UNIVERSITY OF CALIFORNIA, SAN DIEGO

Effects of Processing Dynamics on Social Perception, Judgment, and Action

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Psychology and Cognitive Science

by

Evan Walker Carr

Committee in charge:

Professor Piotr Winkielman, Chair
Professor David Barner
Professor Seana Coulson
Professor Craig R. M. McKenzie
Professor Christopher Oveis

2016

The Dissertation of Evan Walker Carr is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

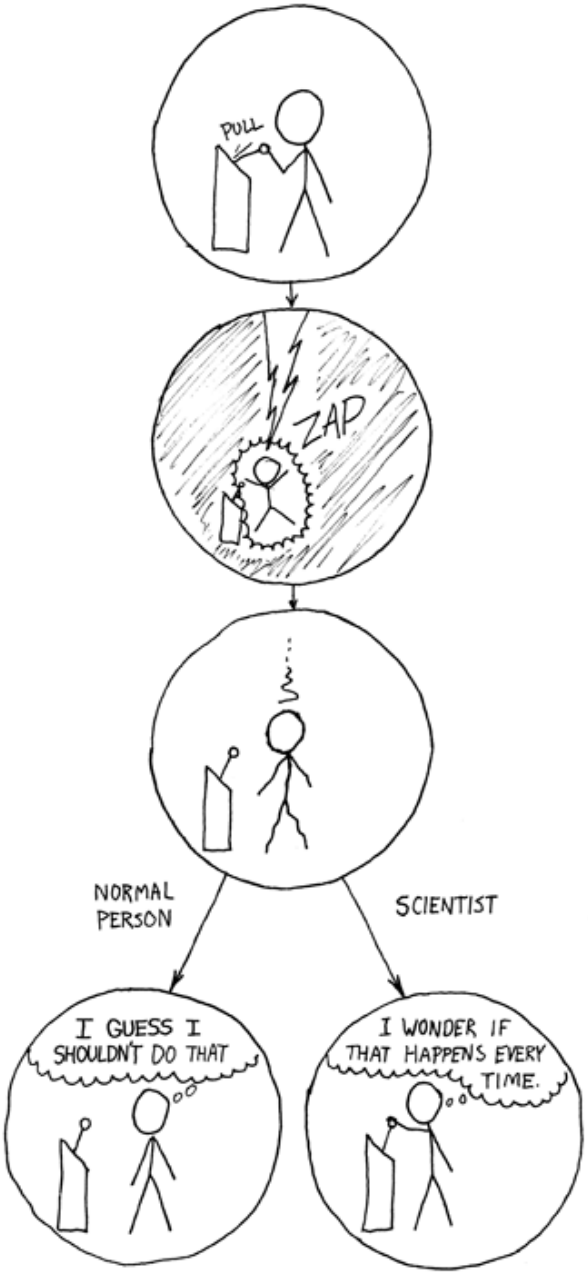
Chair

University of California, San Diego

2016

DEDICATION

For all those that have helped to mold my mind over the past 5 years.



Credit: http://xkcd.com/242/

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

To Piotr and Chris O. — for being incredibly generous mentors that provided me with invaluable thoughts, experiences, and opportunities. My appreciation is infinite.

To the other members of my dissertation committee — Craig, Dave, and Seana — for investing your time and effort in improving my work. I hope I can somehow return the favor someday.

To my loving family — Ryan, Jeff, and JoAnna — for supporting my every whim (no matter how ridiculous) and acting as one of the few constants in my life. You're all an inspiration.

To Christina (Teenies // Pepper) — for making my life over the past couple years a million times better than the 25 that came before it. I love you.

To Scott — for being a best friend, as well as an incredible example. Our conversations are among my most treasured memories. I'm 100% certain that you'll be immensely successful, and I can't wait to see where your creativity takes you. Looking forward to more awesome experiences with you.

To the other amazing friends that I've made in SD — especially Jordie, Jon, Gruberg, Brad, Andy, Rob, Liam, Camille, and Sirawaj — who made the journey more memorable and entertaining than I could have imagined.

To my lab-mates — including Liam, Andy, Galit, and Rob — for acting as a reliable resource and sounding board for basically everything I needed.

To the Department of Defense (DoD), American Society for Engineering Education (ASEE), and Army Research Office (ARO) for supporting my PhD work through the National Defense Science and Engineering Graduate (NDSEG) Fellowship. This award was a wonderful opportunity.

And to the creators of open-source packages that compose software like R, Python, Octave, and GIMP. Without you, my work would not have been possible.

Chapter 1 is, in full, under review for publication of the material. Carr, Evan W.; Pecher, Diane; Zeelenberg, Rene; Halberstadt, Jamin; Winkielman, Piotr. The dissertation author was the primary investigator and author of this material.

Chapter 2 is, in full, under review for publication of the material. Carr, Evan W.; Brady, Timothy, F.; Winkielman, Piotr. The dissertation author was the primary investigator and author of this material.

Chapter 3 is, in full, in press for publication in *Journal of Experimental Psychology: Human Perception and Performance*. Carr, Evan W.; Hofree, Galit; Sheldon, Kayla; Saygin, Ayse P.; Winkielman, Piotr. The dissertation author was the primary investigator and author of this material.

Chapter 4 is, in full, a reprint of the material as it appears in *Emotion*. Carr, Evan W.; Rotteveel, Mark; Winkielman, Piotr, 2016. The dissertation author was the primary investigator and author of this paper.

VITA
Evan Walker Carr

Education

2016	Doctor of Philosophy: Psychology & Cognitive Science *University of California, San Diego*
2013	Master of Arts: Psychology *University of California, San Diego*
2011	Bachelor of Science: Marketing, Strategy, and Information Systems (specialization) *Cornell University (School of Hotel Administration)*

Selected Publications

Vogel, T., **Carr, E. W.**, & Winkielman, P. (under review). Think global, prefer local: Category structure determines attractiveness of global and local prototypes.

Carr, E. W., Brady, T. F., & Winkielman, P. (under review). Are you smiling or have I seen you before? Familiarity makes faces look happier.

Carr, E. W., Pecher, D., Zeelenberg, R., Halberstadt, J., & Winkielman, P. (under revision). The ugliness-in-averageness effect: Tempering the warm glow of familiarity. *Journal of Personality and Social Psychology*.

Carr, E. W., Hofree, G., Sheldon, K., Saygin, A. P., & Winkielman, P. (in press). Is that a human? Categorization (dis)fluency drives evaluations of agents ambiguous on human-likeness. *Journal of Experimental Psychology: Human Perception and Performance.*

Carr, E. W., Kever, A., & Winkielman, P. (in press). Embodiment of emotion and its situated nature. In A. Newen, L. de Bruin, & S. Gallagher (Eds.), *The Oxford Handbook of Cognition: Embodied, Embedded, Enactive, and Extended.* Oxford University Press.

Farmer, H., **Carr, E. W.**, Svartdal, M., Winkielman, P., & Hamilton, A.F.C. (2016). Status and power do not modulate automatic imitation of intransitive hand movements. *PLoS One*, *11*(4), e0151835.

Owen, H. E., Halberstadt, J., **Carr, E. W.**, & Winkielman, P. (2016). Johnny Depp, reconsidered: How category-relative processing fluency determines the appeal of gender ambiguity. *PLoS One, 11*(2), e0146328.

Carr, E. W., Rotteveel, M., & Winkielman, P. (2016). Easy moves: Perceptual fluency facilitates approach-related action. *Emotion, 16*(4), 540-552.

*Kogan, A., *Oveis, C., **Carr, E. W.**, Gruber, J., Mauss, I. B., Shallcross, A., Impett, E. A., van der Löwe, I., Hui, B., Cheng, C., & Keltner, D. (2014). Vagal activity is quadratically related to prosocial traits, prosocial emotions, and observer perceptions of prosociality. *Journal of Personality and Social Psychology, 107*(6), 1051-1063. [*co-first authors]

Carr, E. W. & Winkielman, P. (2014). When mirroring is both simple and "smart": How mimicry can be embodied, adaptive, and non-representational. *Frontiers in Human Neuroscience, 8*(505).

Carr, E. W., Korb, S., Niedenthal, P., & Winkielman, P. (2014). The two sides of spontaneity: Movement onset asymmetries in facial expressions influence social judgments. *Journal of Experimental Social Psychology, 55*, 31–36.

Carr, E. W., Winkielman, P., & Oveis, C. (2014). Transforming the mirror: Power fundamentally changes facial responding to emotional expressions. *Journal of Experimental Psychology: General, 143*(3), 997-1003.

ABSTRACT OF THE DISSERTATION

Effects of Processing Dynamics on Social Perception, Judgment, and Action

by

Evan Walker Carr

Doctor of Philosophy in Psychology and Cognitive Science

University of California, San Diego, 2016

Professor Piotr Winkielman, Chair

Information processing is required for any social thought, decision, or action. Most past and current research in social cognition focuses solely on the content of the information being conveyed. While this is clearly important, in this dissertation, I investigate how basic social responses (i.e., rapid perceptions, judgments, and actions) are impacted by the dynamics of information processing (i.e., its ease, speed, or coherence). To do this, I examine two key determinants of processing dynamics — *familiarity* (prior stimulus experience) and *fluency* (ease of stimulus processing). First, Chapter 1 provides a systematic investigation of how familiarity influences the appeal of facial blends. Even though facial blends are usually deemed more attractive than their constituent individuals, Chapter 1 demonstrates that this classic beauty-in-averageness effect reverses when the individuals are highly familiar (thus, an

ugliness-in-averageness effect). Second, Chapter 2 extends the examination of familiarity to basic effects on perception, in showing that facial expressions from familiar individuals appear happier. These results also suggest that the familiarity-positivity effect functions by selectively enhancing positive stimulus features, rather than reducing negative stimulus features. Third, Chapter 3 moves towards gauging how categorization fluency (or the effort needed to determine category membership) influences the seemingly automatic discomfort people feel towards "mixed" agents (or those containing human and non-human features, like androids). Chapter 3 shows that classifying on the ambiguous human-likeness dimension makes the mixed agents (androids) more disfluent, and in turn, more disliked. Therefore, these results offer evidence that negative reactions to mixed agents are not obligatory, but instead are dependent on the surrounding judgment and context. Finally, Chapter 4 explores the link between fluency and motivation-related action. These experiments demonstrate that fluency elicits context-sensitive approach action-tendencies (i.e., RTs to initiate arm flexion), which are accompanied by physiological responses indicative of positive affect (i.e., increased smiling and reduced frowning, via facial electromyography). Taken together, the current dissertation shows that familiarity and fluency are flexibly embedded into our rapid perceptions, judgments, and actions towards social stimuli.

GENERAL INTRODUCTION

One of the most important signatures of human cognition involves the role of subjective experience — or the ability to recognize, filter, and apply how it "feels" to undergo or interact with something (Dolan, 2002; Gover, 1996; James, 1890; Solms & Turnbull, 2002). Certainly, many of these processes can occur at the level of conscious awareness, as with introspection (Boring, 1953; Ellis, 1991; Wilson, 2003), problem solving (Baars & Franklin, 2003; Mayer, 1992; McLeod & Adams, 2012), or emotion-regulation strategies (Gross, 1997, 2008). However, some aspects of subjective experience can also be unconscious (or at the level of "fringe consciousness"; Reber, Wurtz, & Zimmermann, 2004). In this case, indistinct feelings signal background information on the contents of conscious attention, leading to changes in subsequent affect and cognition (Brown, 2000; Feldman-Barrett, Niedenthal, & Winkielman, 2005; James, 1890; Schooler, Mrazek, Baird, & Winkielman, 2015; Winkielman & Schooler, 2011).

Critically though, these subtle experiences can have downstream effects on social evaluations, judgments, motivations, and actions (Winkielman & Berridge, 2004, 2009). One important example of this comes with *processing dynamics*. Information processing is required for any social thought, decision, or action. However, most past (and even current) research focuses solely on the content of the information being conveyed. While this is clearly important, in this dissertation, I examine on how the dynamics of information processing (i.e., its ease, speed, or coherence) can influence how that information is perceived, evaluated, and acted upon. More specifically, I will investigate two key factors for the dynamics of information processing (along with their effects on social perceptions, judgments, and actions) — *familiarity* and *fluency*.

Familiarity

Familiarity (or the amount of experience and "sense of knowing" with a stimulus) is among the most important factors for social information processing (Zajonc, 2001). Simply put,

the greater the familiarity one has with a stimulus, the greater their liking and preference for it. This *mere exposure effect* (or the phenomenon of increased preference from unreinforced stimulus repetition) dates back over a century to Titchener's (1915) first observations about the "warm glow of familiarity." Since then, familiarity and mere exposure have been consistently examined and applied within psychology, neuroscience, sociology, and business (Baker, 1999; Balogh, & Porter, 1986; Obermiller, 1985; Pettigrew & Tropp, 2008; Tremblay, Inoue, McClannahan, & Ross, 2010). Mere exposure effects are also robust across different stimuli (e.g., words, images, sounds, and faces) and modalities (e.g., vision, audition, touch, and smell) (Zajonc, 1968, 2001), while being subject to some important boundary conditions (Bornstein, 1989).

The connection between familiarity and liking could occur for many reasons (for reviews, see Fang, Singh, & Ahluwalia, 2007; Moreland & Topolinski, 2010). One possibility has to do with learning and uncertainty, where unreinforced repetition associates the stimulus with an absence of negative consequences (Lee, 2001; Zajonc, 1968). However, most of the prominent explanations ascribe to cognitive models, where repetition facilitates ease and efficiency in processing (i.e., increased fluency; Bornstein & D'Agostino, 1992; Butler & Berry, 2004; Klinger & Greenwald, 1994; Whittlesea & Price, 2001; Winkielman, Schwarz, Fazendeiro, & Reber, 2003). There is also evidence for affective models of mere exposure, such that familiar stimuli elicit hedonic physiological responses indicative of liking (Harmon-Jones & Allen, 2001).

Note that familiarity can also vary in its specific nature. Objective familiarity refers to the actual stimulus history (i.e., how many times it has actually been encountered), while subjective familiarity refers to a "sense of knowing" for the stimulus (i.e., a feeling for if and how much it has been encountered before). These distinctions are important because the relation between familiarity and preference usually concerns subjective familiarity. As mentioned, subjective familiarity is also often (though not always) linked to processing fluency, as discussed

later. Given that a previous encounter with an item is thought to increase its activation, it follows that subsequent efficiency or coherence of processing would also increase upon reactivation.

Moreover, familiarity has a wide-ranging impact on social judgment and evaluation. Not only do preferences increase for previously encountered social stimuli (like faces; Peskin & Newell, 2004), but these preferences generalize to stimuli that are similar to ones seen previously, yet objectively new (Gordon & Holyoak, 1983; Whittlesea, 2002). Such generalization effects have also been obtained for social stimuli, such as faces (Rhodes, Halberstadt, & Brajkovich, 2001), where 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). These findings are discussed as evidence for "structural mere exposure effects" — with increasing levels of distortion from the mere-exposed pattern associated with reduced liking for those new stimuli. These generalization effects are theoretically important, as they suggest the role of cognitive factors in this seemingly basic phenomenon (Zajonc, 2001). They also offer a path towards changing real-world social preferences that extend beyond the specific individuals engaged in personal interactions, such as with the positive effects of intergroup contact (Pettigrew & Tropp, 2008).

In turn, it is vital to understand the nature, mechanisms, and limitations of familiarity effects on social stimuli — especially with effects that seem highly replicable (e.g., the attractiveness of facial blends; Chapter 1) or perceptually low-level (e.g., rapidly judging the emotional content of others' facial expressions; Chapter 2).

Fluency

The connection between emotion, cognition, and psychophysiology has also been successfully explored by examining *fluency* (i.e., objective or subjective ease in processing effort associated with a stimulus; Schwarz, 1998). Much evidence shows that fluency increases positive

affective evaluations, whereas disfluency leads to devaluation, on both behavioral and physiological levels (Winkielman, Schwarz, Fazendeiro, & Reber, 2003). Essentially, this idea proposes that the facilitation of perceptual or conceptual processing is associated with a genuine hedonic boost in positive affect (Winkielman & Cacioppo, 2001). Here, objective fluency refers to a concrete outcome of easy processing (e.g., faster recognition RTs, reduced cognitive load or resource demands, increased accuracy, etc.), while subjective fluency indicates the experience associated with such efficient processing (e.g., decreased mental effort; Reber, Wurtz, & Zimmermann, 2004). To account for these findings, the *hedonic fluency model* posits that (dis)fluency generates diffuse affect that can be used to form a range of related social judgments and evaluations. More specifically, easy processing elicits mild positive affect, which is then (mis)attributed to the target stimulus (Winkielman, Schwarz, Fazendeiro, & Reber, 2003). This positive affect presumably emerges because fluency probabilistically indicates lower conflict and greater coherence. As mentioned before, note that liking effects from familiarity are often tied to fluency, since processing ease should be facilitated with repeated exposure to the stimulus.

Fluency can also be manipulated both on the level of percepts and concepts. With perceptual fluency, this usually denotes lower-level changes associated with stimulus structure or form (e.g., priming, duration, clarity, contrast, or typicality; Reber, Fazendeiro, & Winkielman, 2002; Winkielman, Halberstadt, Fazendeiro, & Catty, 2006). For instance, when patterns (from dots to faces) deviate from a category "prototype," they are relatively devalued, compared to more average stimuli (Halberstadt & Rhodes, 2000, 2003; Halberstadt & Winkielman, 2013). Therefore, manipulating perceptual fluency involves changing stimulus features to make it more or less difficult to process. With conceptual fluency, this usually refers to higher-level changes associated with stimulus meaning (e.g., semantic priming, predictability, or categorization; Winkielman, Huber, Kavanagh, & Schwarz, 2012). Here, fluency is manipulated by changing how easy or difficult it is to extract meaning from the stimulus, rather than actually changing its

features. As an example, the effort needed to determine category membership (also called categorization fluency; Halberstadt & Winkielman, 2013) is ultimately task-dependent, and processing difficulty depends on which (un)ambiguous feature dimensions are highlighted by the current task. In other words, if a stimulus is ambiguous on some dimension, it will elicit *dis*fluency (and negative affect), but only in contexts requiring categorization on that particular dimension. To illustrate, Owen, Halberstadt, Carr, & Winkielman (2016) demonstrated that mixed-gender faces are only deemed relatively unattractive when first categorized on the central ambiguous dimension (gender), rather than an ancillary unambiguous dimension (race).

Crucially though, fluency affects a variety of real-world social responses, on both the behavioral and physiological levels. Behaviorally, greater perceptual and conceptual fluency enhances positive evaluative judgments, like basic preferences (Winkielman & Cacioppo, 2001), product choices and decisions (Novemsky, Dhar, Schwarz, & Simonson, 2007), ratings of attractiveness and trustworthiness (Winkielman, Olszanowski, & Gola, 2015), brand assessments (Lee & Labroo, 2004), and even stock purchases (Alter & Oppenheimer, 2006). Physiologically, previous research has demonstrated fluency to induce a genuine positive hedonic response, as with low-level incipient smiling (via increased reactivity over the zygomaticus major, or the "smiling muscle" that pulls up the corners of the mouth; Tassinary, Cacioppo, & Vanman, 2007; Winkielman & Cacioppo, 2001). Parallel effects have also been reported with reduced frowning (via decreased reactivity over the corrugator supercilii, or the "frowning muscle" that furrows the brow; Tassinary, Cacioppo, & Vanman, 2007), which likely indexes reduced negative affect and relaxed mental effort (Topolinski, Likowski, Weyers, & Strack, 2009). These physiological measures are especially valuable because they allow for the dissociation between objective task demands and subjective evaluative judgments (Von Helversen, Gendolla, Winkielman, & Schmidt, 2008), as well as the measurement of timing differences between perceptual and conceptual fluency (Wang, Li, Gao, Xiao, & Guo, 2015).

Given these findings, it is important to gauge when, why, and how fluency effects translate to basic social responses. For example, while most of the previous fluency research focuses on higher-level evaluations of stimuli (e.g., liking ratings), it has never been tested whether fluency can also impact motivation-related action (i.e., the tendency to approach or avoid the stimulus; Chapter 4). Further, new experiments need to assess the boundary conditions for fluency effects to emerge, such as with type of stimulus features in-question (e.g., judgments of human-likeness vs. judgments of color ambiguity; Chapter 3) or the type of judgment context (e.g., affective decisions of "good or bad" vs. non-affective decisions of "living or non-living"; Chapter 4).

Aims of the current dissertation

The main goal of this dissertation is to examine the roles of familiarity and fluency in shaping basic social responses (i.e., perceptions, judgments, and actions that occur quickly and/or with little conscious awareness). Across 13 experiments, I investigate how familiarity and fluency can transform classic phenomena in social cognition (Chapters 1 and 3) and rudimentary reactions to neutral and emotional stimuli (Chapters 2 and 4). Overall, these studies provide robust evidence that even low-level effects and processes in social cognition can be mitigated, amplified, or reversed based on the dynamics of information processing. To do this, the current dissertation is structured according to the following four chapters.

Chapter 1 provides a systematic inquiry into how familiarity impacts the appeal of facial blends. Blends (morphs) of individual face stimuli are usually deemed more attractive than their constituent individuals — known as the classic *beauty-in-averageness (BiA) effect*. The attractiveness of facial blends is also presumably linked to their perceived familiarity, which leads to greater liking. However, based on modern theories of memory, we predicted that the BiA effect should only occur when the contributing individuals are weakly encoded (thus prioritizing

a global prototypical representation, thereby making blends appear more familiar and attractive). When the individuals are strongly encoded, memory theories would instead predict a relative decrease in familiarity and preference for blends — or a novel phenomenon we term the *ugliness-in-averageness (UiA) effect.* In four experiments, we show that the BiA effect emerges with weak learning on individual faces, and the increased attractiveness for blends is driven by their familiarity (Experiment 1). In contrast, when participants are first strongly trained on a subset of individual faces (using a social name-learning task), a UiA effect emerges for trained faces (i.e., blends of trained individuals are rated as *less* attractive than the trained individuals; Experiment 2). We also demonstrate the mechanistic role of familiarity in the UiA effect (Experiment 3) and show that simple perceptual (as opposed to social) learning is sufficient to generate the UiA effect (Experiment 4). These results highlight that familiarity-based memory processes can reshape seemingly immutable patterns of facial attractiveness, when combining the effects of mere exposure (stimulus repetition) and blending (stimulus averaging).

Chapter 2 extends the examination of familiarity to basic effects on perception. Mere exposure is a standard effect in social cognition, yet the basic nature for how familiarity creates positivity remains largely unknown. Here, we use two different tasks to measure early perceptual effects (Experiment 1) and rapid classification judgments (Experiment 2) on affective facial expressions. In Experiment 1, using a paradigm where participants' responses were orthogonal to happiness to avoid response facilitation, we found that trained (familiar) faces were deemed happier than untrained (novel) faces. In Experiment 2, we replicated this effect with a rapid "happy or angry" categorization task. Using psychometric function fitting, we found that participants needed less actual happiness to be present in trained (compared to untrained) faces in order to classify them as happy. With both experiments, we show that the familiarity-positivity effects operate through selective enhancement of positive stimulus features (rather than reduction

of negative stimulus features). Critically, our results help to dissociate prominent models of mere exposure, in demonstrating how familiar faces can appear happier.

Chapter 3 gauges how categorization disfluency influences evaluations of "mixed" agents (or those that contain both human and non-human features). Essentialism theories suggest that mixing human and non-human categories violates "natural kinds," while perceptual theories propose that such mixing creates incompatible cues or "mismatch" effects. Most theories suggest that mixed agents should obligatorily elicit discomfort. Alternatively, in three experiments, we demonstrate that the discomfort associated with mixed agents is partially driven by disfluent categorization on ambiguous features that are pertinent to that agent (i.e., whether they are human or non-human). Participants classified three different agents (humans, androids, and robots) either on the human-likeness dimension or a control dimension and then evaluated them. Human-likeness categorization made the mixed agent (android) more disliked, and disfluency mediated this negative affective reaction. Crucially, devaluation only resulted from disfluency on human-likeness and not from an equally disfluent color dimension. We argue that negative consequences on evaluations of mixed agents arise from integral disfluency (or features that are relevant to the judgment at-hand, like ambiguous human-likeness), whereas no negative effects stem from incidental disfluency (or features that do not bear on the current judgment, like ambiguous color backgrounds). These findings support a top-down account for why, when, and how mixed agents elicit conflict and discomfort.

Chapter 4 explores the link between fluency and motivation-related action. Many studies have already shown that processing fluency impacts liking judgments and physiological reactions, but it remains unknown whether fluency translates to action-tendencies. We used four experiments to investigate this action effect, its boundary conditions, and associated affective and physiological responses. We found faster approach movements (RTs to initiate arm flexion) to perceptually fluent stimuli when participants rapidly classified in an affective judgment context

(i.e., whether the stimulus was “good or bad”; Experiments 1 and 3). Interestingly, this fluency effect on action dissipated within non-affective judgment contexts (i.e., whether the stimulus was “living or non-living”; Experiments 2 and 4), even though perceptual fluency still enhanced liking judgments (all experiments). Finally, while we only observed the fluency-action effect in affective judgment contexts, perceptual fluency led to a physiological hedonic response in both affective and non-affective judgment contexts (i.e., increased smiling and decreased frowning, via facial electromyography; Experiments 3 and 4). This suggests that the affective response can dissociate from the motivation-related action tendency, according to the judgment context. Collectively, these results reveal that perceptual fluency can flexibly and implicitly shape motor responses.

References

Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences, 103*(24), 9369–72.

Baars, B. J., & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences, 7*(4), 166-172.

Baker, W. E. (1999). When can affective conditioning and mere exposure directly influence brand choice? *Journal of Advertising, 28*(4), 31-46.

Balogh, R., & Porter, R. H. (1986). Olfactory preferences resulting from mere exposure in human neonates. *Infant Behavior and Development, 9*(4), 395-401.

Boring, E. G. (1953). A history of introspection. *Psychological Bulletin, 50*(3), 169.

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. (1992). Stimulus recognition and the mere exposure effect. *Journal of Personality and Social Psychology, 63*, 545-552.

Brown, S. R. (2000). Tip-of-the-tongue phenomena: An introductory phenomenological analysis. *Consciousness and Cognition, 9*(4), 516-537.

Butler, L. T. and Berry, D. C. (2004). Understanding the relationship between repetition priming and mere exposure. *British Journal of Psychology, 95*(4). 467-487.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science, 298*(5596), 1191-1194.

Ellis, C. (1991). Sociological introspection and emotional experience. *Symbolic Interaction, 14*(1), 23-50.

Fang, X., Singh, S., & Ahluwalia, R. (2007). An examination of different explanations for the mere exposure effect. *Journal of Consumer Research, 34*(1), 97-103.

Feldman-Barrett, L., Niedenthal, P., & Winkielman, P. (2005). *Emotion and Consciousness*. New York, NY: Guilford Press.

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.

Gover, M. R. (1996). The embodied mind: Cognitive science and human experience. *Mind, Culture, and Activity, 3*(4), 295-299.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*(3), 271.

Gross, James J. (Ed.) (2007). *Handbook of emotion regulation*. New York, NY: Guilford Press.

Halberstadt, J., & Rhodes, G. (2000). The attractiveness of non-face averages: Implications for an evolutionary explanation of the attractiveness of average faces. *Psychological Science, 11*(4), 285-289.

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*(1), 149-156.

Halberstadt, J. & Winkielman, P. (2013). When good blends go bad: How fluency can explain when we like and dislike ambiguity. In C. Unkelbach & R. Greisfelder. *The experience of thinking: How feelings from mental processes influence cognition and behavior* (pp. 133–151). New York, NY: Psychology Press.

Harmon-Jones, E., & Allen, J. J. (2001). The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality and Social Psychology Bulletin, 27*(7), 889-898.

James, W. (1890). *The principles of psychology* (Vol. 1). New York, NY: Holt.

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.

Lee, A. Y. (2001). The mere exposure effect: An uncertainty reduction explanation revisited. *Personality and Social Psychology Bulletin, 27*(10), 1255-1266.

Lee, A. Y., & Labroo, A. A. (2004). The effect of conceptual and perceptual fluency on brand evaluation. *Journal of Marketing Research, 41*(2), 151–165.

Mayer, R. E. (1992). *Thinking, problem solving, cognition*. WH Freeman/Times Books/Henry Holt & Co.

McLeod, D. B., & Adams, V. M. (Eds.) (2012). *Affect and mathematical problem solving: A new perspective*. Springer Science & Business Media.

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

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of Marketing Research, 44*, 347–356.

Obermiller, C. (1985). Varieties of mere exposure: The effects of processing style and repetition on affective response. *Journal of Consumer Research, 12*(1), 17-30.

Owen, H. E., Halberstadt, J., Carr, E. W., & Winkielman, P. (2016). Johnny Depp, reconsidered: How category-relative processing fluency determines the appeal of gender ambiguity. *PLoS ONE, 11*(2), e0146328.

Peskin, M., & Newell, F. N. (2004). Familiarity breeds attraction: Effects of exposure on the

attractiveness of typical and distinctive faces. *Perception, 33*(2), 147-157.

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*(6), 922–934.

Reber, R., Fazendeiro, T. A., & Winkielman, P. (2002). Processing fluency as the source of experiences at the fringe of consciousness. *Psyche, 8*(10), 175-188.

Reber, R., Wurtz, P., & Zimmermann, T. D. (2004). Exploring "fringe" consciousness: The subjective experience of perceptual fluency and its objective bases. *Consciousness and Cognition, 13*(1), 47-60.

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition, 19*(1), 57-70.

Schooler, J. W., Mrazek, M. D., Baird, B., & Winkielman, P. (2015). Minding the mind: The value of distinguishing among unconscious, conscious, and metaconscious processes. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA handbook of personality and social psychology, Vol. 1. Attitudes and social cognition* (pp. 179-202). Washington, DC: American Psychological Association.

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*(1), 50-64.

Solms, M., & Turnbull, O. (2002). *The brain and the inner world: An introduction to the neuroscience of subjective experience*. Karnac Books.

Tassinary, L. G., Cacioppo, J. T., & Vanman, E. J. (2007). The skeletomotor system: Surface electromyography. In J. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 267–302). New York, NY: Cambridge University Press.

Topolinski, S., Likowski, K. U., Weyers, P., & Strack, F. (2009). The face of fluency: Semantic coherence automatically elicits a specific pattern of facial muscle reactions. *Cognition and Emotion, 23*, 260–271.

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*(4), e10283.

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science, 21*(6), 779-785.

Von Helversen, B., Gendolla, G. H., Winkielman, P., & Schmidt, R. E. (2008). Exploring the hardship of ease: Subjective and objective effort in the ease-of-processing paradigm. *Motivation and Emotion, 32*(1), 1-10.

Wang, W., Li, B., Gao, C., Xiao, X., & Guo, C. (2015). Electrophysiological correlates associated with contributions of perceptual and conceptual fluency to familiarity. *Frontiers in Human Neuroscience, 9*, 321.

Whittlesea, B. W. A. & Price, J. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory and Cognition, 29*, 234-246.

Whittlesea, B.W.A. (2002). False memory and the discrepancy attribution hypothesis: The prototype-familiarity illusion. *Journal of Experimental Psychology: General, 131*, 96–115.

Wilson, T. (2003). Knowing when to ask: Introspection and the adaptive unconscious. *Journal of Consciousness Studies, 10*(9-10), 131-140.

Winkielman, P. & Berridge, K. C. (2004). Unconscious emotion. *Current Directions in Psychological Science, 13*, 120-123.

Winkielman, P. & Berridge, K. (2009). Unconscious emotion. In D. Sander & K. R. Scherer (Eds.), *Oxford Companion to the Affective Sciences* (pp. 395-396). Oxford, UK: Oxford University Press.

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology, 81*(6), 989–1000.

Winkielman, P., Halberstadt, J., Fazendeiro, T. & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17*(9), 799-806.

Winkielman, P., Huber, D.E., Kavanagh, L. & Schwarz, N. (2012). Fluency of consistency: When thoughts fit nicely and flow smoothly. In B. Gawronski & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 89-111). New York, NY: Guilford Press.

Winkielman, P. & Schooler, J.W. (2011). Splitting consciousness: Unconscious, conscious, and metaconscious processes in social cognition. *European Review of Social Psychology, 22*, 1–35.

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.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*, 1–27.

Zajonc, R.B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science, 10*, 224–228.

CHAPTER 1:

The ugliness-in-averageness effect: Tempering the warm glow of familiarity

Evan W. Carr, Diane Pecher, Rene Zeelenberg, Jamin Halberstadt, & Piotr Winkielman

(manuscript under review for publication)

Abstract

Mere exposure (i.e., stimulus repetition) and *blending* (i.e., stimulus averaging) are classic effects known to increase social preferences, including facial attractiveness. Both effects presumably occur because familiarity enhances liking. Prominent memory theories assume that target familiarity depends on the strength of its memory trace, similarity to the specific exposed exemplars, and similarity to any global representation. If so, blends (morphs) of individual stimuli should have greater familiarity and liking — or a *beauty-in-averageness effect (BiA).* However, this should only occur when the contributing individuals are weakly encoded, thus prioritizing a global prototypical representation. When the individuals are strongly encoded, memory theories predict a relative *decrease* in familiarity and preference for their blends — or a novel phenomenon we term the *ugliness-in-averageness effect (UiA).* We tested this novel theoretical prediction in four experiments, using the same stimulus set. Experiment 1 showed that with weak learning, participants rated morphs as more attractive than contributing individuals (i.e., BiA effect). Experiment 2 demonstrated that when participants were first strongly trained on a subset of individual faces (using a social name-learning task), a UiA effect emerged for trained faces — where morphs of trained individuals were rated as *less* attractive than the trained individuals. Experiment 3 showed that declines in familiarity for the trained morph (rather than inter-stimulus conflict) drove the UiA effect. Experiment 4 demonstrated that simple perceptual (as opposed to social) learning is sufficient to generate the UiA effect. Overall, these results highlight that memory processes can fundamentally reshape classic, seemingly immutable social preference phenomena.

