II. Expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 14–33. http://dx.doi.org/10.1037/0278-7393.27.1.14

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17,* 799–806. http://dx.doi.org/10.1111/j.1467-9280.2006.01785.x

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Mahwah, NJ: Erlbaum.

Winkielman, P., Ziembowicz, M., & Nowak, A. (2015). The coherent and fluent mind: How unified consciousness is constructed from cross-modal inputs via integrated processing experiences. *Frontiers in Psychology, 6,* 83.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121,* 262–276. http://dx.doi.org/10.1037/a0035940

Wolk, D. A., Schacter, D. L., Berman, A. R., Holcomb, P. J., Daffner, K. R., & Budson, A. E. (2004). An electrophysiological investigation of the relationship between conceptual fluency and familiarity. *Neuroscience Letters, 369,* 150–155. http://dx.doi.org/10.1016/j.neulet.2004.07.081

Wöllner, C., Deconinck, F. J. A., Parkinson, J., Hove, M. J., & Keller, P. E. (2012). The perception of prototypical motion: Synchronization is enhanced with quantitatively morphed gestures of musical conductors. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 1390–1403. http://dx.doi.org/10.1037/a0028130

Young, D. R., & Bellezza, F. S. (1982). Encoding variability, memory organization, and the repetition effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 545–559. http://dx.doi.org/10.1037/0278-7393.8.6.545

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences, 17,* 263–271. http://dx.doi.org/10.1016/j.tics.2013.04.004

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9,* 1–27. http://dx.doi.org/10.1037/h0025848

Zajonc, R. B. (1998). Emotions. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 591–632). Boston, MA: McGraw-Hill.

Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science, 10,* 224–228. http://dx.doi.org/10.1111/1467-8721.00154

Zebrowitz, L. A., White, B., & Wieneke, K. (2008). Mere exposure and racial prejudice: Exposure to other-race faces increases liking for strangers of that race. *Social Cognition, 26,* 259–275. http://dx.doi.org/10.1521/soco.2008.26.3.259

Zeineh, M. M., Engel, S. A., Thompson, P. M., & Bookheimer, S. Y. (2003). Dynamics of the hippocampus during encoding and retrieval of face-name pairs. *Science, 299,* 577–580. http://dx.doi.org/10.1126/science.1077775

Received May 3, 2016
Revision received February 12, 2017
Accepted February 13, 2017 ■