Keywords: mere exposure, blending, attractiveness, familiarity, faces

Introduction

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 explicitly and implicitly shaped by the beauty associated with a person, group, object, or idea (Reber, Schwarz, & Winkielman, 2004; Rhodes & Zebrowitz, 2002). Consequently, understanding such preferences not only helps to illuminate the mechanisms underlying affect and cognition, but this also informs practical applications. The current paper addresses two classic determinants for preferences that have been discussed in psychology — *mere exposure* (i.e., stimulus repetition) and *blending* (i.e., stimulus averaging). Both processes have been shown to shape preference judgments, usually by increasing them. Here, we use these fundamental phenomena to shed light on the mechanisms linking familiarity and preference. More specifically, we explore predictions generated by modern memory models, which link familiarity (and thus, preference) to the similarity of the target to memory representations.

To test these predictions, we apply both of these preference manipulations to the same stimuli (i.e., when mere-exposed stimuli are blended together). Several alternative models predict that both mere exposure and blending should increase preferences in an additive fashion. However, our theoretical framework uniquely predicts that relative to the exposed exemplars, preferences for blends of mere-exposed stimuli should *decrease,* due to the loss of familiarity. Our findings demonstrate that mere exposure generates a familiarity-based preference, and blending actually reduces the preference for highly familiar individuals. To preview the current work, we offer background information on mere exposure, blending effects, and modern memory models.

Mere exposure

The *mere exposure effect* — or the phenomenon of increased preference from unreinforced stimulus repetition — is a psychology classic. It goes back at least 100 years, when Titchener (1915) made his observations about the "warm glow of familiarity." The landmark paper by Zajonc (1968) renewed the field's interest, and since then, mere exposure has been continually discussed in psychology textbooks, investigated in the labs, applied in social interventions, and utilized in business settings (Baker, 1999; Balogh, & Porter, 1986; Kouchaki, Smith-Crowe, Brief, & Sousa, 2013; Obermiller, 1985; Pettigrew & Tropp, 2008; Tremblay, Inoue, McClannahan, & Ross, 2010; Zajonc, 2001). The effect is robust across a wide range of stimuli (e.g., faces, words, sounds, images, etc.) and modalities (e.g., vision, audition, touch, smell, etc.), though subject to important boundary conditions (Bornstein, 1989). Aside from its practical importance, the mere exposure effect also offers a theoretical 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). Zajonc (1968) proposed a non-cognitive model where unreinforced repetition associates the stimulus with an absence of negative consequences. Indeed, there is some evidence that some novel stimuli induce negative affect, which gets reduced with mere repetition (e.g., Zebrowitz & Zhang, 2012). Nevertheless, there is much more evidence for cognitive models, in which 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). Interestingly, while the mere exposure effect is tied to the 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, but are actually objectively *new* (Whittlesea, 2002). Such generalization

effects have also been obtained for social stimuli, such as faces (Rhodes, Halberstadt, & Brajkovich, 2001), where 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). Interestingly, the generalization effect from mere exposure follows a similarity gradient between the original and test stimulus. Gordon & Holyoak (1983) systematically tested this idea, by first exposing participants to a certain subset of abstract and arbitrary stimuli (i.e., letter strings and complex visual patterns). Next, participants evaluated new stimuli — these new stimuli were similar to those exposed previously, but they were systematically distorted to different gradients away from the training stimuli. Even though all test stimuli were objectively new, participants' liking ratings showed a "structural mere exposure effect" — with increasing levels of distortion from the mere-exposed pattern associated with reduced liking for those new stimuli. These generalization effects are theoretically important, as they suggest the role of cognitive factors in this seemingly basic effect (Zajonc, 2001). They also offer a path towards changing real-world social preferences that extend beyond the specific individuals engaged in personal interactions. In fact, generalization of the mere exposure effect is discussed as one mechanism behind the positive effects of intergroup contact (Pettigrew & Tropp, 2008). As such, it is very important to understand the nature, mechanisms, and limitations of such effects.

Stimulus blending

The mere exposure effect relates to another classic phenomenon in the domain of preferences — *blending* (or stimulus averaging). Since the original observations by Galton (1879) on composite portraits, psychologists have explored how preferences are influenced by averaging or blending the stimuli (often through a method of morphing). Generally, averaging makes the stimulus more attractive, and this effect is most robustly established with faces

(Halberstadt, 2006; Langlois & Roggman, 1990; Rhodes & Tremewan, 1996). However, it also occurs within a variety of different modalities and stimuli (e.g., abstract dot patterns, colors, birds, cars, watches, fish, voices, gestures, etc.; Bruckert et al., 2010; Halberstadt & Rhodes, 2003; Winkielman, Halberstadt, Fazendeiro, & Catty, 2006; Wöllner et al., 2012).

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. Specifically, Langlois & Roggman (1990) point out that blending several faces makes the average face more similar to the facial prototype — or the central tendency of a local population of faces encountered by the participants. In fact, averaged faces become more or less attractive as a function of exposure to different populations of faces, suggesting the importance of learning processes (Principe & Langlois, 2012; Rubenstein, Kalakanis, & Langlois, 1999). Consistently, attractiveness of average faces is associated with their implicit familiarity (Peskin & Newell, 2004; Rhodes, Halberstadt, & Brajkovich, 2001). This fits with many studies that use abstract patterns (e.g., random dots), which show that exposure to multiple exemplars of a category allows participants to implicitly extract the prototype (i.e., category average). Such prototypes are also later preferred, as reflected in 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 memory for social psychological theories of preference. In this section, we argue that the relevant memory literature not only explains why these classic preference phenomena occur, but it also helps us 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, an interested reader can explore the memory literature, including its quantitative, computational, neuroscientific, and applied aspects, 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, the reader may refer to a review by Smith (1996).

Let us start with a couple of clarifications on terminology. First, *objective familiarity* refers to the actual stimulus history (i.e., how many times it 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*. Second, 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, re-processing 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.

Now, we will shift focus to a critical question — what elicits subjective familiarity? Prominent memory theories suggest that familiarity of a probe depends on the overall match between the probe and the set of items in memory to which it is compared (Gillund & Shiffrin, 1984; Hintzman, 1984). Familiarity of a probe is calculated from the similarity values of the probe with all traces in memory (or a relevant subset of traces). The strength of a memory trace (or the accuracy of the information stored in memory) determines the similarity between the probe and the memory trace. If the memory trace is weak (because only a few item features were stored correctly), the similarity between the probe and the memory trace will be lower than when the memory trace contains many correctly stored features. Thus, familiarity will be higher for

strong items than for weak items. In addition, memory traces for other items may also be more or less similar to the probe, contributing to the global familiarity of the probe in proportion to their degree of similarity. The similarity calculation is based on the number of matching features between the probe and each of the memory traces. Although the general methods vary, some kind of multiplicative function is used so that a close match to one item leads to higher familiarity than a moderate match to many items (Hintzman, 1986; Murphy, 2002). This helps to explain the mere exposure effect, because stronger memory traces for actually studied items will result in a better match, and thus, higher familiarity values.

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. 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. Hintzman (1986) proposed that each event leads to a trace in memory, assuming that sufficient attention was paid to that event (also see Nosofsky, 1986). On this view, semantic and episodic memory are one system. To identify a probe as a particular item, the similarity of the probe to all traces of that item is compared to the similarity of the probe to all traces of all other items — and when that ratio is high enough, the probe is identified as that item. For example, to identify a picture of a cat as your neighbor's cat, the similarity of the picture to all traces of your neighbor's cat is compared to the similarity of the picture to memory traces for all other items (e.g., cat traces, dog traces, bike traces, etc.). If the similarity to your neighbor's cat is higher than the similarity to the other items, the picture is identified as your neighbor's cat. With categorization, the same mechanism works to compare probe similarity of all central category exemplar traces to the probe similarity of all traces for other category exemplars.

Although it is sometimes assumed that prototypes (or "gist" representations) are formed

to represent categories, many have argued that models with only exemplar representations might perform just as well (or perhaps even better) at explaining categorization performance (compared to models that assume the prototype is stored in memory; Barsalou, 1990; Johansen & Palmeri, 2002; Love, 2013; Medin & Schaffer, 1978; Murphy, 2002). The reason that the prototype and exemplar models differ only slightly is that a prototype is a summary representation of the exemplars that have been encountered — and thus, to a large extent, the prototype contains the same information as the collection of exemplar traces. However, exemplar models might fare better in cases where individual exemplars contribute to performance, given that the particular combination of features is retained (whereas in a prototype representation, the particular feature combinations are lost).

This mechanism explains the BiA effect as resulting from the following process. First, participants are incidentally exposed to many exemplars using minimal exposure, which results in formation of very weak individual traces. Later, participants are presented with the blend probe (or morph) that is more similar to all memory traces than any probe of individual faces. The more similarity the blend has to all other face traces, the more familiar (and preferred) it is compared to the weakly learned individual faces. Consistent with this account, traditional BiA paradigms use only single incidental exposure to individuals. Further, evidence shows that BiA effect increases with the number of faces that go into the blend. In fact, the classic Langlois & Roggman (1990) paper only observed a clear BiA effect when averaging many individuals (i.e., eight or more), which may make the morph appear very familiar (compared to a morph that only averages two individuals).

How do exposure and blending effects interact to drive familiarity and preferences?

With the above principles in mind, we can now derive several predictions regarding the combined effects of exposure and blending on familiarity and preferences. First, a more basic

prediction is that strong learning on individual exemplars should result in the traditional mere exposure effect (i.e., a growth in familiarity and liking). Second, our central (and more nuanced) prediction is that the effects of blending two unknown faces should depend on the larger memory context. More specifically, when participants have weak memory traces for individual exemplars, there should be a traditional BiA effect because the blend will be more familiar than the individuals. When participants have no memory traces for individual exemplars, there should be no BiA effect (since the blend probe is not similar to anything). Critically, if the memory traces for these individual exemplars are strengthened (i.e., through learning), the BiA effect should be relatively reduced, since the probe for an individual face now has a high similarity to its *own* memory trace. Moreover, familiarity increases differentiation, which describes the ability to distinguish items from similar distractors (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Therefore, as memory traces for individual faces become stronger, people improve at differentiating that face from other similar faces.

Our central prediction concerns this blending of highly familiar exemplars. When the blend of highly familiar individuals is presented, the benefit of familiarity due to higher global similarity should be weighed against the cost of *dis*similarity between the blend and the well-learned individual exemplars. As a result, the morph will be *less* familiar than the exposed individuals, so our theoretical perspective derived from the memory literature actually predicts an *ugliness-in-averageness (UiA) effect*. Note that the morph should still benefit from some similarity to the exposed exemplars, and thus have greater familiarity and liking than unfamiliar stimuli. As such, "ugliness" is defined here as a relative decline, rather than an absolute loss. In other words, blending familiar exemplars should reduce the benefits of their exposure, rather than bring the morph below the original attractiveness level of unfamiliar face blends.

This prediction fits well with the previous memory literature. The most relevant examples come from paradigms that investigated the effects of item strength (i.e., learning) on

responses to the original items and "blended" items (i.e., items that are objectively new but include features of the individual items). In one memory paradigm, participants first studied words like "blackmail" and "jailbird," and then are 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 (i.e., weak pairs) or several times (i.e., strong pairs). Next, they were asked about intact pairs, rearranged pairs, and control items (Kelley & Wixted, 2001). The key finding across these paradigms is that participants showed an elevated false alarm rate to the "blended items" (e.g., "blackbird" or *fish-clock*). And crucially, the false alarm rate is both lower than the recognition of actual presented items and reduced (but not eliminated) when participants have 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.

Moreover, this prediction also aligns with the previously discussed work on mere exposure generalization, which found reduced liking benefits with increased dissimilarity of the probe (Gordon & Holyoak, 1983). Note, however, that this work only used graded (increased or decreased) distortions from an abstract prototype, such that there was no feature combination across familiar or unfamiliar stimuli. A more recent and direct example comes from a study that blended faces of celebrities and showed them to participants either in the country where these celebrities were known or in a different country where they were unknown (Halberstadt, Pecher, Zeelenberg, Wai, & Winkielman, 2013). The study found that morphs of two celebrity faces were more attractive than the individual celebrities used to generate them (i.e., traditional BiA effect) but only when those celebrity individuals were unknown in the participants' country (i.e., were famous in another country). Critically, when the morphs were composed of two celebrities from the participants' home country, those morphs were judged as *less* attractive than the individual celebrities. This initial study is consistent with our hypothesis. Crucially though, this

study did not manipulate prior exposure to the individual faces — rather, participants were assumed to have extensive real-world experience with certain celebrity faces over others. Therefore, this study cannot tell us anything about the question of whether averaging familiar individuals makes them truly disliked (i.e., reducing liking for the blend to a level that is below liking for the individual) or if the mere exposure effect is reduced. As we will discuss shortly, this prediction is a key difference between various theoretical models. Further, the observed dislike for "celebrity morphs" could also be explained by a variety of other factors. For instance, participants may simply not like when any image manipulation is applied to their local celebrities (i.e., "Don't mess with my hero!") or participants may dislike the blending of individuals on opposite sides of the social or societal spectrum (i.e., "Don't mix liberals and conservatives!", as with the case of the famous "Bushama" blend). Finally, it is also possible that the effects obtained in this study require massive experience with the individuals, over many years and exposures. Given that the Halberstadt et al. (2013) study relied on such "naturalistic" exposure, it simply cannot answer these questions, nor can it provide evidence for any mechanism or boundary conditions underlying any similar effects.

Our memory-based prediction also differs from several alternative accounts. The most intuitive alternative prediction is that the effects 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 and mere exposure (Monahan, Murphy, & Zajonc, 2000). The prediction of an additive effect from mere exposure and BiA manipulations follows from assumptions that these two effects involve separate mechanisms. As an example, if the attractiveness of a blend is driven by the elimination of "mutant-like" features, participants should prefer blends made from unfamiliar *and* familiar individuals.

Interestingly, other accounts make the complete opposite *mismatch* prediction. Here,

mere exposure and blending should combine negatively, making the morph of two familiar individuals especially unattractive and reducing the liking for the blend below the level of the contributing individuals. This prediction aligns with frameworks for ambiguity aversion, cognitive conflict, and prediction error, given that the morph falls in-between two established categories and represents a cognitive mismatch (Arnal & Giraud, 2012; Dreisbach & Fisher, 2015; Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005; Neta, Kelley, & Whalen, 2013). A similar prediction would also be made by the literature on the "uncanny valley effect," since the morph of two familiar categories represents a distortion to the features of the contributing individual (Kätsyri, Förger, Mäkäräinen, & Takala, 2015; Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012).

Again, note that these *additive* and *mismatch* predictions are different than our familiarity-based prediction from memory models — which expect blends of highly learned individuals to generate familiarity and preference values in-between actually exposed individuals and novel individuals (Jones & Jacoby, 2001; Kelley & Wixted, 2001).

Current Research

The current experiments offer the first systematic investigation of the idea that the attractiveness of facial blends varies as a relative function of their prior exposure. We expected that blends of familiar faces (which we experimentally manipulated), but *not* blends of *un*familiar faces, would be *less* attractive than their constituents (i.e., UiA effect).

First, we wanted to establish the BiA effect under standard conditions, where all the stimuli are initially unknown and exemplars are weakly learned. This also allowed us to examine the link between attractiveness and familiarity (Experiment 1).

Second, we wanted to directly test for the UiA effect under experimental conditions that simultaneously gauged both the mere exposure effect and the blending of highly learned

exemplars (Experiment 2).

Third, we wanted to probe the underlying mechanism for any observed effects, by delineating between different theoretical explanations (Experiment 3). Recall that the mismatch accounts propose that morphing two familiar individuals causes a conflict when perceiving the morph (which leads to an absolute "dip" in attractiveness for trained morphs). Our alternative familiarity account proposes that any reversals observed with the UiA effect are driven by a reduction in morph similarity, compared to its constituent individuals. These accounts can be distinguished across experiments, by morphing two familiar individuals (i.e., high conflict; Experiments 2 and 4) and morphing one familiar individual with one unfamiliar individual (i.e., no conflict, but low similarity; Experiment 3).

Finally, we wanted to examine what type of familiarity with the trained stimulus plays a role in these effects. For instance, to create a UiA effect for certain faces, does the familiarity have to be social in nature (e.g., learning names), or can it be purely perceptual (e.g., seeing the face as a background image during another task)? The memory models suggest that our predictions should be obtained with stimuli that have mere perceptual familiarity. After all, mere exposure effects and blending effects have also been obtained with non-social stimuli (e.g., random dot patterns, Chinese ideographs, etc.), and visual similarity is a powerful driver in the neural processing of faces (Natu & O'Toole, 2011). On the other hand, other research suggests that the processing of familiar faces in the social context may be unique, compared to faces that are novel or only perceptually familiar (Cloutier, Kelley, & Heatherton, 2011). Furthermore, perhaps the nature of the mismatch or conflict must be social? (as with the famous "Bushama" blend, which morphs a conservative George W. Bush with a liberal Barack Obama). We answered these questions by implementing a purely perceptual training paradigm in Experiment 4.

We used four experiments to address our main questions (described above). In

Experiment 1, participants simply rated familiarity and attractiveness for a large set of unfamiliar individuals and dual-person morphs that were generated using those individuals. We observed the standard BiA effect, and the increased attractiveness for morphs was mediated by their perceived familiarity. In Experiment 2, participants were "trained" on a subset of faces (i.e., either "set A" or "set B," but never both), using a free-recall task that required pairing names with individual faces. Thus, 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. After training, participants rated the attractiveness and familiarity of trained and untrained morphs and individuals. Here, we observed 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 address whether this UiA effect for trained morphs was driven by cognitive conflict (i.e., mismatch account) versus a relative reduction in similarity (i.e., 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). Finally, in Experiment 4, we used a perceptually based learning paradigm without names, to assess the relative importance of social and perceptual familiarity for the UiA effect. Our perceptual training task replicated the UiA effect for trained faces, suggesting it is driven by low-level visual familiarity cues (rather than any social information that is paired with the trained faces).

Experiment 1

In Experiment 1, we wanted to test whether our stimulus set generates a standard BiA effect (i.e., morphs rated as more attractive than individuals) using a design with minimal exemplar learning. We expected that when many individual exemplars are presented, without

strong learning of any 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 many exemplars, leading to limited item-specific memory, generates familiarity for a prototypical representation (Posner & Keele, 1968; Winkielman, Halberstadt, Fazendeiro, & Catty, 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, Pecher, Zeelenberg, Wai, & Winkielman, 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 (i.e., 56 individuals and 28 morphs) one-at-a-time, in a randomized order. Note that one feature of this standard design is that many morphs are statistically preceded by their constituting exemplars (making the morphs somewhat familiar). For each face, participants were asked to rate each image on attractiveness and familiarity, using 1 (*not at all attractive // familiar*) to 9 (*very attractive // familiar*) scales.

Results

Attractiveness and familiarity. As predicted, participants rated morphs as more attractive (M = 4.32, SD = 1.17) than individuals (M = 4.20, SD = 1.15), $t(150) = 5.14$, $p < .001$. This confirms that our stimulus set yields a traditional BiA effect in the standard paradigm, when only weak exemplar learning occurs. Consistently, we observed that the morphs were also rated as more familiar (M = 2.46, SD = 1.44) than the individuals (M = 2.36, SD = 1.37), $t(150) = 2.57$, $p = .01$ (see Figure 1.1a). It is also worth noting that the familiarity values are rather low, towards 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, 2014). Such a strategy is appropriate for repeated-measures designs to account for observations nested within participants, since they allow for model-based estimation of the average total, direct, and indirect mediation effects using hierarchical data structures (Bauer, Preacher, & Gil, 2006).

Here, 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. To conduct the multilevel mediation analyses for Experiment 1, 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 5,000 samples per estimate, after which quasi-Bayesian confidence intervals were calculated around the average total, direct, and causal mediation effects.

Figure 1.1b displays the mediation results. We observed clear evidence for mediation.

The total effect (b = .12, $CI_{95\%}$ = [.07, .17], p < .01) and average direct effect (b = .11, $CI_{95\%}$ = [.06, .15], p < .01) on attractiveness ratings were both significant. Target type was a significant predictor of familiarity (*a*-path: b = 0.09, SE = .04, t = 2.58, p = .01), and familiarity was a significant predictor of attractiveness (*b*-path: b = .16, SE = .04, t = 3.73, p < .001). When controlling for familiarity (*c'*-path), the original *t*-value estimate of target type on attractiveness (*c*-path: b = .12, SE = .02, t = 5.14, p < .001) was pushed to non-significance (b = .06, SE = .05, t = 1.24, *ns*), while familiarity was still significant (b = .11, SE = .04, t = 2.56, p = .01). And critically, the average causal mediation effect was also significant (b = .01, $CI_{95\%}$ = [.002, .02], p = .02), revealing familiarity as a mediator.

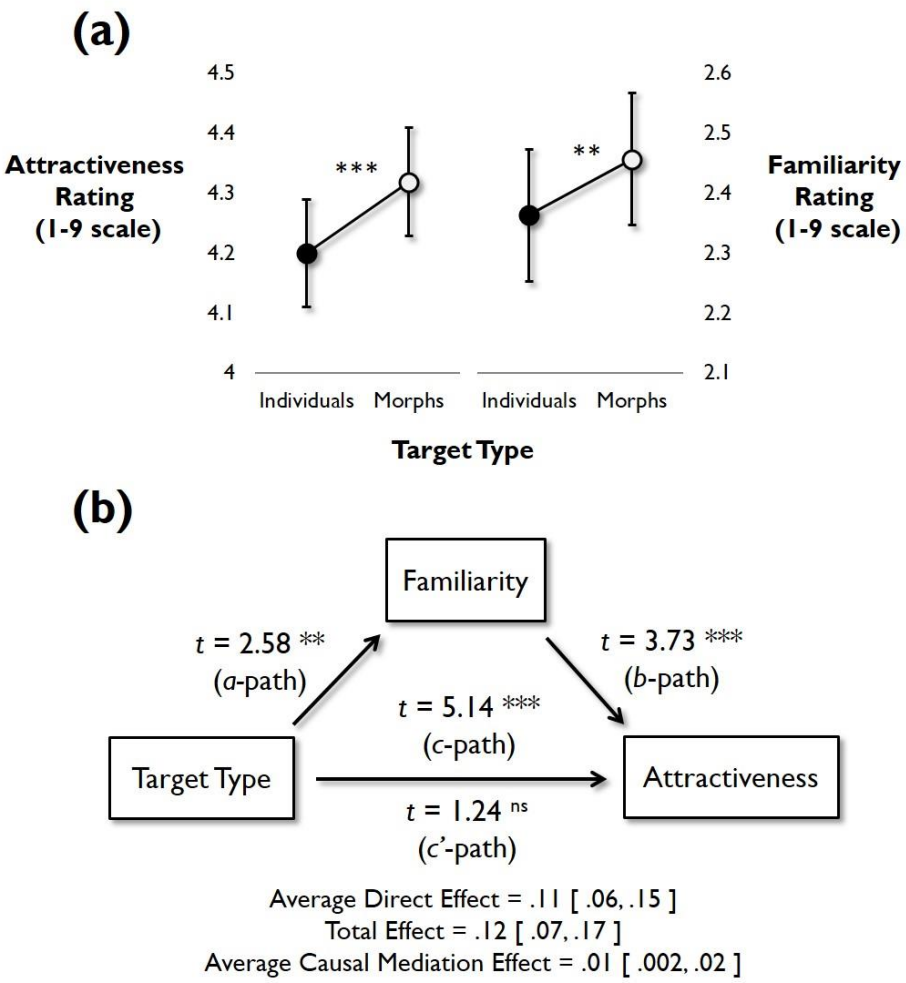


Figure 1.1: Experiment 1 results for attractiveness and familiarity (panel *a*), along with multilevel mediation (panel *b*). We demonstrated that when weak exemplar learning occurs, our stimulus set generates a standard beauty-in-averageness (BiA) effect, where morphs were rated as more attractive than individuals. Morphs are also rated as more familiar than individuals, and this familiarity mediates the relationship between target type (i.e., individuals vs. morphs) and attractiveness ratings (panel *b*).

Experiment 2

Experiment 1 demonstrated that with weak exemplar learning, morphs are judged as more attractive and familiar than individuals (i.e., a traditional BiA effect). These results fit with the memory literature, where in the absence of any strong individual memory traces, participants rely only on the global familiarity of the faces.

We designed Experiment 2 to address our main question. Namely, we wanted to test the idea that a novel 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 these individual exemplars are strengthened via learning, participants should now have higher familiarity for the individuals. Therefore, when blends of highly familiar individuals are presented, the morph 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 2, we "trained" participants on a subset of faces (set A vs. set B), using a free-recall task that required pairing names with individuals. Thus, 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. We observed a UiA effect for trained stimuli — where morphs of trained individuals were rated as *less* attractive than the trained individuals themselves.

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.

Materials. The stimuli were the same as the 56 individuals and 28 morphs used in Experiment 1. Using these stimuli, we wanted to create an experimental situation where we had

individuals and morphs that were both trained and fully untrained (according to each participant). To do this, we created two different sets of images (set A and set B) that contained half the total number of individual faces (i.e., 28 in each set) and half the total number of morph faces (i.e., 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 average attractiveness ratings for 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) (see Figure 1.2a). Importantly, for this study, the morphs were never composed "cross-set" (i.e., there were never 50/50 A-B morphs) (see Figure 1.2b). Note that the images included in sets A and B were the same in Experiments 2 and 4, but this was modified for empirical and theoretical purposes in Experiment 3.

Critically, the advantage of this setup is that each participant rated individuals and morphs after training, according to four important conditions (i.e., untrained individuals, trained individuals, morphs of untrained individuals, and morphs of trained individuals). As an example, if a participant was assigned to study set A individuals (not set B) in Experiment 2, they would be exposed to set A individuals during training, after which they would give ratings to set A individuals (trained individuals), set B individuals (untrained individuals), A-A morphs (morphs of trained individuals), and B-B morphs (morphs of untrained individuals) (see Figures 1.2 and 1.3).

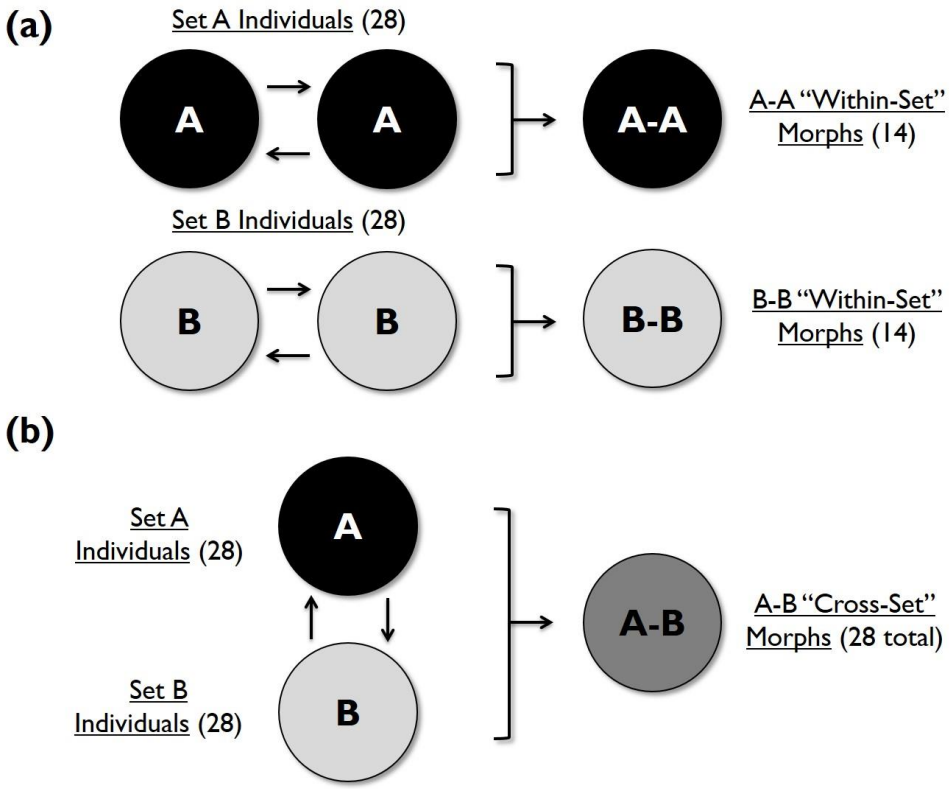


Figure 1.2: Structure of individual and morph setup in Experiment 2 (top panel *a*), Experiment 3 (bottom panel *b*), and Experiment 4 (top panel *a*). These experiments used 56 individuals, which were split into two different sets (set A vs. set B), each containing 28 individuals. Experiments 2 and 4 used "within-set morphs," where 28 dual-person morphs were created by averaging two set A individuals or by averaging two set B individuals (i.e., there were only A-A morphs and B-B morphs, but never A-B morphs). Experiment 3 used "cross-set morphs," where 28 dual-person morphs were created by averaging one set A individual with one set B individual (i.e., there were only A-B morphs, but never A-A or B-B morphs).

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 (never both), before progressing through seven rounds of a free-recall task.

Figure 1.3a 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" phase, where they were told to recall the name that was paired with each training face (i.e., they were presented with a response box on screen, where they would type the name), during which feedback was given. During test phases, RTs were measured from stimulus onset to the final submission of the participants' 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 seven 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 seven training rounds, to keep the level of exposure consistent).

After participants finished the seven rounds of training, they rated each stimulus (i.e., 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 — (1) morph attractiveness, (2) individual attractiveness, (3) morph familiarity, and (4) individual familiarity. On attractiveness ratings, participants were asked "How 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. Within each of the four different rating blocks, stimulus presentation was completely randomized.

Note that in each rating block, the morph ratings always came first. This was done to ensure that ratings of morphs were not influenced by exposure to untrained individuals, since we predicted that any UiA effect would occur after exposure to trained individuals. Also keep in

mind that for this design, the memory literature predicts an elimination of the BiA effect for morphs of unfamiliar individuals, since participants never see the exemplars before rating morph attractiveness.

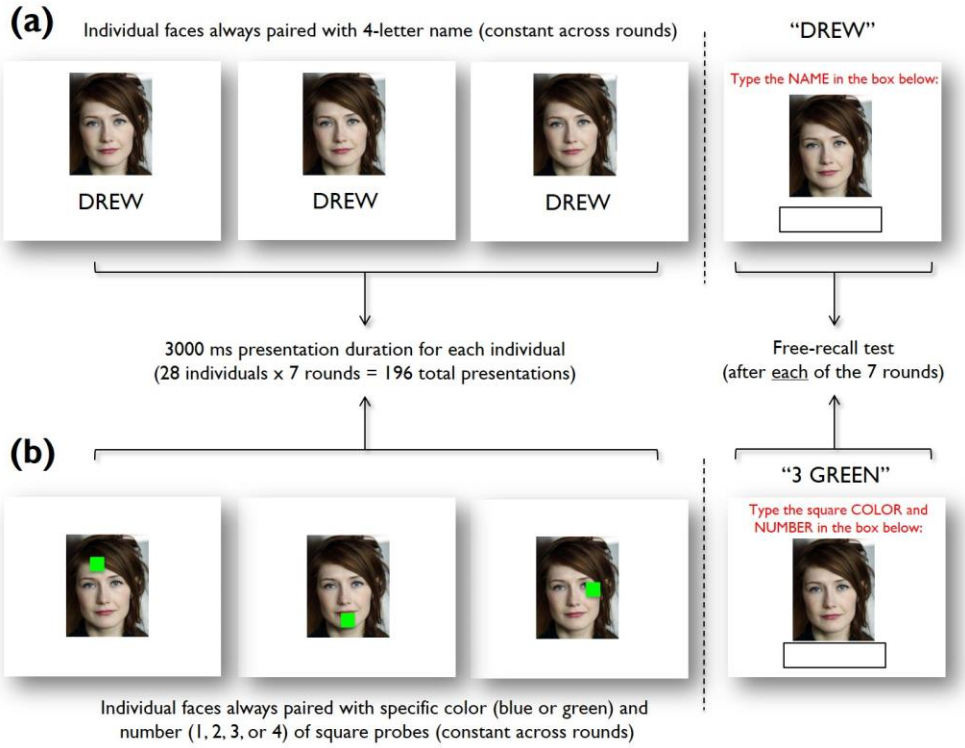


Figure 1.3: Design of the training task for Experiments 2 (top panel *a*), 3 (top panel *a*), and 4 (bottom panel *b*). Experiments 2-3 used a name-learning task, where all 28 individuals in the participants' respective training condition (set A vs. set B) were paired with a four-letter name. Across seven 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 after each "study" round). Experiment 4 used a similar training task, but it was changed to remove the names, in order to create training that was perceptually based. Here, instead, participants were 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). Since the names in Experiments 2-3 stayed the same across all rounds, the square probe color/number assigned to each face was also constant across rounds in Experiment 4. All other timing/exposure parameters for Experiment 4 training were the same as Experiments 2-3.

Results

Analysis strategy. All repeated-measures analyses (including ratings, RTs, and accuracy) in Experiment 2 used mixed-effects modeling via maximum likelihood, since this

method offers numerous analytical advantages — including more effective handling of unbalanced data with missing observations, reliance on fewer assumptions regarding covariance structures, and increased parsimony and flexibility between models (Bagiella, Sloan, & Heitjan, 2000). All models were built using packages the *lme4* (Bates, Maechler, Bolker, & Walker, 2014) and *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2014) in R. To obtain *p*-value estimates for fixed-effects, we used Type III Satterthwaite approximations, which can sometimes result in decimal degrees of freedom, based on the number of observations (West, Welch, & Galecki, 2014).[1]

Training performance (name memory task). We assessed participants' accuracy and RT performance, when recalling names paired with individuals over the seven test rounds (according to whether they were randomly assigned to study set A or set B). We analyzed this using a Training Condition (2: set A, set B) x Testing Block (7) fixed-effects structure, on both accuracy and RTs. To normalize the RT distribution and reduce the impact of outliers, all incorrect RTs were excluded, and the remaining correct RTs were $\log_{10}$-transformed.

Figure 1.4 depicts the results for participants' training performance across rounds. In short, our training task was effective, since participants became progressively more quick and accurate at the free-recall task. This task also standardized the level of exposure for each of the different individual face sets (set A vs. set B), depending on the participants' training condition (see footnote for detailed results of these analyses).[2]

---

[1] Final mixed-effects models were selected based on top-down model building. Maximal random intercept and maximal random slope models were created (using all by-participant effects). Next, the two model fits were tested against one another via χ^2 likelihood-ratio tests (i.e., nested model comparison). If there was *no* significant difference in model fit, the model with fewer random effects 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 effects 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.

[2] In Experiment 2, we observed a main effect of Testing Block on $\log_{10}$-RTs, $F(6, 129.12) = 64.98$, $p < .001$ (Figure 1.4, top panels), 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, $F(1, 72.02) = .02$, *ns*, or a Training Condition x Testing Block

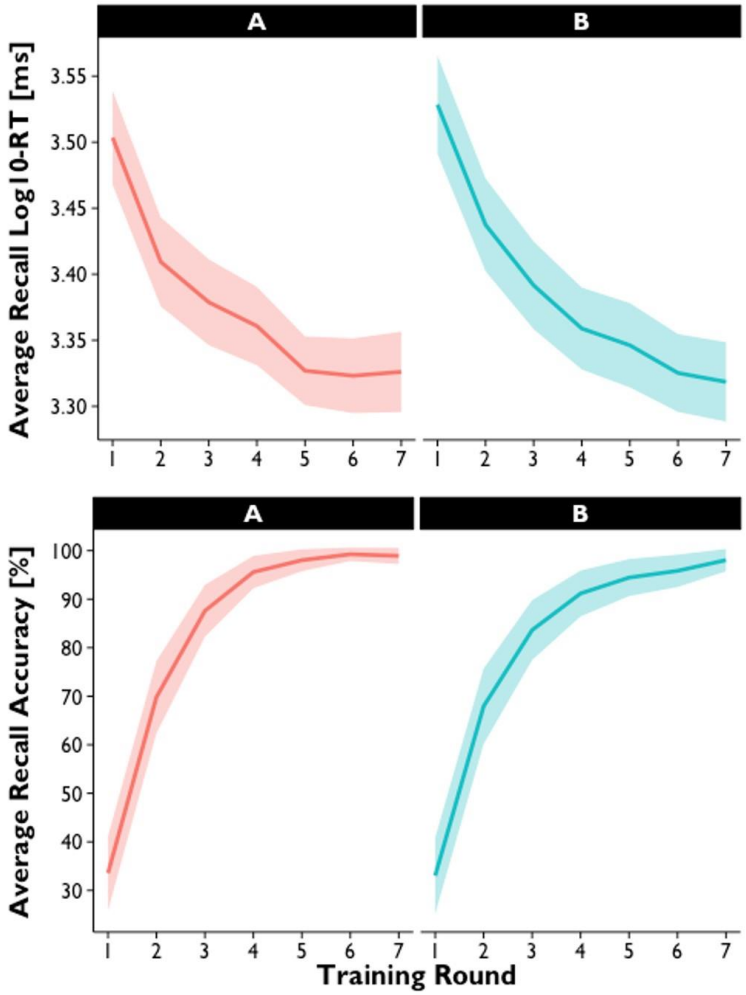


Figure 1.4: Training performance for participants in Experiment 2. Log-10 RTs (top panels) and accuracy performance (bottom panels) during free-recall test phases are shown across all seven training rounds. Training condition (set A vs. set B) is displayed with plot headers (set A = left column; set B = right column) and by line color (set A = red; set B = blue). Confidence bands show standard errors of the mean.

Attractiveness ratings. To answer our main question, we tested how participants'

---

interaction, $F(6, 129.12) = .87$, *ns*. On recall accuracy (Figure 1.4, bottom panels), we again found a main effect of Testing Block, $F(6, 124.51) = 352.27$, $p < .001$, 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, $F(1, 72.01) = 1.22$, *ns*, nor any evidence for a Training Condition x Testing Block interaction, $F(6, 124.51) = .40$, *ns*.

attractiveness ratings depended on training and morphing. We analyzed attractiveness using a mixed-effects model with a Training Type (2: trained, untrained) x Target Type (2: individual, morph) fixed-effects structure.

There was strong evidence for a Training Type x Target Type interaction, $F(1, 5{,}995.00) = 25.14$, $p < .001$. Follow-up tests on this interaction demonstrated that untrained morphs were numerically judged as *more* attractive than untrained individuals, although this effect was not significant, $b < 0.10$, $SE = .08$, $CI_{95\%} = [-.11, .21]$, $t = .58$, *ns*. 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, $b = -0.50$, $SE = .08$, $CI_{95\%} = [-.63, -.31]$, $t = -5.84$, $p < .001$. 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, $b = 0.30$, $SE = .09$, $CI_{95\%} = [.10, .44]$, $t = 3.09$, $p < .01$ (see Figure 1.5a). 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. A main effect of Training Type, $F(1, 90.60) = 94.79$, $p < .001$, reflecting overall higher ratings for trained targets compared to untrained targets. A main effect of Target Type, $F(1, 73.80) = 11.69$, $p = .001$, also demonstrated overall higher ratings for individuals compared to morphs.

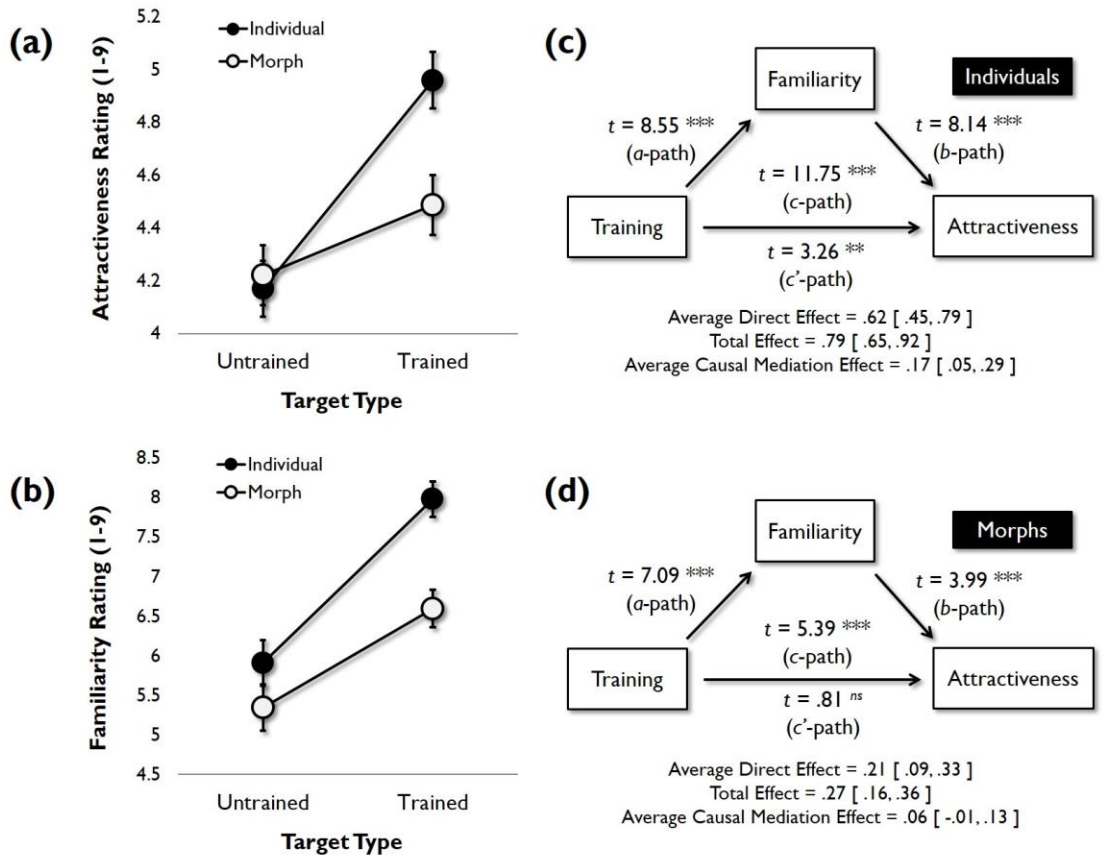


Figure 1.5: Attractiveness ratings (top-left panel *a*), familiarity ratings (bottom-left panel *b*), and multilevel mediation results for individual faces (top-right panel *c*) and morphed faces (bottom-right panel *d*) in Experiment 2. (*a*) We observed an ugliness-in-averageness (UiA) effect after training, such that trained morphs were judged as *less* attractive than trained individuals. (*b*) For familiarity, all effects were significant, but the interaction was driven by the fact that there was a greater increase in familiarity for individuals after training, compared to morphs. (*c*) Multilevel mediation demonstrated that for individual faces, the relationship between exposure training and attractiveness ratings was significantly mediated by familiarity. (*d*) The parallel average causal mediation effect for morphed faces was not quite significant, but trending in the same direction. Error bars = ± 1 standard error of the mean.

Familiarity ratings. To verify that our memory task induced selective familiarity for only trained faces, we analyzed familiarity ratings in the same way as attractiveness ratings, using a mixed-effects model with Training Type (2: trained, untrained) x Target Type (2: individual, morph) fixed-effects structure.

Similar to attractiveness, we observed strong evidence for all effects. First, a main effect

of Training Type, $F(1, 73.00) = 83.04$, $p < .001$, demonstrated that trained targets were judged as more familiar than untrained targets. Second, a main effect of Target Type, $F(1, 73.00) = 19.80$, $p < .001$, showed that individuals were judged as more familiar than morphs. Finally, we detected a Training Type x Target Type interaction, $F(1, 73.00) = 14.25$, $p < .001$. This interaction revealed a greater difference between trained and untrained individuals, $b = 2.10$, $SE = .24$, $CI_{95\%} = [1.59, 2.55]$, $t = 8.55$, $p < .001$, compared to trained and untrained morphs, $b = 1.30$, $SE = .18$, $CI_{95\%} = [.90, 1.60]$, $t = 7.09$, $p < .001$. Consequently, trained individuals were judged to be more familiar than trained morphs, $b = 1.40$, $SE = .22$, $CI_{95\%} = [.95, 1.82]$, $t = 6.32$, $p < .001$. Untrained individuals were seen as a bit more familiar than untrained morphs, $b = 0.60$, $SE = .27$, $CI_{95\%} = [.04, 1.10]$, $t = 2.12$, $p = .04$, but this difference was smaller than the difference between trained individuals and trained morphs (see Figure 1.5b).

It is also worth noting that the familiarity ratings for Experiment 2 were overall greater than those from Experiment 1 (i.e., Experiment 1 familiarity ratings fell mostly between 2 and 3, whereas Experiment 2 familiarity ratings were mostly between 5 and 9). Since strong learning only occurred in Experiment 2 (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 2, 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 2, since they did see those individuals once when rating attractiveness in earlier blocks. Finally, in Experiment 2, we also observed that untrained individuals were rated as slightly more familiar than untrained morphs. This is likely due to the fact that learning on the individual exemplars 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.

Multilevel mediation. We used the same multilevel mediation procedure as Experiment 1, but with some small changes (given the new structure of the data in Experiment 2). Two separate datasets for each target type (individuals vs. morphs) were created, where our main predictor was training type (coded as either 0 [untrained] or 1 [trained]). Our main DV was attractiveness ratings, and our mediator was familiarity ratings. 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 5,000 samples per estimate, after which quasi-Bayesian confidence intervals were calculated around the average total, direct, and causal mediation effects.

Figures 1.5c and 1.5d show the mediation results. We observed clear evidence of mediation for individual faces. The total effect (b = .79, $CI_{95\%}$ = [.65, .92], p < .01) and average direct effect (b = .62, $CI_{95\%}$ = [.45, .79], p < .01) on attractiveness ratings were both significant. Not surprisingly, training was a significant predictor of familiarity (a-path: b = 2.07, SE = .24, t = 8.55, p < .001), and familiarity was a significant predictor of attractiveness (b-path: b = .20, SE = .02, t = 8.14, p < .001). When controlling for familiarity (c'-path), the original t-value estimate of training on attractiveness (c-path: b = .79, SE = .07, t = 11.75, p < .001) was reduced but still significant (b = .86, SE = .26, t = 3.26, p = .001), while familiarity was also significant (b = .10, SE = .03, t = 3.42, p < .001). Moreover, the average causal mediation effect was also significant (b = .17, $CI_{95\%}$ = [.05, .29], p < .01), suggesting familiarity as a mediator.

When following the same process for morphs, the evidence for mediation was in the same direction, albeit not as strong. Specifically, the total effect (b = .27, $CI_{95\%}$ = [.16, .36], p < .01) and average direct effect (b = .21, $CI_{95\%}$ = [.09, .33], p < .01) on attractiveness ratings were still both significant. However, the average causal mediation effect was not quite significant (b =

.06, $CI_{95\%}$ = [-.01, .13], p = .12). When controlling for familiarity (*c'*-path), the original *t*-value estimate of training on attractiveness (*c*-path: b = .27, SE = .05, t = 5.39, p < .001) was reduced to non-significance (*c'*-path: b = .13, SE = .17, t = .81, *ns*). Familiarity ratings were also not a significant predictor, only in this *c'*-path model (b = .04, SE = .03, t = .54, p = .17). Note, however, that the *a*-path model (b = 1.25, SE = .18, t = 7.09, p < .001) and *b*-path model (b = .09, SE = .02, t = 3.99, p < .001) were both significant.

Summary. We found clear evidence that participants' familiarity ratings mediated the relationship between training and attractiveness on individual faces. For morphed faces, familiarity still impacted the relationship between training and attractiveness (since familiarity ratings significantly predicted attractiveness in the *b*-path model, and it reduced other *c'*-path model coefficient estimates) — but the average causal mediation effect was not as strong (p_{morph} = .12 vs. $p_{individual}$ < .01).

Experiment 3

Experiment 2 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. This finding offers a major qualification to the classic beauty-in-averageness (BiA) effect. 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 brief periods of exposure, demonstrating that the UiA effect does not require immense expertise to emerge. Theoretically, the results are consistent with predictions derived from modern memory frameworks, which emphasize the role of exemplar learning in familiarity responses (and in turn, facial attractiveness).

With Experiment 3, we wanted to 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 very attractive. This prediction seems most intuitive, under the assumption that these two manipulations enhance liking via separate and independent mechanisms. However, the results from Experiment 2 offer clear evidence against this idea, since trained morphs were judged as *less* attractive than trained individuals (i.e., a UiA effect).

This leaves two other possibilities. One *mismatch* account suggests that encountering a blend of two familiar individuals causes a cognitive conflict, perhaps not unlike conflict triggered by bi-stable figures (Kornmeier & Bach, 2012; Topolinski, Earle, & Reber, 2015). The negative affect generated from this conflict is then misattributed to subsequent ratings (causing the "dip" in attractiveness for trained morphs). However, if this was the case, we might expect that trained morphs would be judged as not only less attractive than trained individuals, but also less attractive than untrained morphs. But again, this is not what we observed in Experiment 2.

Instead, on the memory framework, the UiA effect is driven by a relative reduction in familiarity-based cues for trained morphs. More specifically, increased memory traces for individual exemplars after training leads to greater familiarity for trained individuals than trained morphs, since the former are exact replicates of what was shown during training. This idea is supported by the multilevel mediation results from Experiment 2, demonstrating that familiarity significantly predicted attractiveness ratings for both individuals and morphs. Moreover, the memory framework is reinforced by the fact that blends of highly learned individuals generated familiarity and preference values in-between actually exposed individuals and novel individuals (Jones & Jacoby, 2001; Kelley & Wixted, 2001).

We used Experiment 3 to further distinguish between these competing mismatch (conflict-based) and familiarity (memory-based) hypotheses, with a simple change to our Experiment 2 design. Recall that in Experiment 2, two different sets of images (set A vs. set B)

contained 28 individuals and 14 morphs each. Crucially though, all morphs were 100% "within-set," meaning that morphs would either be 50/50 morphs of two set A individuals or 50/50 morphs of two set B individuals (i.e., there were never 50/50 "cross-set" A-B morphs; see Figure 1.2a). However, in Experiment 3, we created new versions of "set A" and "set B" (once again based on attractiveness ratings from a previous study; Halberstadt et al., 2013) that instead used *cross*-set morphs (i.e., A-B morphs; see Figure 1.2b). Here, while both sets A and B each still contained 28 individuals and 14 morphs each, they were reorganized so that the two individuals composing each morph were *always* in different sets (that is, one set A individual and one set B individual always comprised each morph). Consequently, in Experiment 3, all morphs were cross-set A-B morphs, rather than the within-set A-A and B-B morphs used in Experiment 2.

Critically, this arrangement directly pits the two remaining alternative explanations against each other:

> **H_1** = This assumes that the UiA effect for trained morphs is driven by *mismatch* — where the conflict of processing a morph of two familiar individuals leads to negative affect, which is then misattributed to lower attractiveness ratings for those morphs. If so, cross-set morphs should *not* be rated as less attractive than trained individuals. 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 removed (and any UiA effect should be eliminated).
>
> **H_2 =** This assumes that the UiA effect for trained morphs is driven by *familiarity* — where the increased exemplar memory leads to greater positivity, so trained individuals receive increased attractiveness ratings than morphs because they are exact replicates of previous exposure during training. If so, cross-set morphs should *still* be rated as less attractive than trained individuals. Since 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 University of California, San Diego undergraduates participated for course-credit, and all participants signed consent forms approved by the UCSD

HRPP.

Materials. To address our main questions for this study, we made only one change to the stimulus setup from Experiment 2. As mentioned above, we created new versions of "set A" and "set B" (once again based on attractiveness ratings from a previous study; Halberstadt et al., 2013) that instead used cross-set morphs (i.e., A-B morphs). More specifically, while both sets A and B each still contained 28 individuals and 14 morphs each, they were reorganized so that the two individuals composing each morph were *always* in different sets (that is, one set A individual and one set B individual always comprised each morph). Therefore, in Experiment 3, all morphs were cross-set A-B morphs, rather than the within-set A-A and B-B morphs used in Experiment 1 (see Figure 1.2) — that is, each morph was the results of blending the faces of a trained individual and an untrained individual.

Design and procedure. The task design and procedure used here was the same as Experiment 2, only using the new stimulus setup created for Experiment 3 (see Figures 1.2 and 1.3).

Results

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

Training performance (name memory task). Similar to Experiment 2, we examined participants' accuracy and RT performance over all seven testing blocks during training. This analysis was structured according to a Training Condition (2: set A, set B) x Test Block (7) fixed-effects design, on both accuracy and RTs. As before, all RTs were $\log_{10}$-transformed to reduce the impact of outliers, after excluding error trials.

Figure 1.6 displays the training results. As with Experiment 2, our training task was effective, since participants became progressively more quick and accurate at the free-recall task,

according to their training condition (see footnote for detailed results of these analyses).[3]

[3] With Experiment 3, we found the predicted main effect of Test Block on $\log_{10}$-RTs, $F(6, 886.11) = 439.87$, $p < .001$ (Figure 1.6, top panels), where participants answered progressively quicker over successive training rounds. Note that we also observed evidence for a main effect of Training Condition, $F(1, 264.29) = 22.31$, $p < .001$, where set A participants had quicker RTs than set B participants throughout the memory task. As with Experiment 2, we observed no evidence for a Training Condition x Test Block interaction, $F(6, 957.61) = 1.88$, *ns*. On accuracy (Figure 1.6, bottom panels), we observed the expected main effect of Test Block, $F(6, 305.73) = 513.58$, $p < .001$. This demonstrated that participants' accuracy performance was similar to that of Experiment 2, where they started at approximately 36% correct responses in block 1, but improved to about 98% by block 7 (with performance beginning to level out at block 5). Similar to Experiment 3 RTs, we observed some evidence for a main effect of Training Condition, $F(1, 255.01) = 10.33$, $p = .001$, such that set A participants ($M_{acc} = 84.71\%$, $SD_{acc} = 35.99\%$) performed better than set B participants ($M_{acc} = 82.27\%$, $SD_{acc} = 38.20\%$) throughout the entirety of the memory task. There was no evidence for a Training Condition x Test Block interaction, $F(6, 373.84) = 1.74$, *ns*.

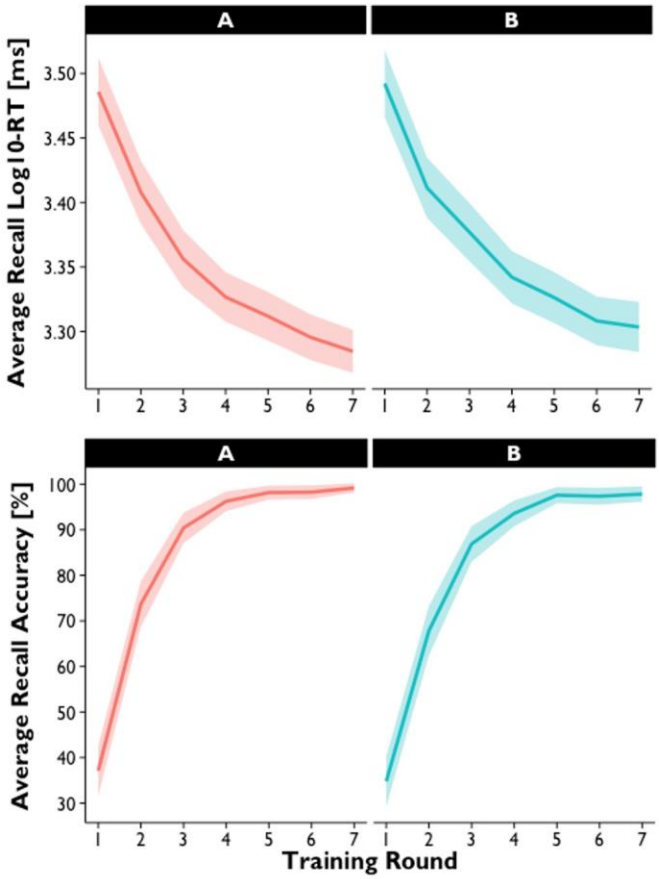


Figure 1.6: Training performance for participants in Experiment 3. Log-10 RTs (top panels) and accuracy performance (bottom panels) during free-recall test phases are shown across all seven training rounds. Training condition (set A vs. set B) is displayed with plot headers (set A = left column; set B = right column) and by line color (set A = red; set B = blue). Confidence bands show standard errors of the mean.

Attractiveness ratings. Our main questions centered on the link between training and attractiveness — specifically with the relative ratings given between trained individuals and cross-set morphs. Given that the cross-set morphs used here were neither 100% trained nor untrained (i.e., they were always composed of one trained and one untrained individual), we analyzed attractiveness differently than Experiment 2, using a mixed-effects model with Target Type (3: morph, trained individual, untrained individual) as the only fixed-effects factor.

We detected strong evidence for a main effect of Target Type, $F(2, 148.04) = 111.08$, $p <$

.001. Critically, a UiA effect still emerged, such that morphs were rated as *less* attractive than trained individuals, $b = .60$, $SE = .06$, $CI_{95\%} = [.44, .68]$, $t = 9.08$, $p < .001$. Interestingly, even though participants did not rate the morphs as more familiar than the untrained individuals (see below), they still rated the morphs as relatively more attractive, $b = .10$, $SE = .06$, $CI_{95\%} = [.01, .24]$, $t = 2.17$, $p = .03$. And as expected, we replicated the mere exposure effect, where trained individuals were judged as more attractive than untrained individuals, $b = .70$, $SE = .05$, $CI_{95\%} = [.59, .77]$, $t = 14.68$, $p < .001$ (see Figure 1.7a).

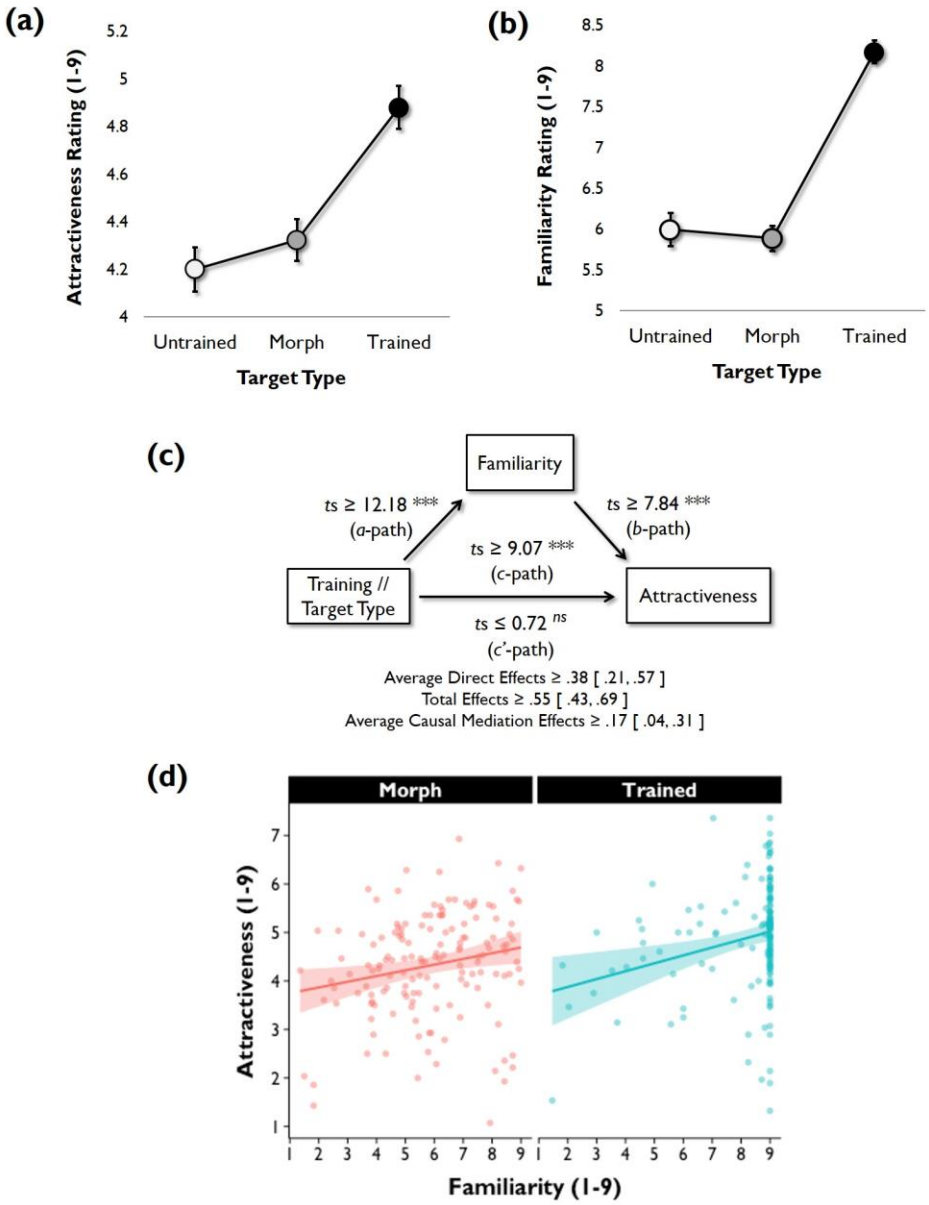


Figure 1.7: Attractiveness ratings (top-left panel *a*), familiarity ratings (top-right panel *b*), multilevel mediation results (middle panel *c*), and correlation analyses (bottom panel *d*) in Experiment 3. (*a*) We still observed an ugliness-in-averageness (UiA) effect after training using cross-set morphs (rather than the within-set morphs from Experiment 2), such that morphs were judged as *less* attractive than trained individuals. (*b*) Trained individuals were judged as more familiar than both untrained individuals and cross-set morphs. (*c*) 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. (*d*) Separate correlation analyses within morphs (left panel) and trained individuals (right panel) showed significant positive correlations between familiarity and attractiveness. Linear fits are shown in each plot, along with 95% confidence interval bands. Error bars = ± 1 standard error of the mean.

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

We observed a clear main effect of Target Type, $F(2, 149.00) = 133.83$, $p < .001$. Trained individuals were judged as more familiar than both untrained individuals, $b = 2.20$, $SE = .18$, $CI_{95\%} = [1.83, 2.53]$, $t = 12.18$, $p < .001$, and morphs, $b = 2.30$, $SE = .15$, $CI_{95\%} = [1.99, 2.58]$, $t = 15.48$, $p < .001$. Note that there was also no difference when comparing mean familiarity ratings between morphs and untrained individuals, $b = .1$, $SE = .18$, $CI_{95\%} = [-.44, .23]$, $t = .63$, *ns* (see Figure 1.7b), though as discussed below, familiarity still played a role in the attractiveness ratings of those targets.

Multilevel mediation. We built multilevel mediation models using the same procedure as Experiment 2, only with one important change. Since the cross-set morphs in Experiment 3 differed from the within-set morphs in Experiment 2 in that they were neither 100% trained nor untrained, this was collapsed into one three-level factor for Training Target Type (3: 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 to 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 1.7c displays a summary of the mediation results. Critically, we detected evidence

for mediation with both models. As in Experiment 2, training target type was a significant predictor of familiarity (*a*-path [M_1]: $b = 2.18$, $SE = 0.18$, $t = 12.18$, $p < .001$; *a*-path [M_2]: $b = 2.28$, $SE = 0.15$, $t = 15.48$, $p < .001$), and familiarity was a significant predictor of attractiveness (*b*-path [M_1]: $b = 0.18$, $SE = 0.02$, $t = 10.55$, $p < .001$; *b*-path [M_2]: $b = 0.16$, $SE = 0.02$, $t = 7.84$, $p < .001$). When controlling for familiarity, the original *t*-value estimate of training on attractiveness (*c*-path [M_1]: $b = 0.68$, $SE = 0.05$, $t = 14.61$, $p < .001$; *c*-path [M_2]: $b = 0.56$, $SE = 0.06$, $t = 9.07$, $p < .001$) was reduced to non-significance (*c'*-path [M_1]: $b = 0.16$, $SE = 0.22$, $t = 0.72$, *ns*; *c'*-path [M_2]: $b = 0.06$, $SE = 0.30$, $t = 0.21$, *ns*).

We also formally tested the total, indirect, and average causal mediation effects in both models. For M_1 (comparing trained individuals vs. untrained individuals), the total effect ($b = .68$, $CI_{95\%} = [.58, .78]$, $p < .01$), average direct effect ($b = .50$, $CI_{95\%} = [.37, .62]$, $p < .01$), and average causal mediation effect ($b = .18$, $CI_{95\%} = [.09, .28]$, $p < .01$) were all highly significant. M_2 (comparing trained individuals vs. morphs) showed similar results, with a significant total effect ($b = .55$, $CI_{95\%} = [.43, .69]$, $p < .01$), average direct effect ($b = .38$, $CI_{95\%} = [.21, .57]$, $p < .01$), and average causal mediation effect ($b = .17$, $CI_{95\%} = [.04, .31]$, $p = .01$).

In short, multilevel mediation demonstrated that familiarity mediated the relationship between training and attractiveness (both when specifically comparing trained individuals to untrained individuals and 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 1.7d 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, $r(149) = .24$, $CI_{95\%} = [.09, .39]$, $p < .01$. Importantly, this demonstrated that not only did familiarity significantly impact 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).

Summary. Overall, both the attractiveness and mediation results support the familiarity-based predictions made by memory models (H_2 described previously), where 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 — and the "dip" in attractiveness ratings for trained morphs is actually due to the relative reduction of those familiarity cues (i.e., trained individuals feel more familiar than trained morphs, since they are "pure" replicates of what was shown during the memory task).

Experiment 4

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 2 revealed that brief periods of training generate a mere exposure effect for those training individuals. Critically, this training also elicits a UiA effect, where trained morphs are judged as *less* attractive than trained individuals. 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).

In Experiment 4, we looked to build on the previous two experiments by investigating a different type of training. More specifically, the face-name memory task from Experiments 2 and

3 contained some social element, since participants were tasked with remembering identities for the trained individuals. But is this social element *necessary* to promote this increased familiarity, which in turn leads to reversals in attractiveness ratings between trained individuals and morphs? In other words, with a non-social version of the training task (where participants are not required to recall identity information, but endure similar levels of exposure), will we observe similar effects on attractiveness? The differences between perceptual and conceptual modes of familiarity have long been studied alongside memory-based processes (e.g., false recognition; Fazendeiro, Winkielman, Luo, & Lorah, 2005), and these distinctions are interesting for several reasons. One is the issue of process specificity, given that person-based knowledge for familiar faces can recruit distinct neural regions, as compared to those faces that are novel or only perceptually familiar (Cloutier, Kelley, & Heatherton, 2011). More critically, according to memory frameworks, the mechanisms for eliciting the UiA effect involve basic and general familiarity (as would be the case with low-level visual cues; Natu & O'Toole, 2011), rather than a conflict between social identities or some other aspect of social knowledge. Thus, the UiA should occur even if learning is kept only to its "pure" perceptual aspects.

We addressed this in Experiment 4 by changing the training to a perceptual detection and memory task (instead of one that focuses on face-name pairs). With this new non-social version of the training, participants were exposed to the same images (individual faces, in either set A or set B) over similar durations (seven blocks of "study" and "test" phases), but they instead had to detect and recall blue and green square probes that randomly appeared on each image. Note that with this task, participants' exposure to each of the images during training is held constant, but we changed the type of information that was required for recall — where Experiments 2 and 3 focused on names (social), while Experiment 4 focused on probes (non-social).

In sum, the predictions for Experiment 4 were as follows. If the UiA effect requires an element of *social* familiarity for trained individuals, then these effects should dissipate in

Experiment 4 (since the training task would not require learning or pairing any social information with those trained faces). If the UiA effect instead only requires *perceptual* familiarity for trained individuals, we should observe similar effects on attractiveness in Experiment 4 (since participants are still receiving the same amount of exposure to each of those faces during training, compared to Experiments 2 and 3).

Method

Participants. One hundred twenty-eight UCSD undergraduates participated for course-credit, and all participants signed consent forms approved by the UCSD HRPP.

Materials. All stimuli and materials were the same as Experiment 2. Note that we used within-set morphs in Experiment 4 (A-A and B-B morphs, as with Experiment 2), rather than the cross-set morphs used in Experiment 3 (A-B morphs) (see Figure 1.2).

Design and procedure. Our main changes focused on the type of memory task we used. As mentioned, we wanted to produce a version of the memory task that removed any social aspects (as there would be with face-name pairs) and focused on perceptual familiarity (using the face stimuli only as background images that participants would be trained on).

Figure 1.3b shows the main revisions to the design of the training task for Experiment 4. This training task had a similar structure to that of Experiments 2 and 3 — where participants would progress through seven rounds of the free-recall task on the 28 individuals in their randomly assigned training set (A or B). However, the type of recall they performed at the test phase during each round was different. Specifically, 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." Therefore, with this version of the memory task, no names were presented with the faces.

Critically, the "images" were the same individual face stimuli used in Experiments 2 and

3. Here, a face was assigned to a constant color (either blue or green) and constant number (between 1 and 4) of square probes (note that this color-number assignment did *not* change across successive rounds of training — similar to the names used in Experiments 2 and 3). During each study phase presentation (3000 ms for each image), these 200 ms square probes would then appear at random intervals, and participants were tasked with remembering the color and number of squares that appeared on each face (see Figure 1.3b).

All attractiveness and familiarity ratings that followed the memory task were the same as Experiments 2 and 3.

Results

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

Training performance (perceptual memory task). As was the case with Experiments 2 and 3, we aimed to gauge participants' accuracy and RT performance over all seven testing blocks during training. We structured this analysis according to a Training Condition (2: set A, set B) x Test Block (7) fixed-effects design, on both accuracy and RTs. Similar to 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.

Figure 1.8 summarizes the training results. Similar to Experiments 2 and 3, our training task was effective, since participants became progressively more quick and accurate over successive training rounds, according to their training condition. Note that there were some less theoretically important effects observed between color and number training performance metrics (see footnote for detailed results of these analyses).[4]

[4] For Experiment 4, we detected main effects of Test Block on both color RTs, $F(6, 126.00) = 67.58$, $p < .001$, and number RTs, $F(6, 126.00) = 105.82$, $p < .001$ (Figure 1.8, top panels). RTs on both color and number decreased over the course of the memory task to block 7. We also saw some evidence for a main

effect of Training Condition on color RTs, $F(1, 126.00) = 4.96$, $p = .03$, and this main effect was marginal for number RTs, $F(1, 126.00) = 3.52$, $p = .06$. While set B participants were more accurate overall, these RT main effects revealed that they took longer to recall both the color and number of square probes, compared to set A participants on color and number. While we did not observe any Training Condition x Test Block interaction for color RTs, $F(6, 126.00) = .29$, *ns*, we did for number RTs, $F(6, 126.00) = 2.70$, $p = .02$. This interaction showed that set A participants had a greater rate of RT improvement from blocks 1 to 7, compared to set B participants. On accuracy (Figure 1.8, bottom panels), the results mirrored the Experiment 4 RT findings. We found a main effect of Test Block for both color, $F(6, 756.00) = 159.43$, $p < .001$, and number, $F(6, 756.00) = 298.65$, $p < .001$, which both showed that overall accuracy improved throughout the memory task. Participants started around chance level in block 1 for both color (53.01%) and number (27.93%), which substantially improved by block 7 (81.72% and 75.00%, respectively), with performance starting to level out around block 5. Similar to RTs, we also observed some evidence for a main effect of Training Condition, on both color, $F(1, 126.00) = 6.04$, $p = .01$, and number, $F(1, 126.00) = 6.69$, $p = .01$. This showed that set B participants were more accurate in recalling the color (M_{acc} = 71.75%, SD_{acc} = 45.02%) and number (M_{acc} = 56.33%, SD_{acc} = 49.60%) of square probes, compared to set A participants on color (M_{acc} = 67.24%, SD_{acc} = 46.94%) and number (M_{acc} = 49.72%, SD_{acc} = 50.00%). While we did not observe any evidence for a Training Condition x Test Block interaction on color, $F(6, 756.00) = .18$, *ns*,, we did see an interaction for number, $F(6, 756.00) = 3.39$, $p < .01$. Post-hoc breakdowns of this interaction revealed that while both set A and B participants started at similar levels of accuracy, set B participants improved more quickly and ended block 7 with greater accuracy performance.

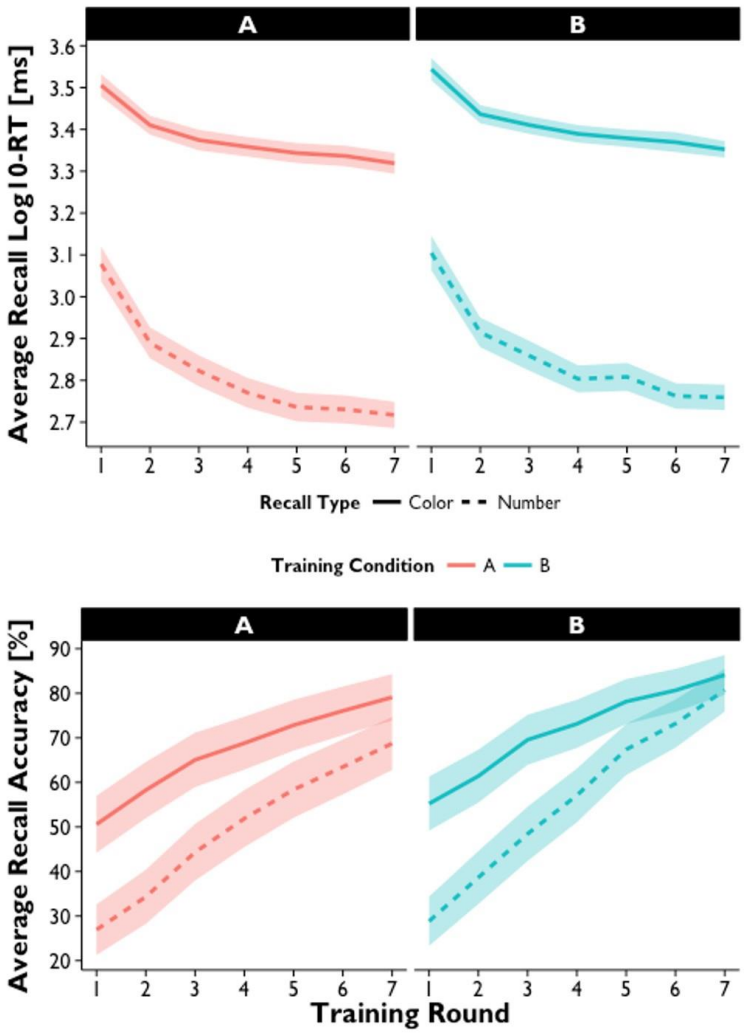


Figure 1.8: Training performance for participants in Experiment 4. Log-10 RTs (top panels) and accuracy performance (bottom panels) during free-recall test phases are shown across all seven training rounds. Training condition (set A vs. set B) is displayed with plot headers (set A = left column; set B = right column) and by line color (set A = red; set B = blue). The type of information for the free-recall (color vs. number of square probes) is shown by line type (color = solid line; number = dotted line). Confidence bands show standard errors of the mean.

Attractiveness ratings. To gauge the effects of training and morphing on attractiveness, we analyzed participants' attractiveness ratings using a mixed-effects model with a Training Type (2: trained, untrained) x Target Type (2: individual, morph) fixed-effects structure.

Figure 1.9a displays the attractiveness results. Most importantly, we found a Training

Type x Target Type interaction, $F(1, 10{,}370.80) = 39.54$, $p < .001$. Follow-up tests on this interaction revealed a similar UiA effect as Experiment 2, with trained morphs judged as *less* attractive than trained individuals, $b = 0.50$, $SE = .08$, $CI_{95\%} = [.33, .65]$, $t = 6.05$, $p < .001$. Untrained morphs were numerically rated as more attractive than untrained individuals, but not significantly so, $b < 0.10$, $SE = .08$, $CI_{95\%} = [-.15, .17]$, $t = .15$, *ns*. Also, similar to Experiment 2, trained morphs were still judged as more attractive when compared to untrained morphs, $b = 0.30$, $SE = .07$, $CI_{95\%} = [.14, .41]$, $t = 3.98$, $p < .001$. We also observed a mere exposure effect, since trained individuals were judged more attractive than untrained individuals, $b = 0.80$, $SE = .05$, $CI_{95\%} = [.67, .87]$, $t = 15.19$, $p < .001$.

The main effects were also significant. The main effect of Training Type, $F(1, 151.8) = 132.67$, $p < .001$, demonstrated that trained targets were judged as more attractive overall, compared to untrained targets. The main effect of Target Type, $F(1, 127.40) = 11.48$, $p < .001$, showed that individuals were judged as more attractive overall, compared to morphs.

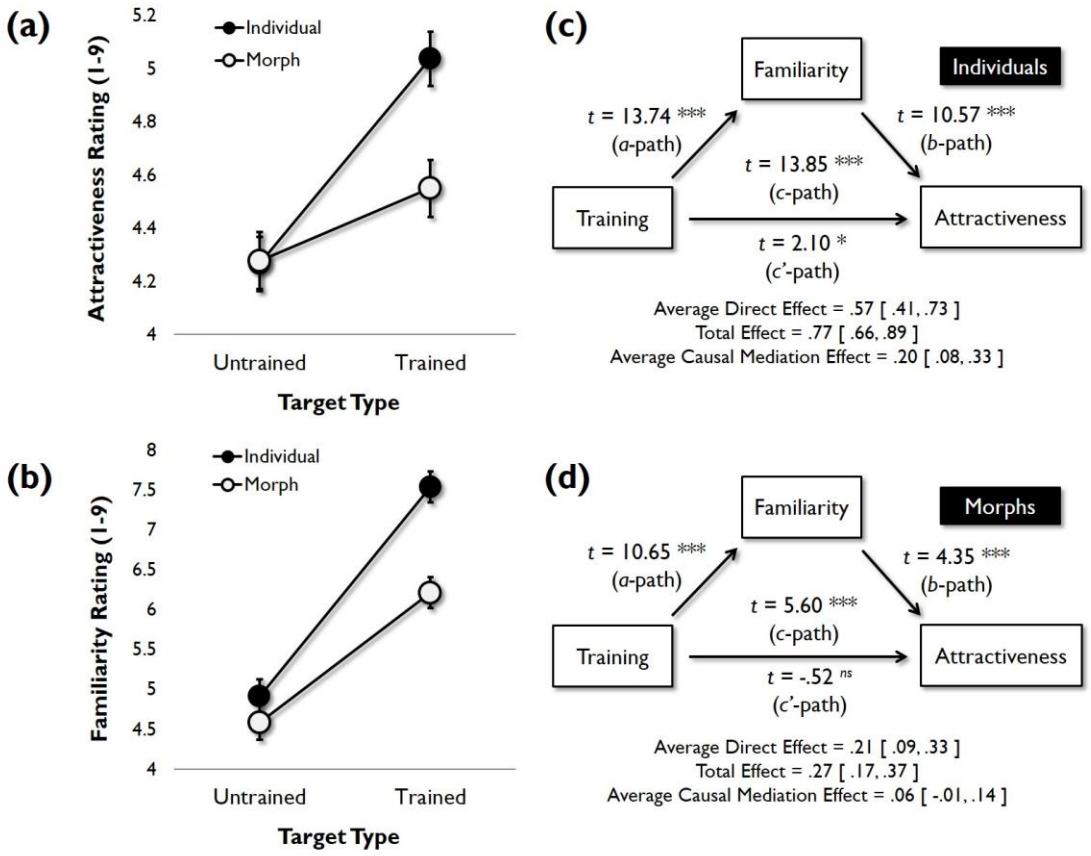


Figure 1.9: Attractiveness ratings (top-left panel *a*), familiarity ratings (bottom-left panel *b*), and multilevel mediation results for individual faces (top-right panel *c*) and morphed faces (bottom-right panel *d*) in Experiment 4. (*a*) We observed an ugliness-in-averageness (UiA) effect after training, such that trained morphs were judged as *less* attractive than trained individuals. (*b*) All familiarity effects were significant, but the interaction was driven by the fact that there was a greater increase in familiarity for individuals after training, compared to morphs. (*c*) Multilevel mediation demonstrated that the relationship between exposure training and attractiveness ratings was significantly mediated by familiarity for individual faces. (*d*) This parallel average causal mediation effect for morphed faces was marginally significant. Error bars = ± 1 standard error of the mean.

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

Figure 1.9b shows the familiarity results. All effects from the mixed-effects model on familiarity were significant. First, and most importantly, we observed strong evidence for a

Training Type x Target Type interaction, $F(1, 127.00) = 36.55$, $p < .001$. Follow-up testing on this interaction revealed the expected effect that trained individuals were rated the most familiar — compared to untrained individuals, $b = 2.60$, $SE = .19$, $CI_{95\%} = [2.25, 3.00]$, $t = 13.74$, $p < .001$, trained morphs, $b = 1.30$, $SE = .12$, $CI_{95\%} = [1.08, 1.57]$, $t = 10.64$, $p < .001$, and untrained morphs, $b = 3.00$, $SE = .21$, $CI_{95\%} = [2.54, 3.36]$, $t = 14.25$, $p < .001$. Critically though, this interaction yielded a similar pattern to Experiment 2, where the difference in familiarity ratings between trained individuals and trained morphs was more amplified, compared to the smaller difference between untrained individuals and untrained morphs, $b = .30$, $SE = .16$, $CI_{95\%} = [.01, .64]$, $t = 2.05$, $p = .04$.

Note that we also observed main effects for both Training Type, $F(1, 127.00) = 195.80$, $p < .001$, and Target Type, $F(1, 127.00) = 50.32$, $p < .001$. These main effects 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 2. Participants judged trained individuals as the most familiar. It is also worth noting that similar to Experiment 2, familiarity ratings in Experiment 4 fell mostly between 5 and 9, and untrained individuals were still judged as more familiar than untrained morphs (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. We used the same multilevel mediation strategy described in Experiment 2, to assess the relative influence of familiarity on the link between training and attractiveness ratings in Experiment 4. As a reminder, we created two separate datasets for each target type (individuals vs. morphs), where our main predictor was training type (coded as either 0 [untrained] or 1 [trained]); our main DV was attractiveness ratings; and our mediator was familiarity ratings.

Figures 1.9c and 1.9d show the mediation results. Critically, we observed strong evidence of mediation for individuals (see Figure 1.9c). The total effect was significant (b = .77, $CI_{95\%}$ = [.66, .89], p < .01), along with the average direct effect of training on attractiveness ratings (b = .57, $CI_{95\%}$ = [.41, .73], p < .01). As expected, training was a significant predictor of familiarity (a-path: b = 2.63, SE = .19, t = 13.74, p < .001), and once again, familiarity was a significant predictor of attractiveness (b-path: b = .19, SE = .02, t = 10.57, p < .001). When controlling for familiarity (c'-path), the original t-value estimate of training on attractiveness (c-path: b = .77, SE = .06, t = 13.85, p < .001) was reduced but still significant (b = .41, SE = .20, t = 2.10, p = .04), while familiarity was also significant (b = .06, SE = .03, t = 2.56, p = .01). Finally, the average causal mediation effect was significant (b = .20, $CI_{95\%}$ = [.08, .33], p < .01), suggesting familiarity ratings as a partial mediator.

When these same analyses were done for morphs, we still observe some evidence of mediation — albeit weaker (see Figure 1.9d). While there was a significant total effect (b = .27, $CI_{95\%}$ = [.17, .37], p < .01) and average direct effect (b = .21, $CI_{95\%}$ = [.09, .33], p < .01), the average causal mediation effect was marginal (b = .06, $CI_{95\%}$ = [-.01, .14], p = .09). Similar to Experiment 2, when controlling for familiarity (c'-path), the original t-value estimate of training on attractiveness (c-path: b = .27, SE = .05, t = 5.60, p < .001) was reduced to non-significance (c'-path: b = -.07, SE = .14, t = -.52, *ns*). Familiarity ratings were also not a significant predictor in this c'-path model (b = .01, SE = .02, t = .57, *ns*). Note, however, that the a-path model (b = 1.63, SE = .15, t = 10.65, p < .001) and b-path model (b = .08, SE = .02, t = 4.35, p < .001) were both significant.

In sum, the mediation results from Experiment 4 looked very similar to those from Experiment 2. We observed convincing evidence of mediation for individual targets, where familiarity mediated the relationship between training and attractiveness ratings. Moreover, for morphs, while familiarity was still connected to both training and attractiveness (attractiveness

was significantly predicted by familiarity, and familiarity reduced other *c'*-path model coefficient estimates), the omnibus causal mediation effect was in the same direction, but marginal (p_{morph} = .09 vs. $p_{individual} < .01$).

Discussion

The current research addressed the mechanisms underlying classic social preference effects and tested predictions generated by modern models of memory. The four experiments 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 (Experiment 1). Critically, they also document a novel *ugliness-in-averageness (UiA) effect*, where morphs of familiar individuals are judged as *less* attractive than contributing individuals (Experiments 2, 3, and 4). The experiments also demonstrate that the UiA effect is due to a relative reduction in familiarity for morphs of trained individuals. As a result, the attractiveness of highly familiar exemplars "trumps" the less familiar morphs. Moreover, consistent with predictions derived from memory theories, the UiA effect does not depend on a conflict between two well-known individuals, but only requires a decrease of familiarity of a single well-known exemplar (Experiment 3) and can be generated by purely perceptual and visual familiarity (Experiment 4). Generally, this research offers the first demonstration for 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). This not only highlights the importance of memory processes in understanding social judgments like attractiveness, but the current findings also represent a major qualification to the classic BiA effect, known since Galton (1879) and confirmed by a multitude studies using a variety of different paradigms, stimuli, and modalities (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 on modern theories of memory. Recall that on those theories, memories contain traces for individual exemplars (e.g., specific faces that are studied). These traces can be accompanied by a "gist" (prototype) representation for those exemplars (Deese, 1959; Posner & Keele, 1968; Whittlesea, 2002; Roediger & McDermott, 1995) or generate "gist" effects without assuming the existence of a unique prototype representation (Barsalou, 1990; Johansen & Palmeri, 2002; Love, 2013; Medin & Schaffer, 1978; Murphy, 2002). Also, the familiarity of a probe (target) is calculated from the similarity values of the probe with all traces in memory (or a relevant subset of traces). The strength of a memory trace determines the similarity between the probe and the memory trace — if the memory trace is weak (because only a few features of the item were stored correctly), the similarity between the probe and the memory trace will be lower than when the memory trace contains many correctly stored features. Thus, familiarity will be higher for strong items than for weak items (i.e., mere exposure effect). With weak learning of multiple items, the blend probe is more similar to all memory traces than any probe of individual faces, predicting the BiA effect. Crucially though, with strong learning, the probe of known individual faces is more similar to the relevant memory traces than the blend probe, predicting the UiA effect. Note that when participants rate morphs made from exemplars without any previous training at all, the memory literature predicts no BiA or UiA effects.

Now, let us move on to our main empirical findings. First, 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 also

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 (i.e., 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 2, 3, and 4 investigated the attractiveness for morphs of highly learned exemplars (i.e., when the individual exemplars have strong traces in memory) and morphs made out of completely unfamiliar exemplars. No BiA effect emerged for morphs made out of completely unfamiliar individuals, while the UiA effect emerged for trained morphs in all three experiments. Theoretically, this follows from our memory-based predictions, since 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 highly learned individuals generate 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 is still some lingering positive effects from partial familiarity (also see Gordan & Holyoak, 1983).

These robust confirmations of our memory-based account of familiarity can be contrasted with alternative theoretical predictions, as previously described in the Introduction. Recall that one prediction was that mere exposure and blending effects would be *additive* — where the positivity from both processes combine to create an even stronger BiA effect (i.e., morphs of highly learned exemplars appear extra attractive). This prediction is intuitive at first, especially under the assumption that the benefits of mere exposure and blending occur via independent mechanisms. Further, this additive logic worked in previous studies combining mere exposure and affective priming (Monahan, Murphy, & Zajonc, 2000). However, the findings from

Experiments 2, 3 and 4 clearly argue against this idea — since morphs of highly learned exemplars were deemed *less* attractive than their constituent individuals (i.e., UiA effect).

Another alternative account predicts the UiA effect, but for a different reason. The *mismatch* account assumes that the morphs of two familiar individuals appear especially unattractive because of negative affect generated by cognitive conflict between two established categories or "attractors" (Arnal & Giraud, 2012; Dreisbach & Fisher, 2015; Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005; Neta, Kelley, & Whalen, 2013). This clash would trigger negative affect, which would then generalize to the morph. This account clearly predicts that the dislike should be eliminated if the conflict is eliminated, by removing one conflicting strong category. In contrast to this prediction, our data supported the familiarity (memory-based) account in Experiment 3, since a UiA effect still emerged when using "cross-set" morphs (composed of one trained individual and one untrained individual, as opposed to the "within-set" morphs in Experiments 2 and 4). Keep in mind, however, that the current results do not challenge the overall validity of the mismatch (conflict-based) account as a mechanism for the generation of negative affect (Dreisbach & Fisher, 2015).

Note that we also found that purely perceptual familiarity in Experiment 4 (without any social information, as with the name-learning task in Experiments 2 and 3) is sufficient by itself to produce a UiA effect. This fits with decades of past research on the mere exposure effect using abstract stimuli (e.g., Chinese ideographs, letter strings, unknown melodies, etc.) and work in computational neuroscience showing the importance of visual cues in facial processing (Natu & O'Toole, 2011). Furthermore, it complements other recent research showing that familiarity can have both perceptual and conceptual components, but at different time stages (i.e., using EEG event-related potentials, perceptual effects emerge at 150–250 ms effects, while conceptual effects arise around 400 ms; Wang, Li, Gao, Xiao, & Guo, 2015). These timing differences would be especially valuable to investigate further with the UiA effect, since judgment and

physiology shift as stimulus processing progresses from perceptual (early) to conceptual (late) stages (Bradley & Lang, 2007). Neurally, this would also be useful in comparing low-level responses between blends of social (e.g., Halberstadt & Winkielman, 2014) and non-social stimuli (e.g., Winkielman et al., 2006).

The current research also observed very strong support for 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 bi-directional 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 judgement 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 the role of affective, motivational, and judgmental contexts in the UiA effect. Mechanistically, the familiarity-preference link could arise due to underlying changes in perceptual fluency (Winkielman, Schwarz, Fazendeiro, & Reber, 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 methodologically challenging, future studies may also attempt to separate the relative contribution of "pure" fluency and "pure" familiarity to the effects obtained in the current studies, though this question is not essential for our central point.

Going forward, the current work prompts many other interesting questions. As an example, the current studies did not investigate the role of valenced 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 (especially within the medial prefrontal cortex; Cloutier, Kelley, & Heatherton, 2011) and more generally between social and non-social stimuli (Gobbini & Haxby, 2007; Haxby, Hoffman, & Gobbini, 2000; Johnson, 2005). Clearly, involving dimensions with social complexity also needs to be considered (e.g., race or gender; Bernstein, Young, & Hugenberg, 2007; Malpass & Kravitz, 1969; Hugenberg & Bodenhausen, 2004). Finally, our results suggest that training modifies blending effects on attractiveness of visual stimuli, but it would also be interesting to gauge whether or not our UiA effect extends to different modalities (e.g., audition, via blended sounds; Bruckert et al., 2010) or works across modalities (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 considering familiarity, blends are not always beautiful.

Chapter 1 is, in full, under review for publication of the material. Carr, Evan W.; Pecher, Diane; Zeelenberg, Rene; Halberstadt, Jamin; Winkielman, Piotr. The dissertation author was the primary investigator and author of this material.

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*(7), 390-398.

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology*, *37*(1), 13-20.

Baker, W. E. (1999). When can affective conditioning and mere exposure directly influence brand choice? *Journal of Advertising*, *28*(4), 31-46.

Balogh, R., & Porter, R. H. (1986). Olfactory preferences resulting from mere exposure in human neonates. *Infant Behavior and Development*, *9*(4), 395-401.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull, & R. S. Wyer (Eds.), *Advances in social cognition, Volume III: Content and process specificity in the effects of prior experiences* (pp. 61-88). Hillsdale, NJ: Lawrence Erlbaum Associates.

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*(2), 142.

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.

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*(8), 706-712.

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. (1992). Stimulus recognition and the mere exposure effect. *Journal of Personality and Social Psychology, 63*, 545-552.

Bradley, M. M., & Lang, P. J. (2007). Emotion and motivation. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 581–607). Cambridge, UK: Cambridge University Press.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, I., & Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology, 26*, 116–120.

Butler, L. T. and Berry, D. C. (2004). Understanding the relationship between repetition priming and mere exposure. *British Journal of Psychology, 95*(4). 467-487.

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*(1), 63-75.

DeBruine, L. M. (2005). Trustworthy but not lust-worthy: Context-specific effects of facial resemblance. *Proceedings of the Royal Society of London B: Biological Sciences, 272*(1566), 919-922.

Deese, J. (1959). Influence of inter-item associative strength upon immediate free recall. *Psychological Reports*, *5*(3), 305-312.

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.*

Dreisbach, G. & Fischer, R. (2015). Conflicts as aversive signals for control adaptation. *Current Directions in Psychological Science, 24*, 255-260.

Fang, X., Singh, S., & Ahluwalia, R. (2007). An examination of different explanations for the mere exposure effect. *Journal of Consumer Research, 34*(1), 97-103.

Fazendeiro, T., Winkielman, P., Luo, C., & Lorah, C. (2005). False recognition across meaning, language, and stimulus format: Conceptual relatedness and the feeling of familiarity. *Memory and Cognition, 33*, 249-260.

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*(6), 636-644.

Gallo, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 120-128.

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.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*(1), 32-41.

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.

Halberstadt, J. B. (2006). The generality and ultimate origins of the attractiveness of prototypes. *Personality and Social Psychology Review, 10,* 166–183.

Halberstadt, J., Pecher, D., Zeelenberg, R., Wai, L.I., & Winkielman, P. (2013). Two faces of attractiveness: Making beauty-in-averageness appear and reverse. *Psychological Science, 24,* 2343-2346.

Halberstadt, J.B., & 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.

Halberstadt, J. & Winkielman, P. (2014). Easy on the eyes, or hard to categorize: Classification difficulty decreases the appeal of facial blends. *Journal of Experimental Social Psychology, 50, 175–183.*

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223-233.

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*(5), 1191.

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

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*(5754), 1680-1683.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization the role of prejudice and facial affect in race categorization. *Psychological Science*, *15*(5), 342-345.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*, 482-553.

Johnson, M. H. (2005). Subcortical face processing. *Nature Reviews Neuroscience*, *6*(10), 766-774.

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.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*.

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.

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.

Kornmeier, J., & Bach, M. (2012). Ambiguous figures—what happens in the brain when perception changes but not the stimulus. *Frontiers in Human Neuroscience, 6*(51).

Kouchaki, M., Smith-Crowe, K., Brief, A. P., & Sousa, C. (2013). Seeing green: Mere exposure to money triggers a business decision frame and unethical outcomes. *Organizational Behavior and Human Decision Processes*, *121*(1), 53-61.

Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language, 37*, 555–583.

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, 37*, 682-683.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1*, 115–121.

Love, B.C. (2013). Categorization. In K.N. Ochsner and S.M. Kosslyn (Eds.) *Oxford Handbook of Cognitive Neuroscience* (pp. 342-358). Oxford Press.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*(4), 330.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*(3), 252–271.

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

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.

Monahan, J.L., Murphy, S.T., & Zajonc, R.B. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science, 11*, 462–466.

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology, 85*, 1035-1048.

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

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*(4), 726-747.

Neta, M., Kelley, W. M., & Whalen, P. J. (2013). Neural responses to ambiguity involve domain-general and domain-specific emotion processing systems. *Journal of Cognitive Neuroscience, 25*(4), 547-557.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Obermiller, C. (1985). Varieties of mere exposure: The effects of processing style and repetition on affective response. *Journal of Consumer Research*, 17-30.

Phaf, R. H., & Rotteveel, M. (2005). Affective modulation of recognition bias. *Emotion, 5*, 309–318.

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

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*(6), 922–934.

Posner, M.I., Keele S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.

Principe, C. P. & Langlois, J. H. (2012). Shifting the prototype: Experience with faces influences affective and attractiveness preferences. *Social Cognition, 30*(1), 109-120.

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.

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition, 19*(1), 57-70.

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.

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological Science, 7*, 105–110.

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

Roediger, H. L., Wixted, J. T. & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In Nadel, Lynn, and Sinnott-Armstrong, Walter, (Eds.). *Memory and Law* (pp. 84-118). New York, NY: Oxford University Press.

Rubenstein, A.J., Kalakanis, L., and Langlois, J.H. (1999). Infant preferences for attractive faces: A cognitive explanation. *Developmental Psychology, 35*, 848–855.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, *7*(4), 413-422.

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25*, 638-656.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychonomic Bulletin and Review, 4*, 145-166.

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70*, 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*(1), 50-64.

Thornhill, R., & Gangestad, S.W. (1993). Human facial beauty: Averageness, symmetry and parasite resistance. *Human Nature, 4*, 237–269.

Topolinski, S., Erle, T. M., & Reber, R. (2015). Necker's smile: Immediate affective consequences of early perceptual processes. *Cognition, 140*, 1–13.

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.

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*(4), e10283.

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science*.

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.

Wang, W., Li, B., Gao, C., Xiao, X., & Guo, C. (2015). Electrophysiological correlates associated with contributions of perceptual and conceptual fluency to familiarity. *Frontiers in Human Neuroscience, 9*, 321.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide*

using statistical software. CRC Press.

Whittlesea, B.W.A. (2002). False memory and the discrepancy attribution hypothesis: The prototype-familiarity illusion. *Journal of Experimental Psychology: General, 131*, 96–115.

Whittlesea, B. W. A. & Price, J. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory and Cognition, 29*, 234-246.

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17*, 799–806.

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.

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.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*, 1–27.

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.

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*(3), 259.

Zebrowitz L. A., Zhang Y. (2012). Neural evidence for reduced apprehensiveness of familiarized stimuli in a mere exposure paradigm. *Social Neuroscience, 7*, 347–358.

CHAPTER 2:

Are you smiling or have I seen you before? Familiarity makes faces look happier

Evan W. Carr, Timothy F. Brady, & Piotr Winkielman

(manuscript under review for publication)

Abstract

Mere exposure is a classic psychological phenomenon. It is well established that repetition leads to a "warm glow of familiarity," but how it occurs remains largely unknown. Across two experiments, we show for the first time that familiarity enhances actual perceived happiness of facial expressions. In Experiment 1, using a paradigm where participants' responses were orthogonal to happiness to avoid response biases, we found that trained (familiar) faces were deemed happier than untrained (novel) faces. In Experiment 2, we replicated this effect with a rapid "happy or angry" categorization task. Using psychometric function fitting, we found that participants needed less actual happiness to be present in trained (compared to untrained) faces in order to classify them as happy. Critically, our results also dissociate prominent models of mere exposure, by demonstrating that familiar faces appear happier through selective enhancement of positive stimulus features (rather than reduction of negative stimulus features).

Keywords: mere exposure, affect, familiarity, perception, facial expressions

Introduction

For more than a century, psychology has been fascinated with the "warm glow of familiarity" (Fechner, 1876; James, 1890; Titchener, 1915). This preference for previously encountered stimuli — better known as the *mere exposure effect* (Zajonc, 1968) — not only occurs across different tasks, modalities, and stimuli, but it also informs numerous emotion-cognition models and applications to real-world settings (e.g., Baker, 1999; Balogh, & Porter, 1986; Obermiller, 1985; Pettigrew & Tropp, 2008; Rhodes, Halberstadt, & Brajkovich, 2001; Tremblay, Inoue, McClannahan, & Ross, 2010). The dominant explanations propose that repetition associates the stimulus with an absence of negative consequences (Zajonc, 2001) and reduces uncertainty (Lee, 2001), or that repetition facilitates processing (Bornstein & D'Agostino, 1994), with such fluency experienced as positive (Winkielman, Schwarz, Fazendeiro, & Reber, 2003).

Interestingly, despite more than 100 years' worth of mere exposure research, the nature of this effect is still mysterious. One key question concerns the processing stage at which familiarity creates positivity: Does mere exposure impact early stimulus processing (during actual perception) or is it purely a judgment phenomenon (occurring at later stages)? Another key question concerns the nature of affective change: Does familiarity change positive affect, negative affect, or both? This paper explores these key questions and tests several novel predictions, using important social stimuli — emotional facial expressions. Our proposal is that familiarity enhances perceived happiness of facial expressions. Further, we suggest that this effect involves the selective amplification of positive stimulus features (rather than the reduction of negative stimulus features).

These predictions are grounded in several areas of previous research on emotion and mere exposure. First, past studies have found that familiarity increases a variety of preference judgments. Many experiments use ratings of liking or attractiveness, but there is some

preliminary evidence that it extends to ratings of happiness, at least for neutral faces (e.g., Claypool, Hugenberg, Housely, & Mackie, 2007). This raises the possibility that familiarity actually makes the face "look" happier. This possibility fits with evidence that mere exposure occurs on physiological measures of affect (e.g., greater right-frontal EEG asymmetry and increased smiling via facial electromyography; De Vries et al., 2010; Harmon-Jones & Allen, 2001) and without any evaluative context (Garcia-Marques, Prada, & Mackie, 2016).

Thus, mere exposure may imbue facial stimuli with intrinsic positivity, which can be picked up in perception and classification tasks. This is related to previous work suggesting that perception depends on perceiver's own affective state (including perception of facial expressions; e.g., Phelps, Ling, & Carrasco, 2006). Here, we use two different tasks to measure early perceptual effects (Experiment 1) and rapid classification judgments (Experiment 2). We also systematically explore familiarity effects using psychometric functions across different levels of emotion expressions.

Current Research

The current experiments tested how familiarity with another individual impacts rapid perceptual judgments of their emotional facial expressions. After participants were exposed to neutral expressions of selected individuals, they then judged the level of happiness in emotional face blends (morphs going from angry to happy) from both familiar and unfamiliar individuals. Importantly, our studies were designed to provide distinctive predictions from prominent models on the connection between familiarity and valence (see predictions in Figure 2.1, which also incorporates weighted familiarity-positivity estimates from exposure on neutral expressions, according to a similarity gradient; Gordon & Holyoak, 1983).

First, the *nonspecific activation* account (Mandler, Nakamura, & Van Zandt, 1987) and a related *fluency amplification* account (Albrecht & Carbon, 2014) assume that repetition enhances

activation or facilitates access to the dominant stimulus features, regardless of their valence. Therefore, in our experiments, familiarity could just enhance the impact of valenced features in the facial expressions — where happy expressions from familiar targets appear happier, but angry expressions from familiar targets also appear angrier.

The second alternative is the *generalized positivity shift* account, which assumes that familiarity elicits broad positive affect that imbues positivity to all expressions, regardless of their valence. This idea is implicit in the notion of generalized "warm glow" (Monin, 2003; Tichener, 1915), with the "glow" functioning like positive mood that makes "everything" better (Schwarz & Clore, 1993).

The next two classes of accounts assume separable effects of familiarity on the positive vs. negative affect system (Cacioppo & Berntson, 1994). On the third account, termed the *negative skew*, repetition selectively dampens negative affect, without enhancing positive affect. This notion is implied by proposals that repetition "unlearns" negative reactions (Zajonc, 2001) and reduces uncertainty (Lee, 2001). Consequently, the negative skew account predicts the greatest repetition benefit for stimuli with negative features (where only angry faces appear less angry).

Finally, a fourth prediction is made by *hedonic skew* accounts. On these views, familiarity selectively impacts positive features. This is consistent with earlier studies showing that familiarity specifically increases positive affect, rather than reduces negative affect (Garcia-Marques, Mackie, Claypool, & Garcia-Marques, 2004; Harmon-Jones & Allen, 2001; Winkielman & Cacioppo, 2001). In the current study, this suggests that only happy expressions from familiar individuals would appear happier (with no change to angry expressions). Theoretically, this could not only indicate dissociable effects of familiarity on the positive vs. negative affect system (Cacioppo & Berntson, 1994), but it could also denote a target-dependent feature attribution (with positive affect reasonably attributed to only positive features; Schwarz,

2014). This is similar to some perceptual effects of earlier adaptation that only express when the correct stimulus features are present (e.g., McCollough, 1965).

Experiment 1

To test these alternatives in Experiment 1, we adapted a paradigm designed to look for influences on perception, independent of decision and response biases (originally developed by Carrasco et al. 2004; adapted to faces by Störmer & Alvarez, 2016). After exposing participants to neutral expressions of certain individuals (but not others), we gauged how familiarity would impact perceptual judgments of happiness on facial expressions from familiar and novel individuals. Our main dependent measure was the probability of selecting a trained (familiar) face as happier than an untrained (novel) face with objectively the same expression on the same trial.

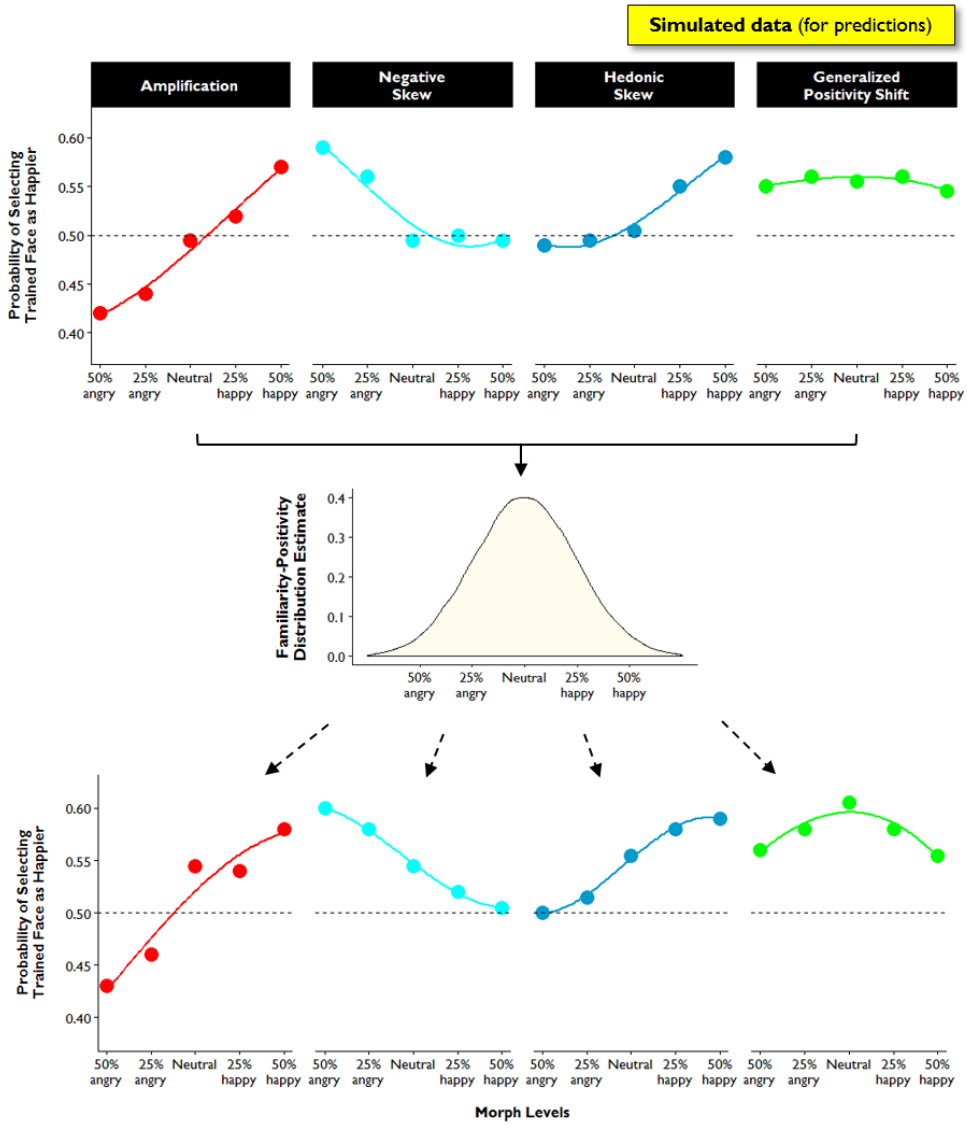


Figure 2.1: Qualitative predictions of different frameworks for Experiment 1. Under the *amplification* framework (red lines), happy expressions from familiar individuals will appear happier, but angry expressions from familiar individuals will also appear angrier (Albrecht & Carbon, 2014; Mandler, Nakamura, & Van Zandt, 1987). Under the *negative skew* framework (lighter blue lines), angry expressions from familiar individuals will appear less angry, with no change to happy expressions (Lee, 2001; Zajonc, 2001). Under the *hedonic skew* framework (darker blue lines), happy expressions from familiar individuals will appear happier, with no change to angry expressions (Garcia-Marques, Mackie, Claypool, & Garcia-Marques, 2004; Harmon-Jones & Allen, 2001; Winkielman & Cacioppo, 2001). Under the *generalized positivity shift* framework (green lines), all expressions from familiar individuals will be selected as happier, regardless of intensity or valence (Monin, 2003; Titchener, 1915). Since participants are trained on neutral expressions, each of these functions (top panel) would also be filtered through a familiarity-positivity distribution (middle panel). More familiarity and positivity would be centered on neutral expressions (since they are exact replicates of the expressions shown during training), with less familiarity-positivity effects as the expressions deviate farther from neutral (here, we assume a simple similarity gradient via mere exposure generalization effects; Gordon & Holyoak, 1983). The integrated predictions of the framework and familiarity gradients are shown in the bottom panel.

Method

Participants and equipment. Fifty University of California, San Diego (UCSD) undergraduates participated for course-credit, and all participants signed consent forms approved by the Human Research Protection Program (HRPP). We aligned our planned sample size according to previous studies on perceptual judgments for faces (e.g., Störmer & Alvarez, 2016), so in order to achieve maximal power, we set our target to $n = 50$.

During the main task, all stimuli were presented using E-Prime 2.0 software, on 17-inch Dell flat-screen PCs running Windows 7 (1,280 × 1,024 pixels; 60 Hz refresh rate).

Materials. We created our facial stimuli using still images from the Amsterdam Dynamic Facial Expression Set (ADFES; Van der Schalk, Hawk, Fischer, & Doosje, 2011). From the ADFES, we selected 12 different models to use for morphing (six males and six females). With the 100% angry, 100% happy, and neutral images for each model, we then generated morph stimuli at five different levels, including 50% angry, 25% angry, neutral, 25% happy, and 50% happy. This created a stimulus set where we had 60 unique stimuli (12 different models displaying five different levels of emotion). Note that these were only single-person morphs — meaning that models were only blended with different images of themselves (there were never blends of multiple models). All the faces were then cropped so that only the facial features were visible (i.e., no hair or neck).

Next, divided the models into "trained" (familiar) or "untrained" (novel), according to each individual participant. To do this, we created two different sets of images (set A and set B), each containing different halves (three males and three females) of the total number of models (six males and six females). Thus, each participant had to respond to each of the different models' emotional expressions, but each model was "trained" (familiar) or "untrained" (novel) for half the participants. As an example, if a participant was assigned to study set A models (not set B), they would be exposed to neutral expressions of set A models during training (phase 1),

after which they would see set A models' emotion morphs (trained targets) and set B models' emotion morphs (untrained targets) in a follow-up task (phase 2).

Design and procedure. There were two main phases in Experiment 1: an exposure training task (phase 1) and a speeded perceptual judgment task on faces (phase 2).

Before phase 1, all participants were told that they would be completing a memory task, where they would have to track and recall "both the color and number of either blue or green square probes that would randomly appear on different images." Critically, these "images" were neutral faces for the specific models that participants were randomly assigned to study, either in set A or set B (never both).

Participants then started phase 1, which involved 20 exposure trials to each of the six models in their assigned training set (120 total training trials). At the start of each exposure trial, participants would see a prompt that said "Remember the color and number of squares!" When they made a button-press, this triggered the start of the exposure trial. During the trial, a neutral expression for one of the training models (in set A or set B, depending on the participant's condition) would appear in the center of the screen for 5000 ms. Over the course of this trial, 200 ms blue and/or green squares would appear at random intervals (the frequency of blue and green squares was set to be anywhere from zero to nine exposures, with equal probabilities). After the 5000 ms trial ended, another screen appeared that asked participants to type a response to "How many BLUE squares did you see (0-9)?" and "How many GREEN squares did you see (0-9)?" To encourage high attention and effort throughout the memory task, participants were told that they would only advance to the next phase of the experiment once they hit a satisfactory level of performance (in reality, participants always completed the same number of training trials, to keep exposure consistent). With this task, we were able to give participants many passive exposures to neutral faces for only certain models, thus giving them selective familiarity for some models over others.

After participants finished the 120 training trials in phase 1, they moved onto phase 2. Our phase 2 paradigm was a modified version of a paradigm originally used to study attentional-cuing by Carrasco et al. (2004). This paradigm usually measures the effects of exogenous attention on perceptual processing, and while we were not interested in attention for the current study, a benefit of this type of task is that it controls for decision and response biases and has therefore been repeatedly used to make claims that effects have a perceptual locus (e.g., Carrasco et al. 2004; Störmer & Alvarez, 2016). For example, in Störmer & Alvarez (2016), two faces were simultaneously presented during each trial on the left and right sides of a computer screen, with one shifted upward and the other one shifted downward along the vertical axis. Participants were then instructed to report the vertical shift (upward or downward, using the up- and down-arrow keys on the keyboard) of the face they perceived as more intense on some dimension (e.g., in Störmer & Alvarez [2016], which face was more attractive). Since the response was orthogonal to the dimension of-interest, this eliminates the possibility of a simple response bias and reduces the likelihood of the effects originating in decision making. We adapted this paradigm to investigate how our phase 1 training task would impact speeded perceptual judgments of happiness in trained (familiar) vs. untrained (novel) faces.

Figure 2.2 shows a schematic of our phase 2 task. Each trial would begin with a reminder prompt for the participant to report whether they think the happier face is above or below the line (using the up- or down-arrow keys, respectively). Once they pressed a key to trigger the trial, a fixation cross would appear for 750 ms. Two lines would then appear on the left and right sides of the fixation for 500 ms, to mark the horizontal axis of the screen. Next, two faces would appear on the left or right side of the screen (each shifted +/- 128 pixels from the center fixation), where one face was shifted slightly upward and the other face was shifted slightly downward (each shifted +/- 154 pixels from the center fixation). If the participant thought the upper face appeared happier, they pressed the up-arrow key; if they thought the lower

face appeared happier, they pressed the down-arrow key. The participant was given up to 3000 ms to respond (any responses not logged within this time were excluded), and after they gave a response, a 1000 ms response confirmation screen was displayed before the next trial.

Critically, we also manipulated the types of faces shown on each trial. We always displayed one trained (familiar) model and one untrained (novel) model. Importantly, the trained and untrained face on each trial were always displaying the same objective level of emotion (i.e., 50% angry, 25% angry, neutral, 25% happy, or 50% happy). Therefore, on each trial, no response could be considered correct or incorrect (given that both faces were displaying the same type and level of emotion). We were interested in how participants' training with certain faces would influence their perceptual judgments of happiness. If training does impact perceptions of happiness, participants would choose the trained faces consistently more often than the untrained face, regardless of spatial location on the screen. Our instructions also emphasized that there were no correct or incorrect answers, and participants were told that the tasks in phases 1 and 2 were unrelated.

This task consisted of six blocks of 60 trials each. This was done to match each of the six trained faces with each of the six untrained faces, across five levels of emotion. Each face was also presented on both sides of the screen on different trials (i.e., 6 trained faces x 6 untrained faces x 5 emotion levels x 2 display positions = 360 total trials). In order to get accustomed to the phase 2 task, participants also completed eight practice trials using models from the ADFES that were not incorporated in either of the phase 1 or phase 2 tasks.

.

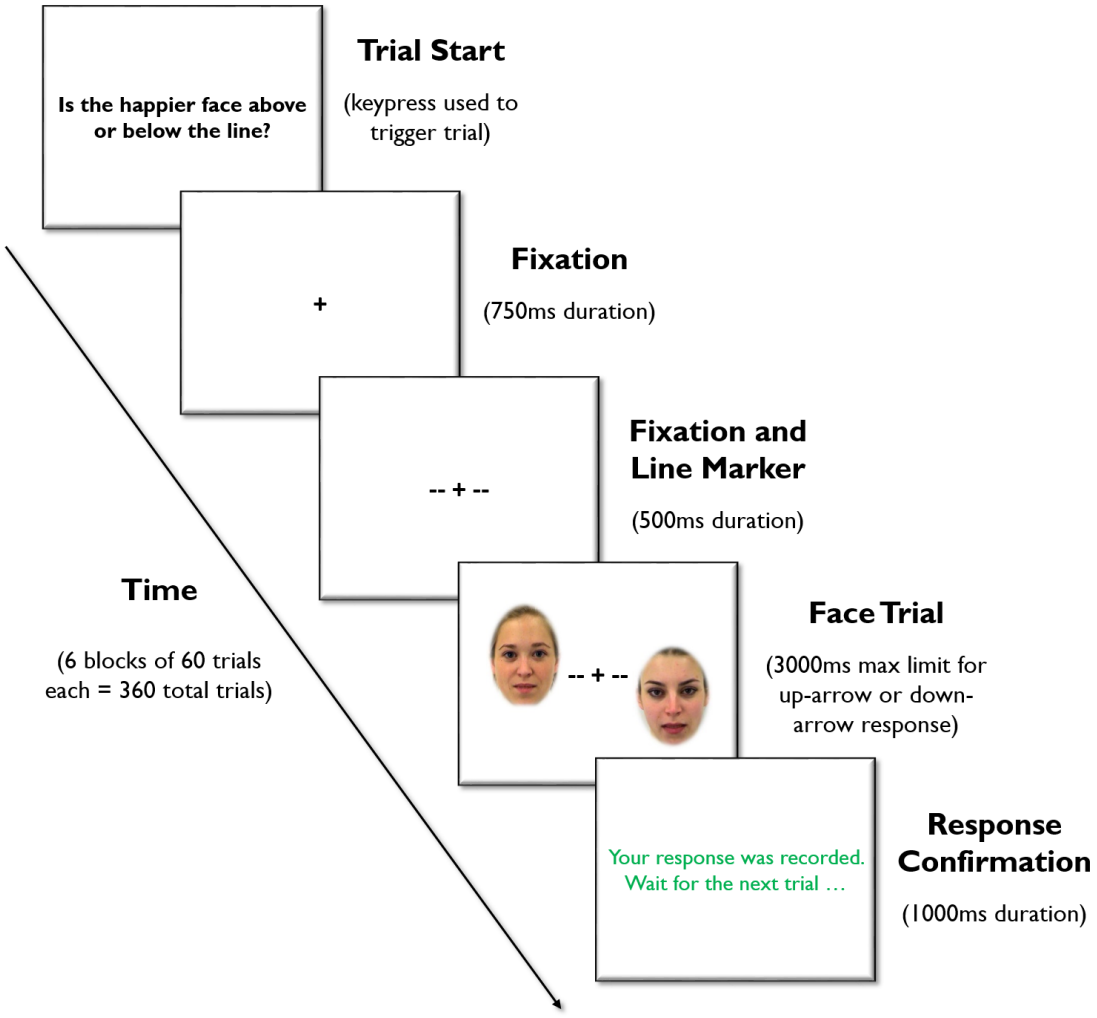


Figure 2.2: Design and procedure for the phase 2 task in Experiment 1.

Results

All repeated-measures analyses used multilevel modeling (MLM) via restricted maximum likelihood, since this method offers numerous analytical advantages — including more effective handling of unbalanced data with missing observations, reliance on fewer assumptions regarding covariance structures, and increased parsimony and flexibility between models (Bagiella, Sloan, & Heitjan, 2000). All models were built with the *lmerTest* package in R (Kuznetsova, Brockhoff, & Christensen, 2014), using the maximal random-effect structure that

would allow for model convergence (Barr, Levy, Scheepers, & Tily, 2013). To obtain *p*-value estimates for fixed-effects, we used Type III Satterthwaite approximations, which can sometimes result in decimal degrees of freedom (West, Welch, & Galecki, 2014). Before analysis, we excluded all trials with RTs less than 200 ms (recall that participants had a max limit of 3000 ms to respond).

In Experiment 1, we analyzed happiness response probabilities using an MLM with Stimulus Emotion (5: 50% angry, 25% angry, neutral, 25% happy, 50% happy) as a fixed-effects factor. Random-effects were fitted across participants.

Figure 2.3 shows the probability of the trained face being chosen as happier than the untrained face at the same level of emotion (across all trials). We plotted this as a function of emotion morph levels and fitted a bootstrapped logistic psychometric function to these response probabilities, using the *quickpsy* (Linares & López-Moliner, 2015) and *ggplot2* (Wickham, 2009) packages in R.

Overall, participants judged trained faces as happier than untrained faces (when compared against 50% chance level), $t(49) = 2.35$, $p = 0.02$, but this also varied as a function of the emotion, resulting in a main effect of Stimulus Emotion, $F(4, 84.82) = 3.59$, $p = .009$. This demonstrated that as the positive features in the faces increased (going from 50% angry to 50% happy), participants were more likely to judge the trained face as happier. Specifically, while participants did not differ from chance at 50% angry, $t(49) = -0.35$, *ns*, or 25% angry, $t(49) = 1.46$, *ns*, they were significantly above chance in judging trained faces as happier for neutral expressions, $t(49) = 2.56$, $p = .01$, 25% happy expressions, $t(49) = 2.20$, $p = .03$, and 50% happy expressions, $t(49) = 3.35$, $p = .002$. Note that while the final curve in Figure 2.3 does appear linear across morph levels, it is indeed a logistic psychometric function. Theoretically, these results mirror the prediction made by *hedonic skew* frameworks (see Figure 2.1).

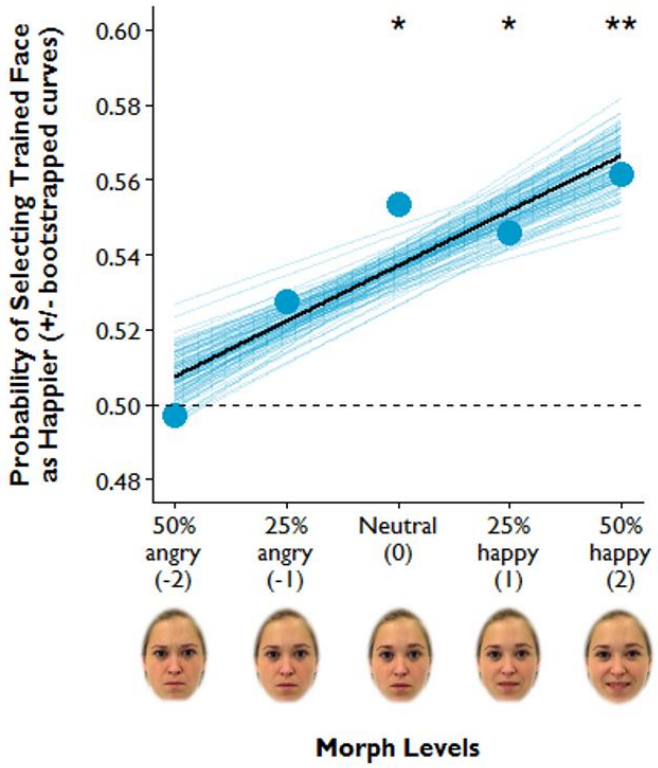


Figure 2.3: Results for phase 2 in Experiment 1. The *y*-axis indicates participants' response probabilities for selecting trained faces as happier (over untrained faces) at the same level of emotion on the same trial. Stimulus emotion morph levels are shown on the *x*-axis. The dots show overall means for each emotion level (across all trials). The blue lines show bootstrapped curves from psychometric fitting of a logistic function (generated using 100 bootstrapped samples), with the final estimate of this curve shown as the black line. While the final curve does appear linear across morph levels, note that this is indeed a logistic function.

Experiment 2

In Experiment 1, we demonstrated that mere exposure influences perception of happiness, suggesting that familiarity increases positivity at early processing stages. Importantly, this familiarity-positivity effect increased in intensity as the test expressions became happier (with no effects for 25% or 50% angry). This supports *hedonic skew* frameworks, which argue that the "warm glow of familiarity" operates via changes in positive affect (i.e., enhancement of

positive features), rather than negative affect.

In Experiment 2, we aimed to replicate and extend this effect with a categorization and judgment task. After training, participants were required on each trial to quickly decide whether a single face (from a familiar or novel individual) was "happy or angry." The key benefit of this "single-face" design over the "dual-face" design from Experiment 1 is that it allows for a direct measure of actual happiness level at each level of emotion (via psychometric function fitting for trained and untrained faces), rather than a judgment relative to another simultaneously presented face. We also had participants give a percentage estimate (0-100%) for how happy they thought the face looked on each trial. These percentage ratings not only provided a secondary measure, but they also focused on a more deliberative judgement (no time limit), rather than a first impression (rapid categorization response).

Method

Participants, materials, and equipment. Forty University of California, San Diego (UCSD) undergraduates participated for course-credit, and all participants signed consent forms approved by the Human Research Protection Program (HRPP). We planned our sample size to align with previous studies on classification judgments for faces (e.g., Winkielman, Olszanowski, & Gola, 2015), so in order to achieve adequate power, we decided on a target of $n = 40$. Our face stimuli, equipment hardware, and software were the same as Experiment 1.

Design and procedure. Our main changes for Experiment 2 were with the phase 2 task design. The phase 1 exposure training task was the same as Experiment 1. Instead of the dual-face perceptual task used in phase 2 of Experiment 1, we changed this to a speeded forced-choice classification paradigm in Experiment 2.

Figure 2.4 shows a schematic of the phase 2 task. Here, there were five blocks of 60 trials (300 total trials), where each block presented all 60 unique morph stimuli across individual

trials (12 models x 5 emotion morphs each). On each trial, a fixation cross was displayed for 1000 ms, after which one face stimulus would appear in the center of the screen, which could be either a trained or untrained model (displaying a 50% angry, 25% angry, neutral, 25% happy, or 50% happy expression). Participants were instructed to categorize, as quickly and accurately as possible, whether they thought the face was happy or angry, using the "Z" and "M" keys on the keyboard (response key pairs randomized across trials). They were told that they would only have up to 3000 ms to respond, and any response longer than that time limit would be counted as incorrect. After the classification, another question would appear that asked "How happy did that face appear to you?" On this question, participants could type answers that varied anywhere from 0 to 100% in a response box shown on the screen (0% = *no happiness at all*; 100% = *as happy as possible*).

Once participants completed all 300 trials for phase 2, they were debriefed and given credit for their participation.

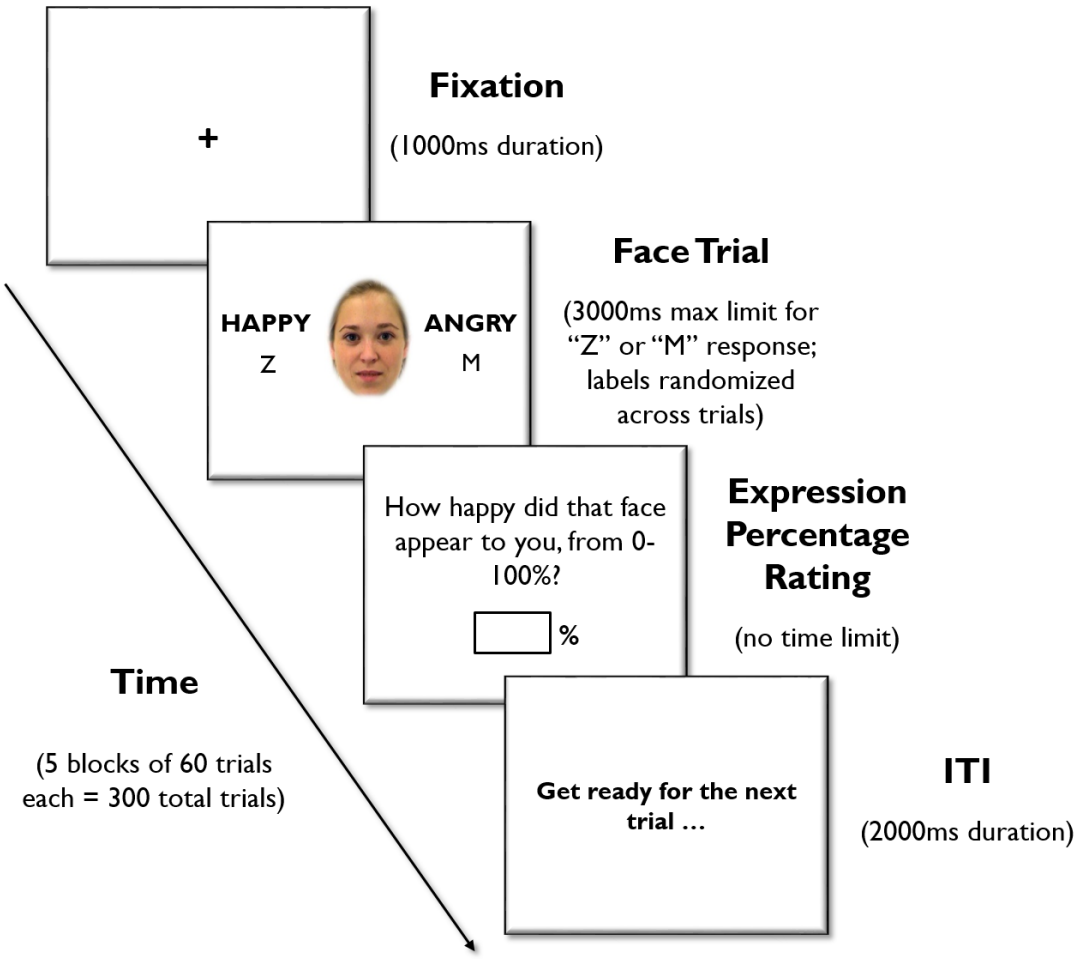


Figure 2.4: Design and procedure for phase 2 task in Experiment 2.

Results

As with Experiment 1, we used multilevel modeling (MLM) strategies to analyze data from Experiment 2 (Bagiella, Sloan, & Heitjan, 2000; Barr, Levy, Scheepers, & Tily, 2013; Kuznetsova, Brockhoff, & Christensen, 2014; West, Welch, & Galecki, 2014). Similar to before, we excluded all trial responses where RTs were less than 200 ms (recall that participants were only given a max limit of 3000 ms for their face classifications).

We also fit logistic psychometric functions for participants' forced-choice happy/angry

classifications on the trained and untrained faces, using the *quickpsy* (Linares & López-Moliner, 2015) and *ggplot2* (Wickham, 2009) packages in R. These psychometric curves were fitted via direct maximization of the likelihood, according to the following form:

$$\varphi(x) = \gamma + (1 - \gamma - \lambda) * fun([1 + \exp(-\beta * (x - \propto))]^{-1})$$

where γ is the guess rate, λ is the lapse rate, and *fun*( ) is the sigmoidal-shape logistic function with asymptotes at 0 and 1 (Linares & López-Moliner, 2015).

The fitting of these psychometric functions allowed for the calculation of thresholds for different emotion morph levels, which we elaborate on in the next section.

Multilevel modeling on happiness classification probabilities. First, we analyzed the probability of happy classifications, according to a Training (2: trained, untrained) x Stimulus Emotion (5: 50% angry, 25% angry, neutral, 25% happy, 50% happy) fixed-effects structure. Random-effects were fitted by participants and stimulus model IDs.

As predicted, we observed a Training x Stimulus Emotion interaction, $F(4, 50.46) = 3.17$, $p = .02$. Follow-up tests revealed that while participants gave a higher percentage of happy classifications for trained faces at neutral, 25% happy, and 50% happy, this difference was greatest at the 25% happy morph level, $b = 0.10$, $SE = 0.02$, $t(68.40) = 3.29$, $p = .002$. Also replicating the pattern from Experiment 1, there were no differences between trained and untrained happy classifications at the 50% angry morph level, $b < 0.10$, $SE = 0.02$, $t(67.90) = -1.13$, *ns*, or 25% angry morph level, $b < 0.10$, $SE = 0.02$, $t(68.70) = 0.48$, *ns*. Note that we also detected a main effect of Stimulus Emotion, which only indicated that happy classifications were more likely as the faces became more positive (going from 50% angry to 50% happy; see Figure 2.5, top panel).

Group-level psychometric function fitting. As previously mentioned, we also fit

logistic psychometric functions to the happy/angry classification data, according to training condition (i.e., trained vs. untrained faces). We did this in order to calculate morph level thresholds at different response probabilities. More specifically, using the fitted curves, one can obtain a point on the morph continuum (somewhere between 50% angry and 50% happy), which corresponds to a certain percentage of responses for one option over the other (e.g., 20% happy classifications, 40% happy classifications, etc.). By bootstrapping these curves, one can also calculate a 95% confidence interval estimate around these thresholds, in order to compare across training conditions (here, we generated 100 bootstrap samples for each function).

Figure 2.5 displays the results. To gauge how trained vs. untrained thresholds changed across the morph continuum, we assessed four different response probabilities: 20%, 40%, 60%, and 80% happy classifications. Interestingly, similar to the familiarity-positivity pattern from Experiment 1, the logistic function for trained faces showed greater percentage of happy classifications specifically at greater happy response probabilities (i.e., where the faces contained more positive features; see Figure 2.5, top panel). Consequently, for the 60% and 80% happiness response probabilities, participants required *less* actual happiness to be present in the trained emotion morphs (compared to untrained emotion morphs), in order for them to classify them as happy (see Figure 2.5, bottom panel). Note that the differences in trained vs. untrained thresholds at 60% and 80% are significant at $\alpha = .05$, but at the lower 20% and 40% happiness response probabilities (where the faces contained more negative features), there were no significant differences between trained and untrained thresholds (see Figure 2.5, bottom panel).

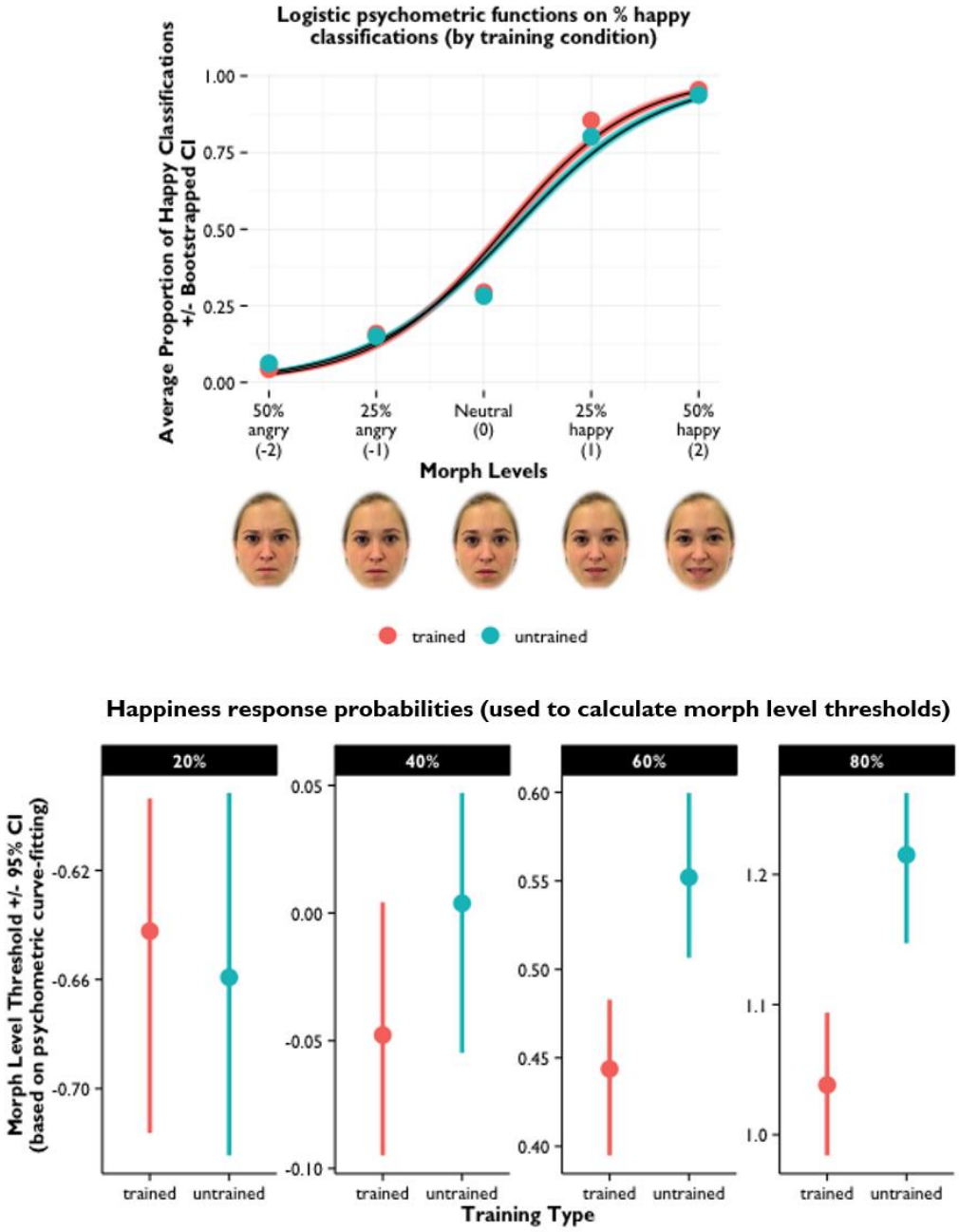


Figure 2.5: Results for psychometric function fitting (top panel) and morph level thresholds (bottom panels) in phase 2 of Experiment 2. The top panel displays separate logistic functions for happy/angry classification data on both trained and untrained faces (red = trained; blue = untrained), across different emotion morph levels on the *x*-axis (i.e., 50% angry, 25% angry, neutral, 25% happy, 50% happy). Bootstrapped 95% confidence intervals are displayed around each logistic function, which were generated using 100 samples. Morph level thresholds were calculated across different response probabilities (i.e., how much happiness was needed in the faces to achieve 20%, 40%, 60%, and 80% happy classifications). These thresholds are shown on the *y*-axis with 95% confidence intervals around the point estimates, along with the training type on the *x*-axis (red = trained; blue = untrained).

Participant-level psychometric function fitting. In addition to the group-level logistic functions and thresholds fit to all happy/angry classification data (see Figure 2.5), we also fit participant-level functions. More specifically, we followed the same steps as with the group-level data, but here, we fit logistic functions to each individual participant's classification data. From this, we were able to calculate morph level thresholds for each participant at the different happiness response probabilities (20% happy, 40% happy, 60% happy, and 80% happy, as before), for both trained and untrained faces. Next, we fit an MLM to predict these participant-level morph thresholds, using a Training (2: trained, untrained) x Response Probability (4: 20% happy, 40% happy, 60% happy, 80% happy) fixed-effects structure. Once again, we fit maximal random-effects according to each participant, to the point that allowed for model convergence.

This showed similar results to the group-level analysis depicted in Figure 2.5. We observed a Training x Response Probability interaction, $F(3, 195.00) = 6.94$, $p < .001$. Thresholds for trained faces were marginally less than untrained faces at the 60% response probability, $b = -0.10$, $SE = 0.07$, $t(72.30) = -1.67$, $p = .10$, and significantly less at the 80% response probability, $b = -0.20$, $SE = 0.07$, $t(72.30) = -3.00$, $p = .004$. There were no differences between trained and untrained faces at the 20% response probability, $b < 0.10$, $SE = 0.07$, $t(72.30) = 0.76$, *ns*, or the 40% response probability, $b < 0.10$, $SE = 0.07$, $t(72.30) = -0.57$, *ns*.

Classification RTs. We also analyzed participants' classification RTs using similar MLM methods, according to a Training (2: trained, untrained) x Stimulus Emotion (5: 50% angry, 25% angry, neutral, 25% happy, 50% happy) fixed-effects structure and random-effects for participants and stimulus model IDs. Recall that we only included RTs between 200 and 3000 ms. We also $\log_{10}$-transformed the remaining valid RTs to normalize the response distribution.

Figure 2.6 (top panel) shows the results. We observed a Training x Stimulus Emotion interaction, $F(4, 151.01) = 3.28$, $p = .01$. Post-hoc breakdowns of this interaction demonstrated that participants had marginally faster RTs both to classify 25% happy trained faces vs. 25%

happy untrained faces, $b < 0.10$, $SE = 0.005$, $t(183.60) = -1.88$, $p = .06$, as well as 50% happy trained faces vs. 50% happy untrained faces, $b < 0.10$, $SE = 0.005$, $t(181.10) = -1.87$, $p = .06$. Interestingly, however, participants were also *slower* to classify neutral expressions for trained models (compared to untrained models), $b < 0.10$, $SE = 0.005$, $t(187.80) = 2.24$, $p = .03$. There were no RT differences by training for 50% angry expressions, $b < 0.10$, $SE = 0.005$, $t(181.50) = -0.53$, *ns*, or 25% angry expressions, $b < 0.10$, $SE = 0.005$, $t(185.20) = 0.43$, *ns*. Note that we also observed a main effect of Stimulus Emotion, $F(4, 68.27) = 18.33$, $p < .001$, which only showed that participants generally had the slowest classification RTs to neutral expressions, compared to faces that displayed more pure emotion (i.e., 50% angry, 25% angry, 25% happy, and 50% happy).

Happiness percentage estimates. Recall that we also had participants give a 0-100% free-response estimate for the level of happiness they saw in each face, after every phase 2 trial. This gave us an alternative metric of happiness perception on a more deliberative judgment (rather than a first impression from a rapid classification). To analyze this, we ran an MLM with a Training (2: trained, untrained) x Stimulus Emotion (5: 50% angry, 25% angry, neutral, 25% happy, 50% happy) fixed-effects structure and random-effects for participants and stimulus model IDs.

Figure 2.6 (bottom panel) displays the main results. These results showed a similar pattern to the Experiment 1 perceptual responses (see Figure 2.3) and Experiment 2 classification responses (see Figure 2.5). We observed a Training x Stimulus Emotion interaction, $F(4, 39.68) = 4.36$, $p = .004$. Follow-up tests revealed that participants estimated trained faces as happier (compared to untrained faces), specifically at the 25% happy level, $b = 2.30$, $SE = 0.72$, $t(63.20) = 3.21$, $p = .002$, and marginally at the 50% happy level, $b = 1.40$, $SE = 0.72$, $t(62.60) = 1.92$, $p = .06$. There were no training differences at the 50% angry level, $b = -0.70$, $SE = 0.72$, $t(62.70) = -0.92$, *ns*, or the 25% angry level, $b = -0.50$, $SE = 0.72$, $t(63.40) = -0.67$, *ns*. Also, note that while

happiness estimates for trained expressions were greater than untrained expressions at the neutral level, this difference did not reach significance, $b = 0.30$, $SE = 0.72$, $t(63.90) = 0.35$, *ns*. We also observed a main effect of Stimulus Emotion, $F(4, 152.63) = 116.14$, $p < .001$, which only showed that participants' happiness estimates increased as the morph levels became more positive (going from 50% angry to 50% happy).

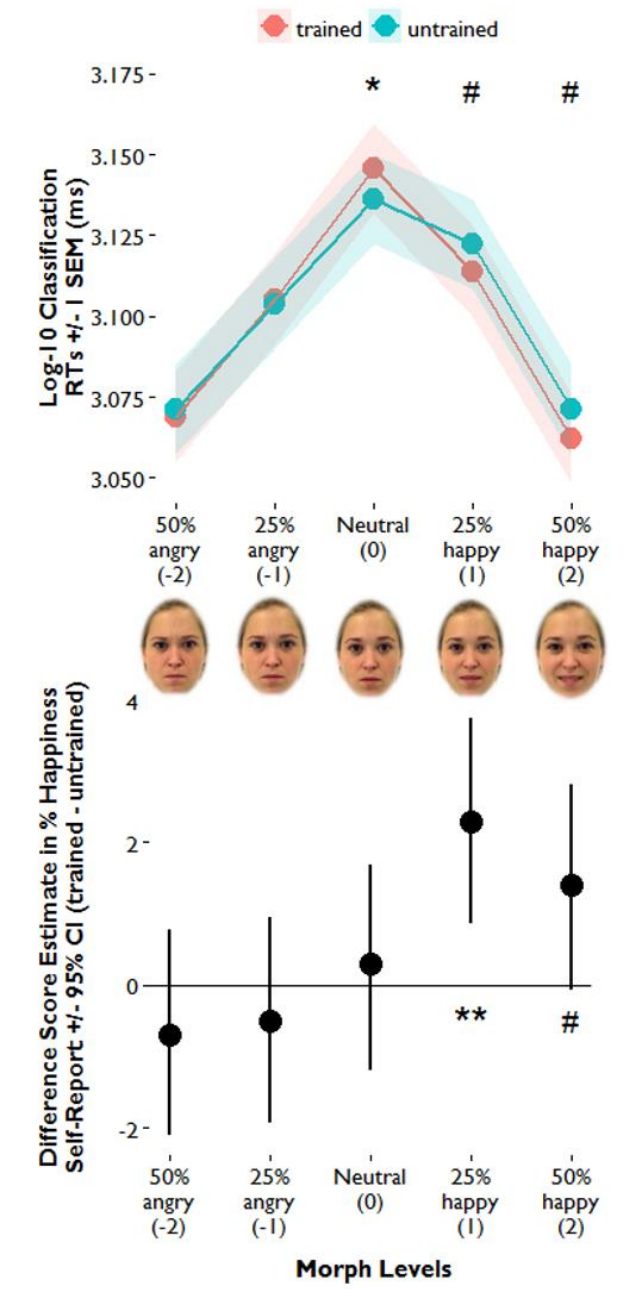


Figure 2.6: Results for classification RTs (top panel) and self-report estimates of happiness percentage (bottom panel) in phase 2 of Experiment 2. The top panel shows $\log_{10}$-transformed classification RTs on the *y*-axis for both trained and untrained faces (red = trained; blue = untrained), across different emotion morph levels on the *x*-axis (i.e., 50% angry, 25% angry, neutral, 25% happy, 50% happy). The bottom panel shows difference scores in happiness estimates on the *y*-axis by training condition (trained face estimates – untrained face estimates), across different emotion morph levels on the *x*-axis (i.e., 50% angry, 25% angry, neutral, 25% happy, 50% happy). ** $p < .01$, * $p < .05$, # $p < .10$.

Discussion

The current results suggest that familiarity (derived via *mere exposure*) impacts early stimulus processing, modifying the perception of others' facial expressions. Across different tasks that involved speeded perceptual judgments (Experiment 1), rapid forced-choice classifications (Experiment 2), and deliberative estimates of happiness (Experiment 2), participants judged familiar individuals' expressions as happier — particularly when the expressions were neutral-to-positive (see Figures 2.3, 2.5, and 2.6). Psychometric function fitting also revealed that exposure training led to lower happiness thresholds at higher happiness response probabilities. Simply put, as the morphs became happier, participants needed less actual happiness to be present in trained faces to classify them as happy rather than angry (see Figure 2.5). Critically, our findings cannot be explained by simple response biases, given that participants only judged trained faces as happier at certain levels of emotion. This specific pattern also emerged across multiple tasks (see Figures 2.3, 2.5, and 2.6) and even when responses were orthogonal to the dimension of happiness (Experiment 1).

Recall the different predictions from prominent models on the familiarity-valence link (see Figure 2.1). Our results offer support for *hedonic skew* frameworks, which posit that familiarity influences positive affect (but not negative affect) and gets expressed via positive features (but not negative features) (Garcia-Marques, Mackie, Claypool, & Garcia-Marques, 2004; Harmon-Jones & Allen, 2001; Winkielman & Cacioppo, 2001). Generally, our results seem inconsistent with models proposing that repetition leads to *amplification* of pre-existing features (or nonspecific activation; Albrecht & Carbon, 2014; Mandler, Nakamura, & Van Zandt, 1987), selective decrease in negative affect (or a *negative skew*; Lee, 2001; Zajonc, 2001), or a *generalized positivity shift* (Monin, 2003; Tichener, 1915). The selectivity in our results for neutral-to-positive faces could be due either to the separation of the positive and negative affect system (Cacioppo & Berntson, 1994) or due to constraints on the attribution of positive affect to

positive features (Schwarz, 2014). In any case, our results show that this effect occurs at early stages of processing. It is worth noting that our results were obtained with human faces, and animal research suggests that mere exposure could reduce distress to novelty (e.g., Zajonc, Markus, & Wilson, 1974). Further, in Experiment 2, participants not only judged trained faces as happier, but they were also faster to classify those faces as such (see Figure 2.6, top panel). This is consistent with the idea that mere exposure works via enhancement of fluency (Winkielman, Schwarz, Fazendeiro, & Reber, 2003), but this interpretation is speculative as more intense positive affect will also speed up classification.

To our knowledge, this is the first evidence that mere exposure modulates perception of facial affect. Our results are consistent with past findings that familiarity enhances ratings of neutral expressions (e.g., Claypool, Hugenberg, Housley, & Mackie, 2007), but they go significantly beyond previous work. First, we used tasks designed to assess early perceptual processes, rather than only scale ratings. Second, our facial expressions varied in the type and intensity of emotion being displayed, allowing us to estimate psychometric functions and thresholds (see Figure 2.5). In turn, we showed that the effects of familiarity on happiness judgments are dependent on the positive features in the test expression. As the expressions become happier, participants were more likely to judge the trained face as happier (both when directly compared to an untrained face [Experiment 1] and when presented alone [Experiment 2]; see Figures 2.3, 2.5, and 2.6). Finally, and perhaps most importantly, these design features allowed us to differentiate theoretical accounts of mere exposure.

Future work should evaluate the boundary conditions of these findings — especially for why familiarity did not reduce the negativity of angry expressions. We suggest this occurs because familiarity works selectively on positive affect and is more easily attributed to positive features. However, some views suggest that anger is "special," perhaps leading it to be gated from familiarity influences (anger superiority effect; e.g., Pinkham, Griffin, Baron, Sasson, &

Gur, 2010). One might be able to answer this question by using different negative emotion morphs. However, note that in the current studies, participants were equally sensitive to neutral-to-happy as to neutral-to-angry transitions (Figure 2.5), suggesting limited evidence for this position (at least in our data).

Finally, claims about the perceptual nature of any effect are hotly debated (Firestone & Scholl, 2015) and we hesitate to claim that familiarity influences early vision. Nevertheless, evidence for affective influences on perception is still reasonably strong (Niedenthal & Setterlund, 1994; Vetter & Newen, 2014). Future research should examine such early effects with tasks that can gauge visual pop-out for trained faces (e.g., continuous flash suppression or visual search paradigms). For now, however, our results suggest that familiar faces do appear happier.

Chapter 2 is, in full, under review for publication of the material. Carr, Evan W.; Brady, Timothy, F.; Winkielman, Piotr. The dissertation author was the primary investigator and author of this material.

References

Albrecht, S., & Carbon, C. C. (2014). The Fluency Amplification Model: Fluent stimuli show more intense but not evidently more positive evaluations. *Acta Psychologica, 148*, 195-203.

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology*, *37*(1), 13-20.

Baker, W. E. (1999). When can affective conditioning and mere exposure directly influence brand choice? *Journal of Advertising*, *28*(4), 31-46.

Balogh, R., & Porter, R. H. (1986). Olfactory preferences resulting from mere exposure in human neonates. *Infant Behavior and Development*, *9*(4), 395-401.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bornstein R.F. & D'Agostino, P.R. (1994). The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Social Cognition, 12*, 103–128.

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*, 401-423.

Claypool, H. M., Hugenberg, K., Housley, M. K., & Mackie, D. M. (2007). Familiar eyes are smiling: On the role of familiarity in the perception of facial affect. *European Journal of Social Psychology*, *37*(5), 856-866.

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.

Fechner, G. T. (1876). *Vorschule der aesthetik* (Vol. 1). Breitkopf & Härtel.

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behavioral and Brain Sciences*, 1-72.

Garcia-Marques, T., Mackie, D. M., Claypool, H. M., & Garcia-Marques, L. (2004). Positivity can cue familiarity. *Personality and Social Psychology Bulletin*, *30*(5), 585-593.

Garcia-Marques, T., Prada, M., & Mackie, D. M. (2016). Familiarity increases subjective positive affect even in non-affective and non-evaluative contexts. *Motivation and Emotion*, *40*(4), 638-645.

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.

Harmon-Jones, E., & Allen, J. J. (2001). The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality and Social Psychology Bulletin, 27*(7), 889-898.

James, W. (1890). *Principles of psychology*. New York: Holt, Rinehart, & Winston.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models. R package version 2.0-20.

Lee, A. Y. (2001). The mere exposure effect: An uncertainty reduction explanation revisited. *Personality and Social Psychology Bulletin, 27*(10), 1255–66.

Linares, D. & López-Moliner, J. (in press). quickpsy: An R package to fit psychometric functions for multiple groups.

Mandler, G., Nakamura, Y., & Van Zandt, B. J. (1987). Nonspecific effects of exposure on stimuli that cannot be recognized. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 646.

McCollough, C. (1965). Color adaptation of edge-detectors in the human visual system. *Science, 149*(3688), 1115-1116.

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology*, *85*(6), 1035-1048.

Niedenthal, P. M., & Setterlund, M. B. (1994). Emotion congruence in perception. *Personality and Social Psychology Bulletin*, *20*(4), 401-411.

Obermiller, C. (1985). Varieties of mere exposure: The effects of processing style and repetition on affective response. *Journal of Consumer Research*, 17-30.

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*(6), 922–934.

Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological science*, *17*(4), 292-299.

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition, 19*(1), 57-70.

Schwarz, N., & Clore, G. L. (2003). Mood as information: 20 years later. *Psychological Inquiry*, *14*(3-4), 296-303.

Schwarz, N. (2014). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Psychology Press.

Titchener, E.B. (1915). *A beginner's psychology*. New York: Macmillan.

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*(4), e10283.

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, *27*, 62-75.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide using statistical software*. CRC Press.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer Science & Business Media.

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation increases positive affect. *Journal of Personality and Social Psychology*, *81*(6), 989-1000.

Winkielman, P., Olszanowski, M., & Gola, M. (2015). Faces in-between: Evaluations reflect the interplay of facial features and task-dependent fluency. *Emotion*, *15*(2), 232-242.

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.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*, 1–27.

Zajonc, R. B., Markus, H., & Wilson, W. R. (1974). Exposure, object preference, and distress in the domestic chick. *Journal of Comparative and Physiological Psychology*, *86*, 581-585.

Zajonc, R.B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, *10*, 224–228.

CHAPTER 3:

Is that a human? Categorization (dis)fluency drives evaluations of agents ambiguous on human-likeness

Evan W. Carr, Galit Hofree, Kayla Sheldon, Ayse P. Saygin, & Piotr Winkielman

In press for publication in

Journal of Experimental Psychology: Human Perception and Performance

Abstract

A fundamental and seemingly unbridgeable psychological boundary divides humans and non-humans. Essentialism theories suggest that mixing these categories violates "natural kinds." Perceptual theories propose that such mixing creates incompatible cues. Most theories suggest that mixed agents, with both human and non-human features, obligatorily elicit discomfort. In contrast, we demonstrate top-down, cognitive control of these effects — such that the discomfort with mixed agents is partially driven by disfluent categorization of ambiguous features that are pertinent to the agent. Three experiments tested this idea. Participants classified three different agents (humans, androids, and robots) either on the human-likeness or control dimension and then evaluated them. Classifying on the human-likeness dimensions made the mixed agent (android) more disfluent, and in turn, more disliked. Disfluency also mediated the negative affective reaction. Critically, devaluation only resulted from disfluency on human-likeness — and not from an equally disfluent color dimension. We argue that negative consequences on evaluations of mixed agents arise from *integral disfluency* (on features that are relevant to the judgment at-hand, like ambiguous human-likeness). In contrast, no negative effects stem from *incidental disfluency* (on features that do not bear on the current judgment, like ambiguous color backgrounds). Overall, these findings support a top-down account of why, when, and how mixed agents elicit conflict and discomfort.

Keywords: emotions, categorization, judgment and decision making, cognitive processing, human-computer interaction

Introduction

A key psychological distinction is the one that divides human and non-human. Treating an agent as human (or "human-like") fundamentally changes how individuals perceive, interpret, behave, communicate, or empathize (Dennett, 1971). Children and adults consider human-likeness to be a deep, unchangeable trait — a type of psychological essentialism (Medin & Ortony, 1989; Prentice & Miller, 2007). Such essentialist beliefs arise early in life (Gelman, 2004), structure social categories and stereotypes (Bastian & Haslam, 2006; Haslam, Rothschild, & Ernst, 2000; Haslam, Bastian, Bain, & Kashima, 2006; Howell, Weikum, & Dyck, 2011), drive attention (Bastian & Haslam, 2007), and guide automatic motor responses (Bastian, Loughnan, & Koval, 2011).

This human/non-human boundary is often investigated with agents that mix human and non-human features. It has long been noticed that mixed stimuli (e.g., chimeras, griffins, hybrids, mannequins, human dolls, etc.) generally elicit a sense of weirdness and discomfort (Frenkel-Brunswik, 1949; Jentsch, 1906). This issue gained renewed importance with the recent proliferation of bionic humans and androids (i.e., robots with human-like features and behaviors; Ishiguro, 2007; Mori, MacDorman, & Kageki, 2012). Here, the key psychological question is what causes such discomfort to mixed agents.

Extant theories propose a variety of processes. From the aforementioned essentialism perspective, individuals perceive the human/non-human boundary as fundamentally unbridgeable, and negative reactions spontaneously arise from the inappropriate blending of two different "natural kinds" or separate "essences" (see Demoulin, Leyens, & Yzerbyt, 2006). Proponents of essentialism argue that certain properties are intrinsic and immutable traits of the agent in-question, and perceivers can essentialize a variety of dimensions (e.g., gender, race, sexual orientation, etc.). Critically though, essentialized categories are usually said to have some key defining characteristics: clear and discrete boundaries from other categories, involuntary and

unchanging membership, and observable features that reflect something about the underlying function of the agent (Prentice & Miller, 2007). While theoretically distinct (but still related), perception research tends to focus on "mismatches," or conflicting cues in visual, auditory, and motion processing of mixed agents (Katsyri, Forger, Makarainen, & Takala, 2015; MacDorman, Green, Ho, & Koch, 2009; Mitchell et al., 2011; Seyama & Nagayama, 2007). Other proposals highlight the potential role of conflict between the apparent lack of conscious experience paired with perceived agency (Waytz, Gray, Epley, & Wegner, 2010). This idea is related to suggestions that negative reactions to mixed agents reflect low-level mechanisms involving disease avoidance (MacDorman & Ishiguro, 2006). Overall, most (if not all) theories suggest that mixed agents (with both human and non-human features) spontaneously elicit conflict and discomfort.

In contrast to these assumptions, we propose that the relative dislike for mixed agents can be modified by contextual factors — providing a major theoretical qualification to these earlier claims. Specifically, the current paper argues that reactions to mixed agents involve an interaction between top-down higher-order cognitive processes and bottom-up perceptual factors. Basically, we suggest that the sense of "weirdness" is not inherent to the perception of mixed agents, but rather is generated when people classify such agents into human versus non-human categories — resulting in the experience of categorization disfluency. This disfluency triggers negative affect, which generalizes to agent evaluations (as we explain next). Note that our top-down framework is not simply another level of analysis for essentialist or perceptual conflict theories. While there are some versions of bottom-up perceptual theories that could be considered compatible with our fluency account, these frameworks still differ on the type of role disfluency serves in generating negative affect (i.e., either as a key component or a mere byproduct). We will return to these theoretical distinctions in the *Discussion*. Generally, we posit that the subjective boundary between human and non-human entities can be markedly

reconstructed via categorization processes, which has downstream consequences on judgments, evaluations, and attitudes.

Our proposal builds on several lines of previous research related to *fluency* — or changes in processing speed and effort (Schwarz, 1998). Here, much of the original research has focused on perceptual fluency (manipulated by enhancing low-level "surface" features, like clarity, contrast, readability, typicality, etc.; e.g., Carr, Rotteveel, & Winkielman, 2016; Reber & Schwarz, 1999; Reber, Winkielman, & Schwarz, 1998) or conceptual fluency (manipulated by facilitating the processing of stimulus meaning, as with semantic priming; e.g., Rajaram & Geraci, 2000; Whittlesea, 1993). Evidence shows that processing ease (fluency) increases evaluations, whereas *dis*fluency lowers them, as reflected in self-reports and physiological measures (Winkielman & Cacioppo, 2001). According to the *hedonic fluency model*, easy processing elicits positive affect, which is then (mis)attributed to the target stimulus (Winkielman, Schwarz, Fazendeiro, & Reber, 2003). This positive affect presumably emerges because fluency reflects (or probabilistically signals) lower conflict and greater coherence in stimulus processing.

Importantly, such effects also extend to categorization (dis)fluency — or the effort needed to determine category membership (Halberstadt & Winkielman, 2013). Note that categorization fluency differs from perceptual fluency in that stimulus features remain unchanged. Rather, categorization fluency is ultimately task-dependent, and processing difficulty instead depends on which (un)ambiguous feature dimensions are highlighted by the current task. In other words, if a stimulus is ambiguous on some dimension, it will elicit *dis*fluency (and negative affect), but only in contexts requiring categorization on that particular dimension. To illustrate, Owen, Halberstadt, Carr, & Winkielman (2016) had participants categorize morphed male-female faces either on the central ambiguous dimension (gender) or an auxiliary unambiguous dimension (race). Devaluation for mixed-gender faces only occurred in the gender-

categorization condition, when the gender-ambiguous faces were made disfluent (i.e., difficult to categorize). Similar effects have also been shown for bi-racial faces (Halberstadt & Winkielman, 2014) and those displaying mixed emotions (Winkielman, Olszanowski, & Gola, 2015).

This theoretical approach raises an important (and previously unexplored) possibility that changing someone's categorization mindset can alter negative responses to mixed agents, which contain supposedly "unbridgeable" human and non-human features. If so, their categorization on the human-likeness dimension should elicit disfluency (and devaluation), but this should be reduced during classification on an alternative feature dimension that is still social, yet unambiguous (e.g., gaze orientation cues). Further, as we explain next, to produce devaluation, perceivers must not only experience disfluency when processing the agent, but this disfluency must be derived from an integral (essential) rather than incidental (non-essential) feature of the agent.

Another contextual factor that dictates fluency-devaluation effects is the underlying relevance of the feature in-question. Some features can be *integral* (or pertinent to the judgment at-hand) while others can be *incidental* (or peripheral to the current task), as discussed by Bodenhausen (1993). Crucially though, evaluative consequences of (dis)fluency depend on the perceived relevance of the experience for the judgment at-hand (see Schwarz, 2010, for a review). For example, disfluency in categorizing someone else's emotional expression may lead the participant to judge the target as less trustworthy (i.e., since emotion is a key factor in trust judgments; Winkielman, Olszanowski, & Gola, 2015). However, one can speculate that an equally difficult categorization experience on a secondary dimension (e.g., ambiguous hair color) will not lower trustworthiness judgments, given that hair color does not bear on trust. Consequently, for the current experiments, devaluation effects may only follow from disfluency that occurs in response to an agent's integral features, rather than an equally disfluent experience on incidental features. We expected integral disfluency (i.e., ambiguous human-likeness) to have

downstream negative consequences on evaluations of mixed agents, but no effects to occur from incidental disfluency (i.e., ambiguous color backgrounds). We will return to this issue in the *Discussion*.

In short, we propose that affective responses to mixed agents are partially driven by top-down mechanisms. Categorization difficulty should trigger negative reactions to agents with ambiguous features when the current categorization task focuses on the ambiguous dimension (and only when the dimension is integral to the nature of the agent).

Current Research

We investigated our predictions in three experiments. In each study, participants saw and rated three different agents — one that was clearly human (a human), one that was clearly not human (a robot), and one that had both human and non-human features (an android). Participants rated these agents under two different conditions that varied on categorization requirements. Some participants categorized the agents as "human or non-human" — a selectively difficult task for the android agent — while others performed a control task on which the mixed agent was not selectively difficult.

To preview the results, categorization fluency impacted evaluative ratings of the different agents. Androids were devalued more so in the human-classification condition — both compared to a control task of speeded stimulus detection (Experiment 1) and an alternative task of social gaze categorization (Experiment 2). Using data from Experiments 1 and 2, multilevel mediation analyses demonstrated that categorization effort mediated the relationship between agent "mixed-ness" and weirdness judgments, but only for the human-classification condition.

Critically, the results show that these effects cannot be explained by simple misattribution of incidental task effort. This is because devaluation effects for mixed agents were eliminated in Experiment 3, which elicited disfluency by having subjects view androids with ambiguous color.

These results suggest that disfluency translates into devaluation only when generated by ambiguous dimensions that are integral (i.e., human-likeness classification) rather than those that are incidental (i.e., color-classification) to the evaluative judgment.

As such, these results challenge the assumption that human and non-human categories are "unbridgeable" — where mixing produces obligatory conflict and devaluation. Instead, they argue for a more cognitively flexible, task-sensitive link between ambiguity, task-relevant fluency, and evaluation.

Experiment 1

Experiment 1 tested whether categorization disfluency impacted evaluative ratings of ambiguous non-human agents, using highly controlled images that placed androids towards the middle of the human-likeness continuum. We focused on weirdness ratings, given that dimensions of "eeriness" and "strangeness" are deemed important in many studies on non-human agents (Ho & MacDorman, 2010).

Method

Participants and stimuli. Fifty-two undergraduates (M_{age} = 21.00 years, SD_{age} = 2.44 years; 42 females) at the University of California, San Diego (UCSD) participated for course credit and signed consent forms approved by the UCSD Human Research Protections Program.

Our stimuli were still images taken from the Saygin-Ishiguro Action Database (SIAD), which includes actions performed by human, android, and robot agents (Saygin & Stadler, 2012). For the android agent, these stimuli featured Repliee Q2 (see Figure 3.1, middle images), which was developed at Osaka University in collaboration with Kokoro Inc. Importantly, with brief exposures, people can mistake Repliee Q2 for a human being (Ishiguro, 2006). Repliee Q2 is an advanced humanoid robot that has 42 degrees of freedom and can make head and upper body

movements. Since Repliee Q2 can make head and body movements, the images displayed both the head and upper body of the agents. For the human condition, the stimuli featured the female adult on whom Repliee Q2's appearance was modeled after (see Figure 3.1, left images). For the robot condition, the surface elements of the Repliee Q2 android were removed and stripped of human-like features, in order to reveal the underlying materials (e.g., wiring, metal limbs, and joints; see Figure 3.1, right images). For the purpose of the current experiments, to further match and standardize these images on perceptual cues (e.g., coloring, clothing, etc.), we edited them using Adobe PhotoShop CS2 to maximally control for each agent's general physical appearance (see Figure 3.1).

To create the stimuli, Repliee Q2 was photographed performing eight different actions (nudging, grasping, drinking, waving, talking, turning, wiping, and lifting), both with and without its original human-like surface features (i.e., android and robot agent conditions, respectively). Critically, the female model for Repliee Q2 then naturally performed the same actions several times, and the version of those actions that most closely matched that of Repliee Q2 were selected for the stimuli (Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012; Saygin & Stadler, 2012). In short, the stimuli were highly controlled images of human, android, and robot actions, which mainly varied on the dimension of the agents' human-likeness.

The stimulus images were grayscale and cropped to 400 × 400 pixels. We used these grayscale images in Experiments 1 and 2, but note that the coloring of the images was edited for theoretical reasons in Experiment 3. Figure 3.1 shows examples of the different stimuli. All agents were photographed in the same room, with the same background, lighting, and camera settings.

Figure 3.1: Example stimuli from Experiments 1-2 (grayscale images; top row) and Experiment 3 (blue/green images; bottom row). Each image depicted one of three agent types (i.e., human [left column], android [middle column], or robot [right column]), doing one of eight actions. Some actions showed the individual with their head oriented towards the camera (e.g., "talking" on the top row) while others showed the individual with their head oriented away from the camera (e.g., "turning" on the bottom row).

Design and procedure. Participants were randomly assigned to one of two classification conditions (human-classification or no-classification). In the human-classification condition, participants were instructed to judge whether or not individuals in the different pictures were "human or non-human." In the no-classification condition, participants were instead told to "hit the spacebar as fast as possible, once the picture appears on the screen" (see Figure 3.2).

Next, participants proceeded through three counterbalanced blocks of 24 trials each (totaling 72 trials), each of which asked for a weirdness rating of the agent in the video. After each trial, participants gave the rating using a 1 (*not at all*) to 7 (*very much*) scale. Prior to rating each image, human-classification participants were told to categorize, "as quickly and accurately as possible, whether the agent in the picture was human or non-human," using the 'A' and 'L'

keys on the keyboard (response labels were randomized across trials). No-classification participants only used the spacebar to indicate the onset of the image. Each trial began with a 500 ms fixation, followed by the 3000 ms stimulus image (as soon as the participant responded, the image was replaced with the rating scale). The ITI for each trial was 5000 ms (see Figure 3.2).

Therefore, on each trial, we logged the participants' RT to classify the image and their weirdness ratings.

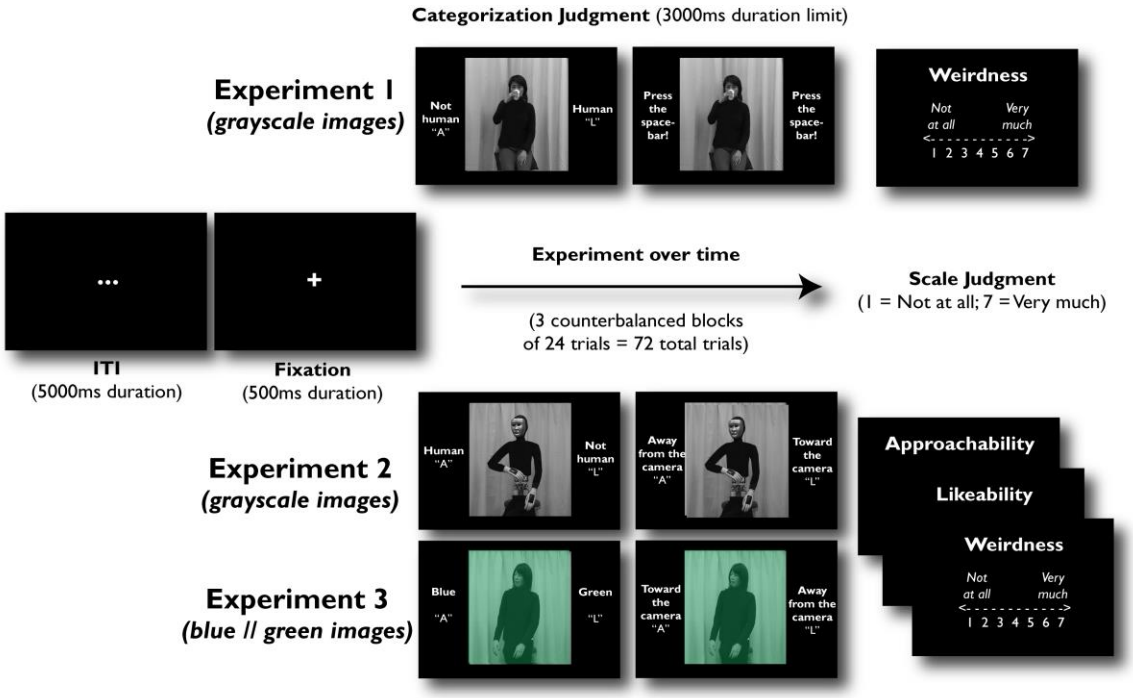


Figure 3.2: Design and procedure for Experiments 1, 2, and 3. Experiment 1 used grayscale images of all agent actions (first row), with a human-classification condition and no-classification condition (human/"toward" action example shown). Experiment 2 used grayscale images of all agent actions (second row) and used both a human-classification condition and orientation-classification condition (robot/"away" action example shown). Experiment 3 (bottom row) used 100% blue, 100% green, and 50/50 blue-green images of agent actions, with both a color-classification condition and orientation-classification condition (android/"away" action example shown). Note that Experiment 1 only involved weirdness ratings (all 72 trials), while Experiments 2 and 3 measured approachability, likeability, and weirdness (24 trials each; 72 trials total).

Results

Analysis strategy. All RTs and ratings were analyzed using trial-level data with multilevel models (MLMs; via restricted maximum likelihood estimation). MLMs more effectively handle hierarchical and unbalanced data with missing observations, relying on fewer assumptions regarding covariance structures and increasing parsimony and flexibility between models (Bagiella, Sloan, & Heitjan, 2000). Note that while we report MLM results here, due to the advantages over traditional ANOVA methods, all reported effects still replicate when using these traditional approaches.

MLMs were built with the *lme4* (Bates, Maechler, Bolker, & Walker, 2014) and *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2014) packages in R, using the maximal random-effects structure appropriate for the data (Barr, Levy, Scheepers, & Tily, 2013). Such a strategy strikes a balance in reducing possible Type I errors, while also avoiding overparameterization in the MLM (Bates, Kliegl, Vasishth, & Baayen, 2015). To obtain *p*-value estimates for fixed-effects, we used Type III Satterthwaite approximations, which can sometimes result in decimal degrees of freedom, based on the number of observations (West, Welch, & Galecki, 2014).

Across all experiments, subjects who performed at ≤ 50% accuracy during the main task were removed from all analyses, and for the remaining subjects, error trials were excluded. For Experiment 1, one subject performed at ≤ 50% accuracy and another subject did not adhere to task instructions — therefore, these two subjects were excluded from the total sample, leaving a final $n = 50$.

We focused on non-error trials to ensure that participants recognized the true human/non-human nature of the agent (i.e., whether it was actually human vs. non-human). We were also primarily interested to what extent participants' evaluations reflect the sheer effort of processing,

as opposed to possible categorization errors. However, all analyses for error trials can be found in footnotes for Experiment 1[5], Experiment 2[6], and Experiment 3[7].

Response RTs. Following previous fluency studies (Winkielman, Halberstadt, Fazendeiro, & Catty, 2006), we excluded trials with extremely fast (less than 200ms) and slow (greater than 3000ms) RTs, and the remaining RTs were $\log_{10}$-transformed to normalize the

[5] We could only analyze error trials in the human-classification condition for Experiment 1 (since there were no incorrect responses in the no-classification condition). To do this, we created a new MLM for only the human-classification condition, with Agent (3: human, android, robot) as the only fixed-effect factor. As expected, within the human-classification condition, participants had lower accuracy in response to mixed (android) agents — both compared to the human agent, b = -.29, SE = .01, t = -3.61, p = .002, dz = .75, and the robot agent, b = -.28, SE = .01, t = -3.50, p = .002, dz = .73.

[6] On Experiment 2, we analyzed accuracy using the same methods as RTs (since both classification conditions could have correct and incorrect responses). As expected, human-classification participants showed a greater number of errors selectively in response to the mixed agent (android). A Condition x Agent interaction, F(2, 167.80) = 43.26, p < .001, demonstrated that human-classification participants had lower accuracy for androids — both compared to the human agent, b = -.41, SE = .03, t = -12.54, p < .001, dz = 1.36, and the robot agent, b = -.41, SE = .03, t = -12.43, p < .001, dz = 1.35. Also, there was no difference between human and robot accuracy within the human-classification condition, b = .01, SE = .01, t = .56, *ns*, dz = .06. Importantly, for the orientation-classification condition, accuracy for the android did not significantly differ from the human agent, b = -.06, SE = .03, t = -1.76, *ns*, dz = .19, or the robot agent, b = .01, SE = .03, t = .34, *ns*, dz = .04. Both main effects were also significant. The main effect of Condition, F(1, 167.14) = 28.97, p < .001, demonstrated that orientation-classification participants were more accurate overall than human-classification participants. The main effect of Agent, F(2, 167.80) = 54.33, p < .001, demonstrated that participants were overall less accurate in response to the android compared to the other agents.

[7] We could not do the same analysis as Experiment 2 for accuracy in Experiment 3, since the android images in the color-classification condition were exactly 50/50 between blue and green (and thus, no response on those trials could be counted as correct or incorrect). However, we did analyze accuracy performance for all agent types in the orientation-classification condition (using similar methods as Experiment 1, by creating a new MLM only for the orientation-classification condition, with Agent [3: human, android, robot] as the only fixed-effect factor). We also checked overall accuracy for the human and robot images in the color-classification condition. First, on the orientation-classification condition, overall accuracy across agent types was high (M = 90.87%, SD = 28.80%). We did observe a main effect of Agent from the MLM, F(2, 111.90) = 12.14, p < .001, which showed that orientation-classification participants were more accurate in responding to the human agent — both compared to the android agent, b = -.04, SE = .01, t = 4.16, p < .001, dz = .52, and the robot agent, b = .05, SE = .01, t = 4.29, p < .001, dz = .54. There were no accuracy differences between the android and robot agents, b = .01, SE = .01, t = 0.69, *ns*, dz = .09. Second, within the color-classification condition, accuracy performance was comparably high for both the human agent (M = 95.71%, SD = 20.27%) and robot agent (M = 96.17%, SD = 19.20%). Once again, note that we could not code accuracy for the android agent in the color-classification condition, since they were exactly 50% blue // 50% green.

response distribution. Next, we created an MLM with a Condition (2: human-classification, no-classification) x Agent (3: human, android, robot) fixed-effects structure.[8]

We observed strong evidence for all RT effects in Experiment 1. Critically, we detected the predicted Condition x Agent interaction, $F(2, 99.61) = 3.55$, $p = .03$. Follow-up tests demonstrated that only the human-classification subject group took longer to respond to the android, both compared to the human agent, $b = .05$, $SE = .01$, $t = 4.42$, $p < .001$, $d_Z = .92$, and robot agent, $b = .05$, $SE = .01$, $t = 4.00$, $p < .001$, $d_Z = .83$. The no-classification group showed no differences between android and human RTs, $b = .01$, $SE = .01$, $t = .82$, *ns*, $d_Z = .16$, and a smaller difference between android and robot RTs, $b = .02$, $SE = .01$, $t = 2.10$, $p = .04$, $d_Z = .40$. Neither condition differed in response RTs between the human and robot agents (see Figure 3.3).

Note that we also observed both main effects of Condition, $F(1, 49.92) = 57.56$, $p < .001$, and Agent, $F(2, 99.61) = 11.18$, $p < .001$. Human-classification subjects had longer RTs, and the android took the most time to categorize across conditions.

Overall, the android was selectively disfluent (compared to both of the other agents). Critically, this occurred only within the human-classification condition.

[8] Note that all of the reported effects for Experiment 1 and 2 RTs still hold, both with and without error trials on the android agent (where participants classified the android as "human").

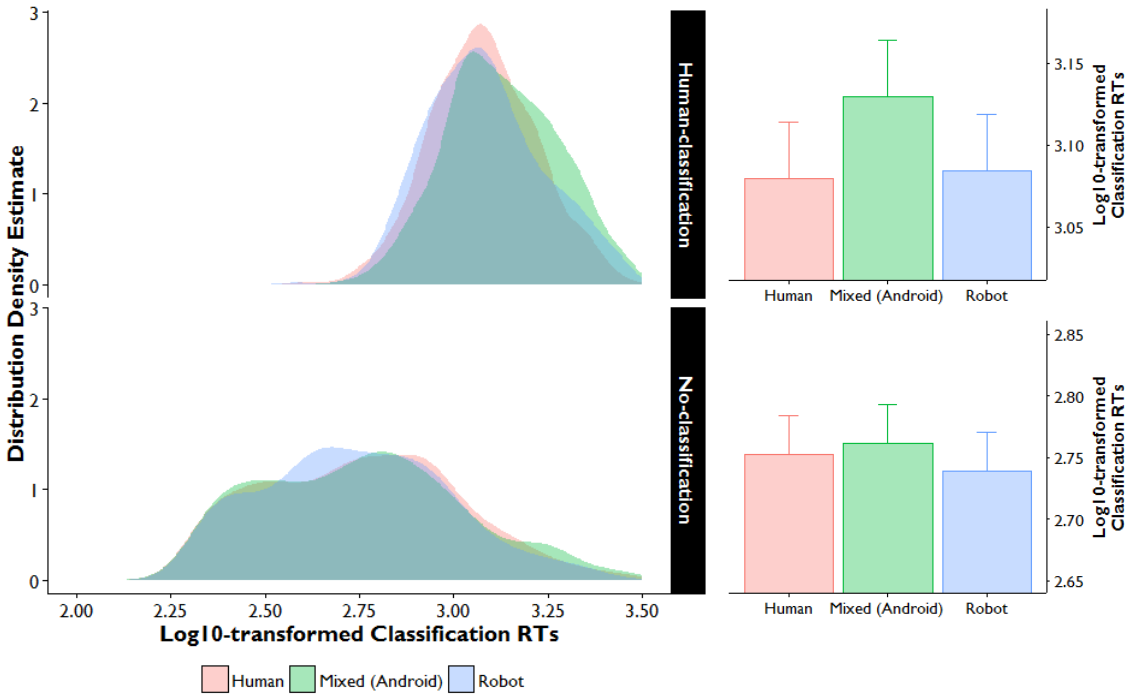


Figure 3.3: Density distributions and means/SEMs for $\log_{10}$-transformed RTs by classification condition (top row = human-classification; bottom row = no-classification) and agent type (indicated by colors) for Experiment 1.

Weirdness ratings. We analyzed weirdness ratings similar to the RTs, using an MLM with a Condition (2: human-classification, no-classification) x Agent (3: human, android, robot) fixed-effects structure.

Crucially, we found evidence for a Condition × Agent interaction, $F(2, 99.06) = 5.42$, $p = .006$. Post-hoc tests revealed that human-classification participants rated the android higher on weirdness (compared to the no-classification group; see Figure 3.4), $b = .79$, $SE = .32$, $t = 2.49$, $p = .01$, $d = .72$. Within the no-classification condition, participants rated the android as weirder than the human, $b = 2.36$, $t = 8.99$, $p < .001$, $d_Z = 1.73$, and the robot as weirder than the android, $b = 1.30$, $SE = .26$, $t = 4.97$, $p < .001$, $d_Z = .96$. Within the human-classification condition, participants still rated the android as weirder than the human, $b = 3.40$, $SE = .30$, $t = 11.45$, $p <$

.001, d_Z = 2.39, but there was no difference between the android and robot agents, b = .09, SE = .30, t = .31, ns, d_Z = .06.

We also detected a main effect of Agent, $F(2, 99.06) = 190.51$, $p < .001$, showing that overall, robots were judged as weirder than both the android and human agents. Importantly, this main effect of Agent Type occurred in both the experimental and control conditions (ps < .001). This indicates that participants in both conditions paid enough attention to the differences between human and non-human agents to form discriminative evaluations.

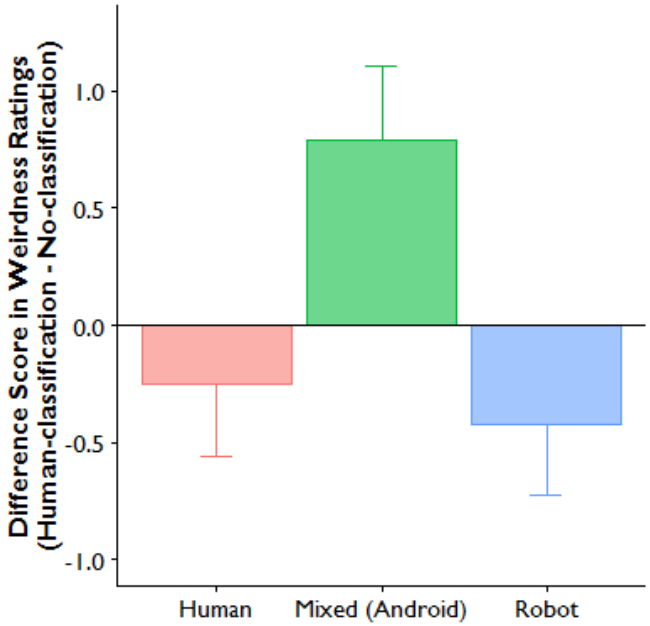


Figure 3.4: Weirdness difference scores by classification condition (human-classification – no-classification) across the different agent types (human, android, and robot). Error bars = ± 1 SEM.

Experiment 2

Experiment 1 showed that the devaluation of mixed agents (androids) was selectively amplified in the human-classification condition (compared to the no-classification condition) —

possibly because of disfluency caused by difficult categorization on the ambiguous human-likeness dimension.

We had three main goals for Experiment 2. First, we wanted to replicate the key effects from Experiment 1 with greater power and sample size. Second, we aimed to replace the no-classification control condition from Experiment 1 with a condition that was more closely matched to the experimental condition. To do this in Experiment 2, we used a social categorization control condition, which was designed to have a reasonable but equal difficulty for all agents, rather than selective difficulty for mixed agents (like in the experimental human-classification condition). Third, we added approachability and likeability scales (along with the weirdness ratings from Experiment 1), to assess the generalizability of fluency-devaluation effects to other dimensions.

Method

Participants and stimuli. One hundred seventy UCSD undergraduates (M_{age} = 20.08 years, SD_{age} = 2.20 years; 114 females) participated for course credit and signed consent forms approved by the UCSD Human Research Protections Program. All experimental stimuli were the same as Experiment 1 (see Figure 3.1).

Design and procedure. The design for Experiment 2 was similar to Experiment 1, but with two main changes.

First, the no-classification condition from Experiment 1 was replaced with the gaze orientation-classification condition in Experiment 2, which required detailed stimulus processing by asking participants to judge whether the agent's head was oriented "toward or away from the camera" (see Figure 3.2 and Li, Florendo, Miller, Ishiguro, & Saygin, 2015). The human-classification condition remained the same as Experiment 1.

Second, similar to Experiment 1, participants proceeded through three counterbalanced blocks of 24 trials each (totaling 72 trials). However, in Experiment 2, each rating block of 24 trials was split by different rating dimensions (i.e., approachability, likeability, *or* weirdness) to gauge the generalizability of the fluency effect. As before, after each trial, participants gave the rating using a 1 (*not at all*) to 7 (*very much*) scale, and both conditions made their respective classifications using the 'A' and 'L' keys on the keyboard (response labels were randomized across trials). All other timing and trial parameters were the same as Experiment 1 (see Figure 3.2).

Thus, in short, participants proceeded through three counterbalanced blocks of 24 trials each (totaling 72 trials). Each block was randomly assigned to one of three judgments (approachability, likeability, or weirdness; see Figure 3.2) and required categorization on the image stimuli from Experiment 1 (human-classification vs. orientation-classification).

Three subjects performed at ≤ 50% accuracy and were thus excluded from the total sample, leaving a final $n = 167$.

Results

Categorization RTs. We analyzed RTs in the same way as Experiment 1, using an MLM according to a Condition (2: human-classification, no-classification) x Agent (3: human, android, robot) fixed-effects structure.

As with Experiment 1, all effects were significant. Most importantly, we clearly replicated the Condition x Agent interaction, $F(2, 167.72) = 13.83$, $p < .001$. Human-classification participants took longer to categorize the android, both compared to the human agent, $b = .05$, $SE = .005$, $t = 9.92$, $p < .001$, $d_Z = 1.08$, and the robot agent, $b = .05$, $SE = .007$, $t = 7.84$, $p < .001$, $d_Z = .85$, but there was no difference between their human and robot RTs. Orientation-classification participants still took longer to classify the android compared to the

human agent, b = .02, SE = .005, t = 3.38, p < .001, d_Z = .37, but not the robot agent, b = .01, SE = .007, t = 1.10, *ns*, d_Z = .12, and they also showed no differences between human and robot RTs (see Figure 3.5).

Also similar to Experiment 1, we detected some less theoretically important effects. Specifically, there were strong main effects of both Condition, $F(1, 167.02) = 16.59$, $p < .001$, and Agent, $F(2, 167.72) = 44.78$, $p < .001$. Generally, these effects showed that orientation-classification subjects had longer RTs, and the android took the most time to categorize when aggregating across conditions.

In sum, once again, the android was selectively disfluent (compared to both the other agents) — but only in the human-classification condition.

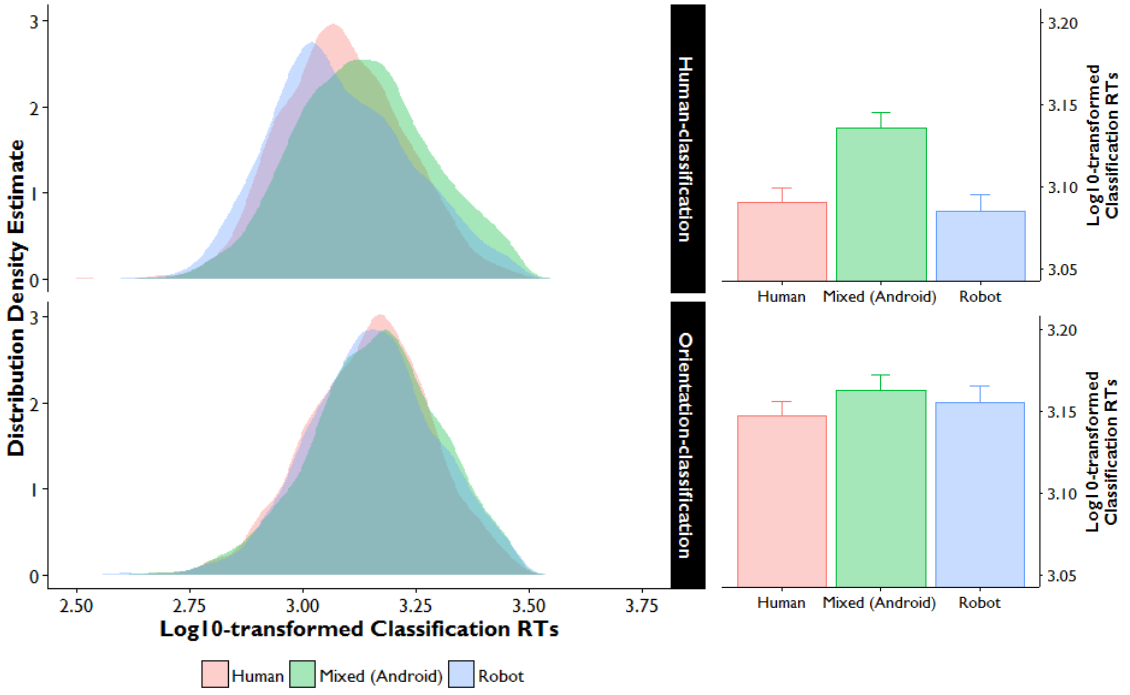


Figure 3.5: Density distributions and means/SEMs for log_{10}-transformed RTs by classification condition (top row = human-classification; bottom row = orientation-classification) and agent type (indicated by colors) for Experiment 2.

Scale ratings. We analyzed all scale ratings in the same way as Experiment 1, using MLMs with Condition (2: human-classification, no-classification) x Agent (3: human, android, robot) fixed-effects structures.

Approachability. On approachability, we detected a Condition x Agent interaction, $F(2, 159.94) = 11.24$, $p < .001$. Critically, human-classification participants rated the android lower in approachability (compared to the orientation-classification group), $b = -.64$, $SE = .19$, $t = -3.35$, $p = .001$, $d = .52$ (see Figure 3.6). Within the orientation-classification condition, participants rated the android as less approachable than the human, $b = -1.24$, $SE = .15$, $t = -8.55$, $p < .001$, $d_Z = .94$, and the robot as less approachable than the android, $b = -.45$, $SE = .16$, $t = -2.85$, $p = .005$, $d_Z = 0.31$. Within the human-classification condition, participants still rated the android as less

approachable than the human, b = -2.27, SE = .16, t = -13.96, p < .001, d_Z = 1.51, but there was no difference between the android and robot agents, b = .21, SE = .17, t = 1.23, *ns*, d_Z = 0.13.

Note that we also observed a strong main effect of Agent, $F(2, 159.94) = 168.87$, p < .001, such that the robot was less approachable than the android, which in turn was less approachable than the human. This replicates the pattern of evaluative ratings from Experiment 1. This main effect of Agent Type occurred in both the experimental and control conditions (ps < .001), suggesting that participants in both conditions of Experiment 2 paid attention to the differences between human and non-human agents.

Likeability. For likeability ratings, we observed very similar effects to approachability. We once again detected a Condition x Agent interaction, $F(2, 165.28) = 5.43$, p < .01. As with approachability ratings, human-classification participants gave lower likeability scores to androids (compared to the orientation-classification condition), b = -.48, SE = .19, t = -2.46, p = .02, d = .38 (see Figure 3.6). Within the orientation-classification condition, participants rated the android as less likeable than the human, b = -1.39, SE = .16, t = -8.68, p < .001, d_Z = .96, and the robot as less likeable than the android, b = -.31, SE = .15, t = 2.07, p = .04, d_Z = .23. Within the human-classification condition, participants still rated the android as less likeable than the human, b = -2.08, SE = .18, t = -11.70, p < .001, d_Z = 1.27, but there was no difference between the android and robot agents, b = .26, SE = .17, t = 1.54, *ns*, d_Z = .17.

Also similar to approachability, there was again significant evidence for a main effect of Agent, $F(2, 165.28) = 181.00$, p < .001, where the robot was less likeable than the android, which in turn was less likeable than the human. This main effect of Agent Type was present in both the experimental and control conditions (ps < .001).

Weirdness. With weirdness ratings, the central effects from Experiment 1 replicated. There was a basic main effect of Agent, $F(2, 166.89) = 253.97$, p < .001, with participants rating

the robot as weirder than the android, which was rated weirder than the human. This main effect occurred in both conditions (*p*s < .001).

Critically, we also observed a Condition × Agent interaction, $F(2, 166.89) = 7.12$, $p = .001$. Post-hoc breakdowns of this interaction revealed that human-classification participants rated the android higher on weirdness (compared to the orientation-classification group), $b = .71$, $SE = .25$, $t = 2.86$, $p = .005$, $d = .45$ (see Figure 3.6). Within the orientation-classification condition, participants rated the android as weirder than the human, $b = 1.94$, $SE = .19$, $t = 10.18$, $p < .001$, $d_Z = 1.12$, and the robot as weirder than the android, $b = .83$, $SE = .19$, $t = 4.30$, $p < .001$, $d_Z = .47$. Within the human-classification condition, participants still rated the android as weirder than the human, $b = 2.96$, $SE = .21$, $t = 13.88$, $p < .001$, $d_Z = 1.51$, but there was no difference between the android and robot agents, $b = .02$, $SE = .21$, $t = 0.10$, *ns*, $d_Z = .01$.

Composite positivity index. We also created a composite rating by averaging approachability, likeability, and reverse-coded weirdness scores — which yielded a general positivity index for each participant, towards each agent. This allowed us to gauge how (and to what magnitude) fluency effects on evaluation generalize more broadly, to overall positive and negative dimensions. Note that while we still used similar MLM methods for the composite rating, this MLM was run on subject means instead of trial-level data (since we needed to obtain a single composite score for each subject by averaging their responses across the other rating dimensions).

Critically, we detected clear evidence for a Condition x Agent interaction, $F(2, 159.08) = 7.19$, $p = .001$. This interaction demonstrated a parallel pattern to the individual rating dimensions. Human-classification participants responded more negatively to androids (compared to orientation-classification participants), $b = -.43$, $SE = .16$, $t = -2.63$, $p = .01$, $d = .40$ (see Figure 3.6). Finally, we also observed a main effect of Agent, $F(2, 159.08) = 310.69$, $p < .001$, where the robot received lower ratings than both the android and human. This main effect

occurred in both conditions (*p*s < .001), showing that participants in both conditions paid attention to the differences between human and non-human agents.

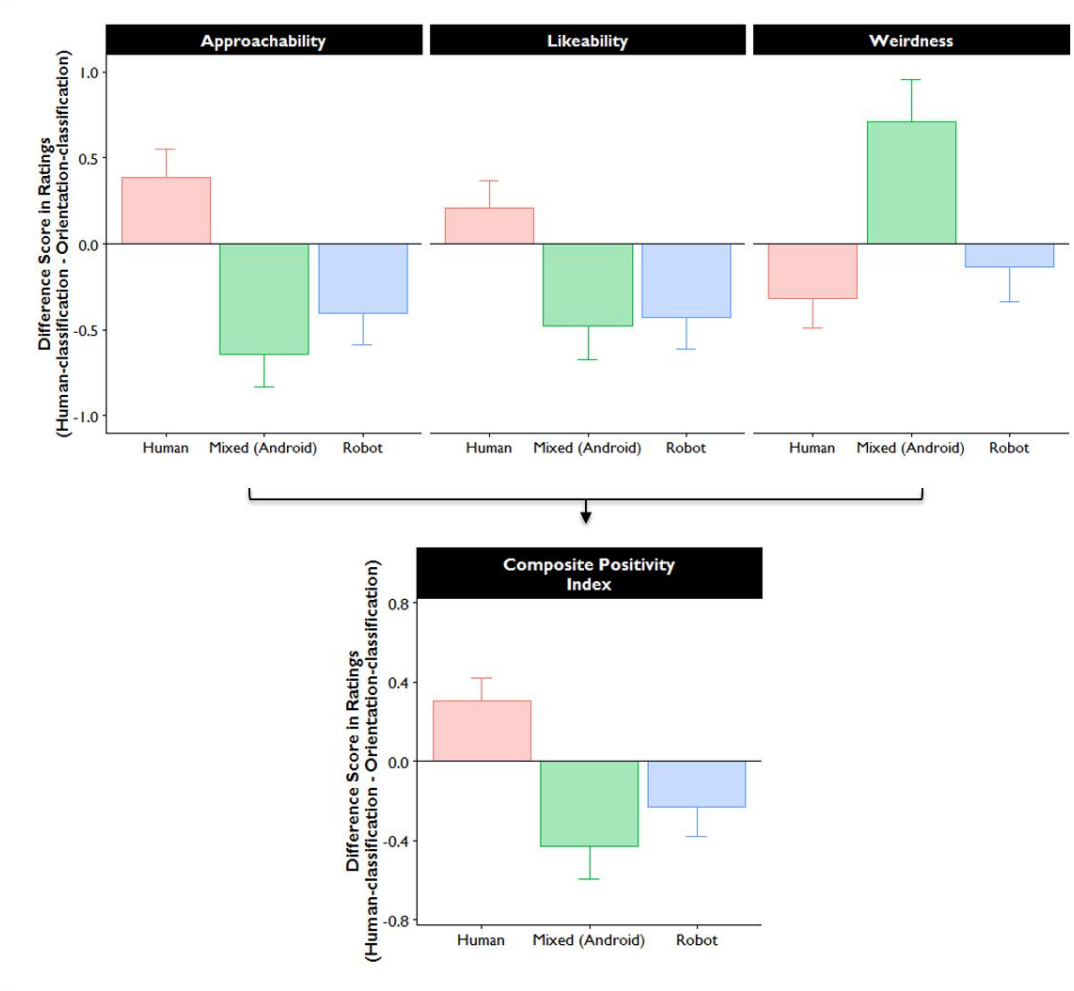


Figure 3.6: Difference scores by classification condition (human-classification – orientation-classification) on scale ratings for the different agent types (human, android, and robot; indicated by bar colors) in Experiment 2. Individual rating dimensions are shown (approachability, likeability, and weirdness), along with the composite positivity index (average of approachability, likeability, and reverse-coded weirdness scores). All graphs plot least squares means, along with standard errors.

Multilevel mediation across Experiments 1 and 2

To gauge whether fluency actually drives changes in weirdness evaluations (aside from merely accompanying them), we collapsed the categorization RTs (log_{10}-transformed) and

weirdness rating data from both Experiments 1 and 2 and applied multilevel mediation analyses for each condition (human-classification vs. orientation/no-classification), via the *mediation* package in R (R Core Team, 2014; Tingley et al., 2013). While the human-classification condition was exactly the same in both Experiments 1 and 2, note that the alternative conditions were different (i.e., in Experiment 1, no-classification participants only had a simple RT task, but in Experiment 2, this was changed to an orientation-classification task). For simplicity, these conditions were collapsed in the main mediation analyses, but separate analyses for each condition also did not yield any effects. Also, note that we only used weirdness ratings for mediation because Experiment 1 did not have any approachability or likeability ratings (compared to Experiment 2, which had all three dimensions).

Essentially, these methods allow for model-based estimation of the average total, direct, and indirect mediation effects using hierarchical data structures. Such a strategy is appropriate for repeated-measures designs to account for observations nested within subjects (Tingley et al., 2013). Our main predictor was agent "mixed-ness," which was dummy-coded as either 0 (not mixed [human and robot]) or 1 (mixed [android]). Our main DV was weirdness ratings, and our mediator was $\log_{10}$-categorization RT. To conduct the multilevel mediation analyses for Experiments 1 and 2, mixed-effects models were constructed for each of the mediation paths, using by-subject random effects parameters. All simulations from the *mediation* package in R were based on 5,000 samples per estimate, after which quasi-Bayesian confidence intervals were calculated around the average total, direct, and causal mediation effects.

Importantly, we observed evidence for mediation only within the human-classification subject group. Agent "mixed-ness" was a significant predictor of $\log_{10}$-RTs (*a*-path: $b = .05$, $SE = .01$, $t = 6.12$, $p < .001$), and $\log_{10}$-RTs were a significant predictor of weirdness ratings (*b*-path: $b = 3.21$, $SE = 1.02$, $t = 3.16$, $p = .002$). The total effect was significant (*c*-path: $b = 1.44$, $CI_{95\%} = [.99, 1.91]$, $p < .01$), along with the average direct effect of agent "mixed-ness" on weirdness

ratings (*c'*-path: $b = 1.31$, $CI_{95\%} = [.85, 1.80]$, $p < .01$). And critically, the average causal mediation effect was also significant ($b = .13$, $CI_{95\%} = [.03, .25]$, $p = .02$), demonstrating $\log_{10}$-categorization RTs as a mediator.

Note that when these same analyses were done for the orientation/no-classification subject groups, we observed no evidence of mediation. Agent "mixed-ness" did not predict $\log_{10}$-RTs (*a*-path: $b = .005$, $SE = .006$, $t = .85$, *ns*), and $\log_{10}$-RTs did not predict weirdness ratings (*b*-path: $b = -.07$, $SE = .41$, $t = -.18$, *ns*). The total effect was still significant (*c*-path: $b = .55$, $CI_{95\%} = [.14, .95]$, $p = .01$) as was the average direct effect (*c'*-path: $b = .55$, $CI_{95\%} = [.15, .96]$, $p = .01$), but there was no average causal mediation effect ($b < .001$, $CI_{95\%} = [-.01, .01]$, *ns*).

Experiment 3

We replicated the key findings from Experiment 1 with Experiment 2 (where androids were selectively devalued in the human-classification condition) and this generalized to all evaluative dimensions (approachability, likeability, and weirdness). Multilevel mediation analyses across data from both Experiments 1 and 2 revealed that categorization fluency ($\log_{10}$-RTs) mediated the effect between agent "mixed-ness" and weirdness ratings.

In Experiment 3, we wanted to investigate the specificity of the fluency-devaluation effects. One key question is whether similar devaluation effects emerge if androids are disfluent on a dimension that is *not* a key feature of the agent (an ambiguous dimension that is *not* human-likeness). This is theoretically important because it addresses a key theoretical (yet underexplored) distinction between disfluency that results from integral versus incidental ambiguity (Bodenhausen, 1993). If fluency-rating effects *do* emerge when the android is selectively disfluent but on incidental features (e.g., mixed color background cues, instead of human-likeness), this would argue that devaluation arises from general misattribution of task difficulty. If *not*, this would suggest that, in order to influence evaluations, the disfluency must

be meaningfully connected to the underlying nature of the target. We return to these theoretical alternatives later in the *Discussion*.

Method

Participants and stimuli. One hundred twenty-two UCSD undergraduates (M_{age} = 20.25 years, SD_{age} = 1.49 years; 91 females) participated for course credit and signed consent forms approved by the UCSD Human Research Protections Program.

Design and procedure. The design for Experiment 3 was similar to Experiment 2, but with two important (and related) changes.

First, we altered the image stimuli from Experiments 1 and 2 to be tinted across different shades of blue and green, but critically, in a way that mirrored the "blending" of the agents themselves. More specifically, the human and robot images were tinted as either 100% blue or 100% green respectively (image colors *not* mixed), while the android images were tinted to be exactly 50% blue // 50% green (image colors mixed). This was done to create a situation where the level of ambiguity (and thereby, classification difficulty) was still tied to each individual agent, but in a manner that did not specifically relate to the human-likeness dimension (see Figures 3.1 and 3.2).

Second, using these color-modified images, we changed the human-classification conditions from Experiments 1 and 2 to a *color*-classification condition in Experiment 3. In Experiment 3, color-classification participants were instead instructed to categorize each of the stimulus images on whether or not they were "blue or green." Note that with this setup, the fluency structure of the human-classification conditions from the previous experiments is still preserved (i.e., robots and humans are easier to categorize, while the android is made selectively difficult) but refers to an incidental dimension (i.e., whether the individual images are "blue or green").

Finally, after the experiment, we also debriefed each participant and asked for their opinions on what they thought the study was investigated. None of the participants mentioned anything related to categorization difficulty impacting their ratings, based on the color or agents in the stimuli.

One subject performed at ≤ 50% accuracy and was thus excluded from the total sample, leaving a final $n = 121$. All other parameters and analysis procedures remained the same as Experiments 1 and 2 (see Figure 3.2).

Results

Categorization RTs. We analyzed RTs using the same MLM methods as Experiments 1 and 2. Aside from a main effect of Agent, $F(2, 121.15) = 104.14$, $p < .001$, we found the predicted Condition × Agent interaction, $F(2, 121.15) = 92.77$, $p < .001$.

Color-classification participants took longer to categorize the android, both compared to the human agent, $b = .08$, $SE = .005$, $t = 15.80$, $p < .001$, $d_Z = 2.09$, and the robot agent, $b = .09$, $SE = .005$, $t = 17.03$, $p < .001$, $d_Z = 2.26$, but there was no difference between their human and robot RTs. Orientation-classification participants took less time to categorize the human agent, both compared to the android agent, $b = -.01$, $SE = .005$, $t = -2.34$, $p = .02$, $d_Z = .29$, and the robot agent, $b = -.02$, $SE = .005$, $t = -3.24$, $p < .01$, $d_Z = .41$, while showing no differences between android and robot RTs, $b < .01$, $SE = .005$, $t = .94$, *ns*, $d_Z = .12$ (see Figure 3.7).

In sum, the pattern of RTs for the color-classification condition was similar to that of the human-classification conditions in Experiments 1 and 2. In Experiment 3, android images were selectively disfluent (i.e., took longer to categorize) only in the color-classification condition, whereas there were no consistent RT differences for android images in the alternative orientation-classification condition.

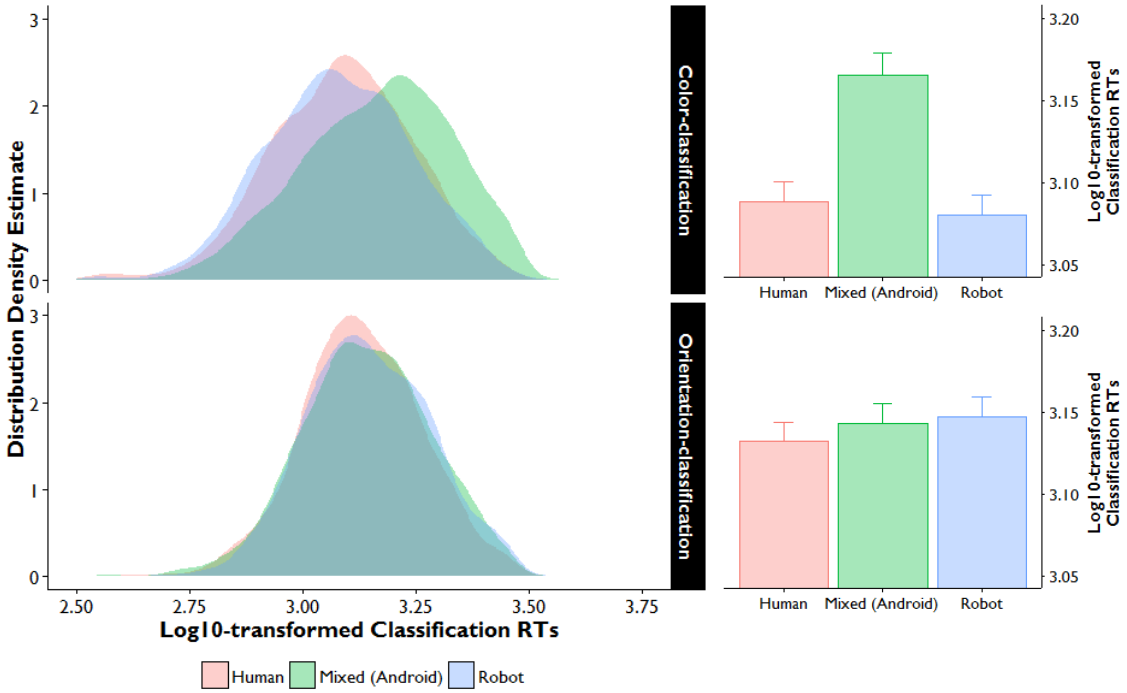


Figure 3.7: Density distributions and means/SEMs for $\log_{10}$-transformed RTs by classification condition (top row = color-classification; bottom-row = orientation-classification) and agent type (indicated by colors) for Experiment 3.

Scale ratings. We analyzed all scale ratings for Experiment 3 using the same MLM methods as Experiments 1 and 2.

Approachability. Intriguingly, with Experiment 3, we did *not* observe a Condition x Agent interaction, $F(2, 120.48) = 0.32$, *ns*. When the difficult human-classification condition was changed to a difficult *color*-classification in Experiment 3, the fluency effects on approachability ratings disappeared. Color-classification participants did not differ from the orientation-classification participants on approachability ratings for the android, $b = .19$, $SE = .19$, $t = 1.03$, *ns*, $d = .19$ (see Figure 3.8).

Also, as was the case with Experiments 1 and 2, we observed a strong main effect of Agent, $F(2, 120.48) = 105.58$, $p < .001$, such that participants rated the robot as less approachable

than the android, which was less approachable than the human (same as in Experiments 1 and 2). Once again, this main effect of Agent Type occurred in both the experimental and control conditions (*p*s < .001). This suggests that participants in both conditions of Experiment 3 noticed differences between human and non-human agents. Other ratings demonstrated this same pattern, as indicated below.

Likeability. The effects on the likeability ratings were very similar to those of the approachability dimension. Crucially, we also did *not* detect a Condition x Agent interaction on likeability ratings, $F(2, 120.87) = .36$, *ns*. Once again, the color-classification group did not differ from the orientation-classification group for likeability ratings on the android, $b = .16$, $SE = .18$, $t = .87$, *ns*, $d = .16$ (see Figure 3.8).

Note that once again, we also saw a main effect of Agent, $F(2, 120.87) = 120.79$, $p < .001$, where the robot was rated as less likeable than the android, which was rated less likeable than the human. This main effect was significant in both the experimental and control conditions (*p*s < .001).

Weirdness. With weirdness ratings, once again, we did *not* find a Condition x Agent interaction, $F(2, 120.50) = 1.14$, *ns*. The color-classification group did not differ from the orientation-classification group for weirdness ratings on the android, $b = .06$, $SE = .23$, $t = .28$, *ns*, $d = .03$ (see Figure 3.8).

Moreover, as with the other rating dimensions, we observed a similar main effect of Agent as Experiments 1 and 2, $F(2, 120.50) = 267.02$, $p < .001$, such that participants rated the robot as weirder than the android, which was rated as weirder than the human. This main effect was significant in both the experimental and control conditions (*ps* < .001).

Composite positivity index. As with Experiment 2, we constructed a composite positivity index by averaging approachability, likeability, and reverse-coded weirdness scores. Once again, we found a main effect of Agent, $F(2, 121.00) = 231.80$, $p < .001$, where the robot was rated

lower than both the android and human agents. This main effect occurred in both classification conditions (*p*s <.001). Crucially, we did not detect any evidence for a Condition x Agent interaction, $F(2, 121.00) = .74$, *ns*, suggesting that differences in agent ratings were not influenced by the categorization condition (color-classification versus orientation-classification; see Figure 3.8).

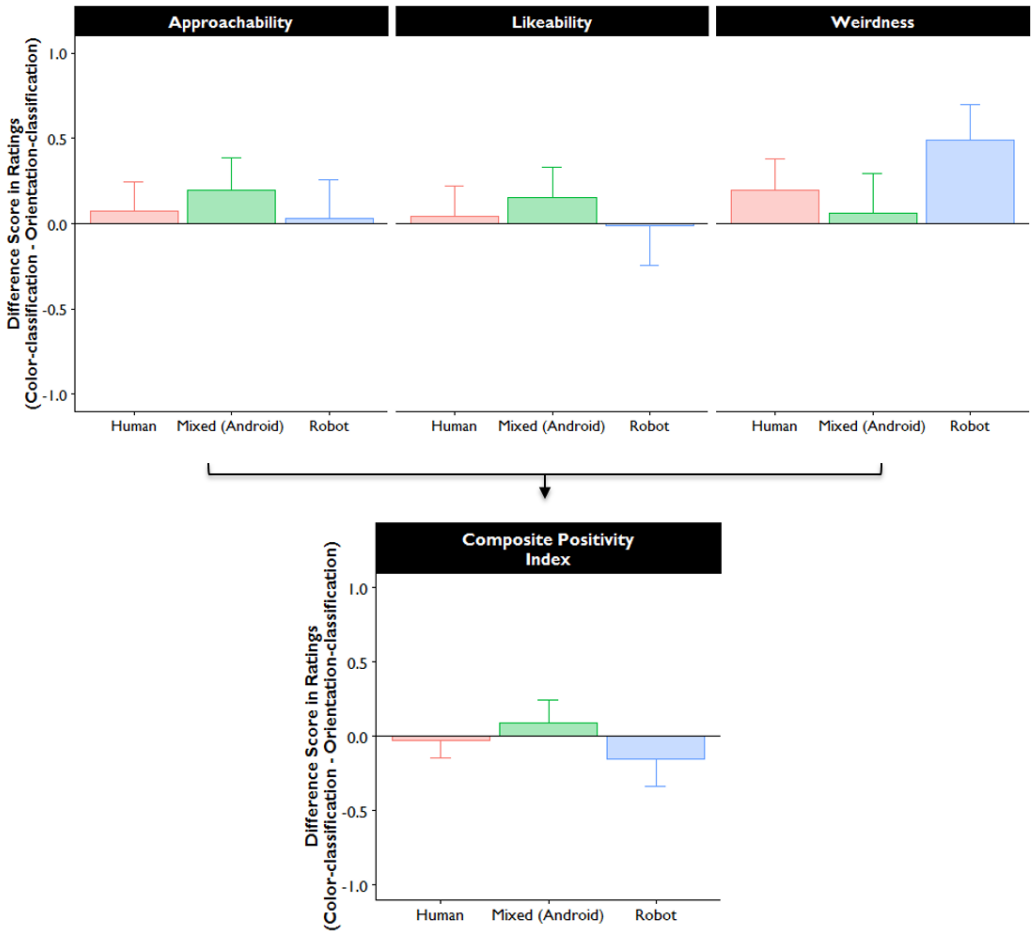


Figure 3.8: Difference scores by classification condition (color-classification – orientation-classification) on scale ratings for the different agent types (human, android, and robot; indicated by bar colors) in Experiment 3. Individual rating dimensions are shown (approachability, likeability, and weirdness), along with the composite positivity index (average of approachability, likeability, and reverse-coded weirdness scores). All graphs plot least squares means, along with standard errors.

Discussion

Our results suggest that negative evaluations of mixed agents can arise from the processing effort exerted to classify such agents on dimensions relevant to human features. Crucially, such disfluency and resulting devaluation occurred only when participants first categorized those agents along the human-likeness dimension on which they were ambiguous (Experiments 1 and 2). These effects did not occur when processing of mixed agents was measured using a generic stimulus detection RT task (Experiment 1) or when these agents were classified along a social orientation dimension on which they were *un*ambiguous (Experiment 2). These effects emerged even though participants devoted an overall comparable amount of time to processing the agents in the control and experimental conditions (Experiment 2). Consistent with this, mediation effects emerged only for participants in the human-classification condition (Experiments 1 and 2). These results cannot be a mere byproduct of general difficulty misattribution, since a color-classification task that made androids selectively disfluent did not yield similar patterns (Experiment 3). Note that our findings also cannot be explained by lack of attention to relevant agent features in the control condition. After all, in each classification condition across all experiments (including control conditions), participants were sensitive to relevant human/non-human features and adjusted their evaluative ratings accordingly.

These findings provide an important qualification to previous essentialist claims about the "unbridgeable" boundary between human and non-human entities. Recall that these essentialized properties are viewed as deep and immutable traits of the agent in-question. While the experience of being human is certainly one example, note that perceivers can also essentialize other social dimensions (e.g., gender, race, sexual orientation, etc.; Bastian & Haslam, 2006; Haslam, Rothschild, & Ernst, 2000; Haslam, Bastian, Bain, & Kashima, 2006; Howell, Weikum, & Dyck, 2011). Regardless of the specific dimension, essentialized categories carry with them a list of defining characteristics: clear and discrete boundaries from other categories, involuntary

and unchanging membership, and observable features that reflect something about the underlying function of the agent (Prentice & Miller, 2007).

On this essentialism view, the spontaneous negative responses to mixed agents arise due to a violation caused by blurring different "natural kinds" for human and non-human (see Demoulin, Leyens, & Yzerbyt, 2006; Medin & Ortony, 1989; Prentice & Miller, 2007). While theoretically distinct, other frameworks similarly posit that "mismatches" spontaneously yield negative responses to mixed agents. These mismatches can be perceptual, resulting in conflicting cues in visual, auditory, and motion processing (Katsyri et al., 2015; MacDorman, Green, Ho, & Koch, 2009; Mitchell et al., 2011; Seyama & Nagayama, 2007). They can also be more conceptual, as with incompatible cues for mind perception (Waytz, Gray, Epley, & Wegner, 2010). Our findings challenge views that push for strong automaticity in negative responses to mixed agents — instead, we show context-sensitivity and top-down control of these effects. This nicely corroborates recent work showing that negative responses can be modified by situational factors (Pollick, 2010), since both behavioral and neural responses can vary based on depth of processing (Cheetham et al., 2013) and subjectivity in judging human-likeness (Cheetham, Suter, & Jancke, 2011). Also, it is worth keeping in mind that the tendency towards essentializing categories is quite variable across different tasks and perceivers (Kalish, 2002). As Prentice & Miller (2007) state, "essentialism is not an all-or-none proposition, but rather is a matter of degree" (pg. 203). Therefore, our results more so argue against "strong" versions of the theory that do not allow for (*i*) dependence of responses to these categories on the specific task and context settings or (*ii*) flexibility in the construal of essential nature for human and non-human categories.

Critically, the current findings also offer an important theoretical extension of previous fluency models. Note that when substituting the human-classification condition for a *color*-classification task with the same "difficulty structure" (where 50/50 blue-green judgments made

android images selectively disfluent), all devaluation effects dissipated. This is a key point, since it highlights the importance of the human-likeness dimension, and it demonstrates that devaluation effects do not simply result from any general categorization difficulty. One explanation for these color-classification results appeals to the distinction between integral versus incidental cues (Bodenhausen, 1993). Devaluation effects may only follow from disfluency that occurs in response to an agent's key central features (i.e., *integral* human-likeness) rather than ancillary features with similar ambiguity (i.e., *incidental* colored backgrounds). More theoretically, we suggest that the evaluative consequences of (dis)fluency depend on the metacognitive processes that arise from monitoring that processing experience (see Schwarz, 2010, for a review). On one level, contextual variables can impact the processing experience itself, as with making information more or less difficult to process (similar to the different classification conditions in our experiments). But on another level, contextual variables can further impact how the metacognitive experience of (dis)fluency is interpreted and used — and this can dictate how later judgments are influenced. In our experiments, even though human-classification and color-classification led to similar experiences of disfluent processing, the interpretation of those metacognitive experiences is likely what drove differences in the evaluation of mixed agents. If one experiences disfluency on an integral feature of the agent (e.g., human-likeness), this will likely have downstream negative consequences on judgment. However, similar disfluency on an incidental feature (e.g., color background) would not be deemed relevant, and thus "gated" from the evaluation. Note that the integral versus incidental distinction is different than the idea that subjects discount blatant cues of cognitive disfluency in their ratings. In our studies, the color-based and human-based disfluency were equally strong and salient. The key difference here lies in participants' beliefs about the relevance of fluency cues to the particular judgment (here, evaluative rating). This also corresponds well to ideas of feelings-as-information, where an experience is used as a cue in making judgments, but only when the

experience is considered to be appropriate and relevant to the judgment at-hand (Schwarz & Clore, 2003).

More broadly, our work also relates to findings on inhibitory devaluation and stimulus-category competition (Fenske & Raymond, 2006; Raymond, 2009). These models argue that the discomfort with mixed agents is a more specific example of cognitive interference, which emerges from resolving multiple competing stimulus representations via inhibition (Ferrey, Burleigh and Fenske, 2015). This inhibition leads to negative evaluation, which has been shown with human faces and bodies (Fenske et al., 2005; Ferrey, Frischen, & Fenske, 2012) and non-human entities (Griffiths & Mitchell, 2008). Interestingly, some of these studies found that negative evaluation of stimuli on a categorical boundary can occur without explicit categorization (e.g., Ferrey et al. 2015). Our results differ from these findings. First, our experiments show strong task sensitivity — that is, in our experiments, the devaluation effects were more pronounced with categorization. Second, our Experiment 3 demonstrates that difficult color-categorization did not lead to devaluation of ambiguous images. One interpretation of this difference is that our stimuli are richer and more complex, thus leading participants' construal of those stimuli to be more dependent on categorical processes. In fact, some other work suggests that with highly similar and familiar stimuli, categorization conflict may spontaneously "pop out" without any categorization task, leading to devaluation (Halberstadt, Pecher, Zeelenberg, Wai, & Winkielman, 2013). Further, notice here that our fluency perspective is theoretically distinct from models of inhibition. While fluency theories focus on category processing effort, inhibition theories focus on resolving cognitive conflict (often by attaching "inhibitory" tags to distractors). These theoretical perspectives should be investigated further, along with other frameworks linking emotion and categorization to judgments of human and non-human agents (e.g., Burleigh and Schoenherr, 2014; Cheetham et al., 2011; Cheetham et al., 2013).

Limitations and future directions

There are also some important limitations to consider for the current experiments. First, we used longer RTs as our main operationalization of disfluency (i.e., greater RTs indicates greater processing difficulty). Note, however, that while processing difficulty would certainly yield longer RTs, longer RTs do not necessarily have to index disfluency (e.g., longer RTs could also emerge from reduced motivation, increased curiosity towards the stimulus, etc.). Thus, in future studies, alternative fluency measures to RTs should also be incorporated to extend our findings.

Second, while our stimuli were highly controlled images of human, android, and robot agents doing different actions (Saygin & Stadler, 2012), our stimulus set only contained one specific example for each agent type. While our experiments were not designed to investigate subtle gradations in human-likeness between human and robot, our android agents were likely not exactly in the middle of this continuum (the perception of which also probably varies across participants). Therefore, future research may also want to include multiple exemplars of human, android, and robot agents, which may be able to offer more precise degrees of human-likeness (e.g., human/non-human morphs; Mathur & Reichling, 2016; Powers, Worsham, Freeman, Wheatley, & Heatherton, 2014).

Third, the directionality of our fluency-devaluation effect on mixed agents remains unclear. More specifically, explicit categorization on the human-likeness dimension (as with Experiments 1 and 2) may amplify disfluency and negative attitudes for mixed agents. Another possibility is that perceivers rapidly and implicitly categorize the agents on the human-likeness dimension as the "default," and forced categorization on an alternative dimension (e.g., orientation-classification) then *reduces* negative attitudes for mixed agents. Our results from Experiment 1 might suggest that perceivers do not spontaneously categorize the agents on human-likeness, since human-classification participants judged the android to be selectively

weirder than participants with only a detection RT task (see Figure 3.4). However, note that participants had much faster RTs without categorization (around 500-600 ms) compared to those in the human-classification condition (around 1200-1400 ms; see Figure 3.3). Implicit categorization on the human-likeness dimension may take longer to emerge, or the no-classification task in Experiment 1 might have distracted participants enough such that spontaneous human-likeness categorization could not occur. This would be interesting to examine in future studies, considering previous papers reporting "spontaneous" negative responses to mixed agents (Mitchell et al., 2011; Tinwell, Grimshaw, Nabi, & Williams, 2011; Złotowski et al., 2015). It may be possible that spontaneous "pop-out" effects from mixed agents are at least partially due to implicit and disfluent categorization on the central human-likeness dimension (also see Burleigh and Schoenherr, 2014; Cheetham et al., 2011; Cheetham et al., 2013; Ito & Cacioppo, 2000; Wiese, Schweinberger, & Neumann, 2008).

Note that our results do not rule out bottom-up effects in processing mixed agents. This seems evident from our RT results in Experiment 1 (see Figure 3.3) and Experiment 2 (see Figure 3.5), where participants still took longer to respond to androids even during the alternative control tasks (albeit the differences were much smaller than human-classification conditions). Also, keep in mind that some versions of bottom-up perceptual theories can be considered compatible with fluency frameworks (as when incompatibility is detected early and leads to obligatory difficulty in feature processing).

Finally, some theories hold that disfluency itself is not the key driver of negative evaluation, but rather it is the implications of such disfluency. As examples, disfluency or inconsistency might only matter if they signal a prediction error (Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012), a gap in knowledge (Kruglanski, 2013; Roets et al., 2015; Viola et al., 2015), or a collapse in the sense of meaning (Proulx & Inzlicht, 2012). Therefore, we argue that

negative responses to mixed agents involve an *interaction* between bottom-up perceptual factors and top-down categorization processes.

Conclusion

In sum, the current findings highlight the broader theoretical point that higher-order processes can modify how we evaluate non-human agents. These influences, which we explored here with "hot" evaluative judgments, are also likely to bear on "cold" cognitive assessments of mind perception, agency, and intentionality (Gray & Wegner, 2012; Waytz, Gray, Epley, & Wegner, 2010). Essentially, categorization (dis)fluency can modify the impact of this fundamental boundary between human and non-human, where we can dehumanize the living (Haslam, 2006) or anthropomorphize the artificial (Chandler & Schwarz, 2010; Epley, Waytz, & Cacioppo, 2007).

Chapter 3 is, in full, in press for publication in *Journal of Experimental Psychology: Human Perception and Performance*. Carr, Evan W.; Hofree, Galit; Sheldon, Kayla; Saygin, Ayse P.; Winkielman, Piotr. The dissertation author was the primary investigator and author of this material.

References

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology, 37*(1), 13-20.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology, 42*(2), 228-235.

Bastian, B., & Haslam, N. (2007). Psychological essentialism and attention allocation: Preferences for stereotype-consistent versus stereotype-inconsistent information. *Journal of Social Psychology, 147*(5), 531-541.

Bastian, B., Loughnan, S., & Koval, P. (2011). Essentialist beliefs predict automatic motor responses to social categories. *Group Processes & Intergroup Relations, 14*(4), 559-567.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*.

Bodenhausen, G. V. (1993). Emotions, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping* (pp. 13-37). San Diego, CA: Academic Press.

Burleigh, T. J., & Schoenherr, J. R. (2014). A reappraisal of the uncanny valley: Categorical perception or frequency-based sensitization? *Frontiers in Psychology, 5*.

Carr, E.W., Rotteveel, M., & Winkielman, P. (2016). Easy moves: Perceptual fluency facilitates approach-related action. *Emotion, 16*(4), 540-552.

Chandler, J., & Schwarz, N. (2010). Use does not wear ragged the fabric of friendship: Thinking of objects as alive makes people less willing to replace them. *Journal of Consumer Psychology, 20*(2), 138-145.

Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": Behavioral and functional MRI findings. *Frontiers in Human Neuroscience, 5.*

Cheetham, M., Pavlovic, I., Jordan, N., Suter, P., & Jancke, L. (2013). Category processing and the human likeness dimension of the uncanny valley hypothesis: Eye-tracking data. *Frontiers in Psychology, 4.*

Demoulin, S., Leyens, J. P., & Yzerbyt, V. (2006). Lay theories of essentialism. *Group Processes & Intergroup Relations, 9*(1), 25-42.

Dennett, D. C., (1971). Intentional Systems. *The Journal of Philosophy*, *68*, 87-106.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864.

Fenske, M. J., & Raymond, J. E. (2006). Affective influences of selective attention. *Current Directions in Psychological Science*, *15*(6), 312-316.

Fenske, M. J., Raymond, J. E., Kessler, K., Westoby, N., & Tipper, S. P. (2005). Attentional inhibition has social-emotional consequences for unfamiliar faces. *Psychological Science*, *16*(10), 753-758.

Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: A novel account of the uncanny valley. *Frontiers in Psychology*, *6*.

Ferrey, A. E., Frischen, A., & Fenske, M. J. (2012). Hot or not: Response inhibition reduces the hedonic value and motivational incentive of sexual stimuli. *Frontiers in Psychology*, *3*, 575.

Frenkel-Brunswik, E. (1949). Intolerance of ambiguity as an emotional and perceptual personality variable. *Journal of Personality*, *18*, 108–143.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404-409.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125-130.

Griffiths, O., & Mitchell, C. J. (2008). Negative priming reduces affective ratings. *Cognition and Emotion*, *22*(6), 1119-1129.

Halberstadt, J., Pecher, D., Zeelenberg, R., Wai, L.I., & Winkielman, P. (2013). Two faces of attractiveness: Making beauty-in-averageness appear and reverse. *Psychological Science, 24*, 2343-2346.

Halberstadt, J. & Winkielman, P. (2013). When good blends go bad: How fluency can explain when we like and dislike ambiguity. In C. Unkelbach & R. Greisfelder. *The experience of thinking: How feelings from mental processes influence cognition and behavior* (pp. 133–151). New York, NY: Psychology Press.

Halberstadt, J. & Winkielman, P. (2014). Easy on the eyes, or hard to categorize: Classification difficulty decreases the appeal of facial blends. *Journal of Experimental Social Psychology, 50, 175–183.*

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252-264.

Haslam, N., Bastian, B., Bain, P., & Kashima, Y. (2006). Psychological essentialism, implicit theories, and intergroup relations. *Group Processes & Intergroup Relations*, *9*(1), 63-76.

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology, 39*(1), 113-127.

Ho, C. C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior, 26*(6), 1508-1518.

Howell, A. J., Weikum, B. A., & Dyck, H. L. (2011). Psychological essentialism and its association with stigmatization. *Personality and Individual Differences, 50*(1), 95-100.

Ishiguro, H. (2006). Android science: Conscious and subconscious recognition. *Connection Science, 18*(4), 319–332.

Ishiguro, H. (2007). Android science. In *Robotics Research* (pp. 118-127). Springer Berlin Heidelberg.

Ito, T. A., & Cacioppo, J. T. (2000). Electrophysiological evidence of implicit and explicit categorization processes. *Journal of Experimental Social Psychology, 36*(6), 660-676.

Kalish, C.W. (2002). Essentialist to some degree: Beliefs about the structure of natural kind categories. *Memory & Cognition, 30,* 340–352.

Kruglanski, A. W. (2013). *Lay epistemics and human knowledge: Cognitive and motivational bases*. Springer Science & Business Media.

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

Li, A. X., Florendo, M., Miller, L. E., Ishiguro, H., & Saygin, A. P. (2015, March). Robot Form and Motion Influences Social Attention. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 43-50). ACM.

MacDorman, K. F., Green, R. D., Ho, C. C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3), 695-710.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*(3), 297-337.

Medin, D.L., & Ortony, A. (1989). Psychological essentialism. In S. Vosnaidou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge, UK: Cambridge University Press.

Mitchell, W. J., Szerszen Sr, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception, 2*(1), 10.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *Robotics & Automation Magazine, IEEE, 19*(2), 98-100.

Owen, H. E., Halberstadt, J., Carr, E. W., & Winkielman, P. (2016). Johnny Depp, reconsidered: How category-relative processing fluency determines the appeal of gender ambiguity. *PLoS ONE, 11*(2), e0146328.

Pollick, F. E. (2010). In search of the uncanny valley. In *User centric media* (pp. 69-78). Springer Berlin Heidelberg.

Powers, K. E., Worsham, A. L., Freeman, J. B., Wheatley, T., & Heatherton, T. F. (2014). Social connection modulates perceptions of animacy. *Psychological Science, 25*(10), 1943-1948.

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. *Current Directions in Psychological Science, 16*(4), 202-206.

Proulx, T., & Inzlicht, M. (2012). The five "A" s of meaning maintenance: Finding meaning in the theories of sense-making. *Psychological Inquiry*, *23*(4), 317-335.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria < http://www.R-project.org/ >.

Rajaram, S., & Geraci, L. (2000). Conceptual fluency selectively influences knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 1070.

Raymond, J. (2009). Interactions of attention, emotion and motivation. *Progress in Brain Research*, *176*, 293-308.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*, 338-342.

Reber, R., Winkielman P., & Schwarz N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science, 9,* 45-48.

Roets, A., Kruglanski, A. W., Kossowska, M., Pierro, A., & Hong, Y. Y. (2015). Chapter four: The motivated gatekeeper of our minds: New directions in need for closure theory and research. *Advances in Experimental Social Psychology, 52*, 221-283.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, *7*(4), 413-422.

Saygin, A. P., & Stadler, W. (2012). The role of appearance and motion in action prediction. *Psychological Research*, *76*(4), 388-394.

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review, 2*, 87–99.

Schwarz, N. (2010). Meaning in context: Metacognitive experiences. In B. Mesquita, L. F. Barrett, & E. R. Smith (eds.), *The mind in context* (pp. 105-125). New York, NY: Guilford Press.

Schwarz, N., & Clore, G. L. (2003). Mood as information: 20 years later. *Psychological Inquiry*, *14*(3-4), 296-303.

Seyama, J. I., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence*, *16*(4), 337-351.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2013). mediation: R package for causal mediation analysis. R package version 4.4.

Tinwell, A., Grimshaw, M., Nabi, D. A., & Williams, A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior*, *27*(2), 741-749.

Viola, V., Tosoni, A., Brizi, A., Salvato, I., Kruglanski, A. W., Galati, G., & Mannetti, L. (2015). Need for cognitive closure modulates how perceptual decisions are affected by task difficulty and outcome relevance. *PLoS One*, *10*(12), e0146002.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383-388.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide using statistical software*. CRC Press.

Wiese, H., Schweinberger, S. R., & Neumann, M. F. (2008). Perceiving age and gender in unfamiliar faces: Brain potential evidence for implicit and explicit person categorization. *Psychophysiology*, *45*(6), 957-969.

Whittlesea, B. W. A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1235-1253.

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation increases positive affect. *Journal of Personality and Social Psychology*, *81*(6), 989-1000.

Winkielman, P., Halberstadt, J., Fazendeiro, T. & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17*. 799-806.

Winkielman, P., Olszanowski, M., & Gola. M. (2015). Faces in between: Evaluative responses to faces reflect the interplay of features and task-dependent fluency. *Emotion*, *15*, 232-242.

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.

Złotowski, J. A., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2015). Persistence of the uncanny valley: The influence of repeated interactions and a robot's attitude on its perception. *Frontiers in Psychology, 6*.

CHAPTER 4:

Easy moves: Perceptual fluency facilitates approach-related action

Evan W. Carr, Mark Rotteveel, & Piotr Winkielman

As it appears in

Emotion

2016

Volume 16, Issue 4, Pages 540-552

Digital Object Identifier (DOI): http://dx.doi.org/10.1037/emo0000146

Emotion

1528-3542/16/$12.00 http://dx.doi.org/10.1037/emo0000146

Easy Moves: Perceptual Fluency Facilitates Approach-Related Action

Evan W. Carr
University of California, San Diego

Mark Rotteveel
University of Amsterdam and Amsterdam Brain and Cognition (ABC) Center, Amsterdam, the Netherlands

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

It is well established that processing fluency impacts preference judgments and physiological reactions indicative of affect. Yet, little is known about how fluency influences motivation-related action. Here, we offer a novel demonstration that fluency facilitates action-tendencies related to approach. Four experiments investigated this action effect, its boundary conditions, and concomitant affective responses. Experiment 1 found faster approach movements (reaction times [RTs] to initiate arm flexion) to perceptually fluent stimuli when participants acted to rapidly classify stimuli as either "good" or "bad." Experiment 2 eliminated this fluency effect on action when participants performed nonaffective classifications ("living" or "nonliving"), even though fluency robustly enhanced liking judgments. Experiment 3 demonstrated that fluency can also facilitate approach action that is not immediate, as long as the delayed action involves affective classification. This experiment also found that fluent stimuli elicit genuine hedonic responses, as reflected in facial electromyography (fEMG) activity over zygomaticus "smiling" muscle. Experiment 4 replicated the physiological (fEMG) evidence for hedonic responses to fluent stimuli, but similar to Experiment 2, we observed no fluency effects on actions involving nonaffective classification. The current studies offer the first evidence that perceptual fluency can facilitate approach-related movements, when such movements are embedded in the context of affective decisions. Generally, these results suggest that variations in processing dynamics can flexibly and implicitly shape action-tendencies.

Keywords: emotion, action-tendencies, cognitive processes, facial expressions, electromyography

What determines whether we smile or frown to others, approach or avoid different objects, or judge people positively or negatively? These are classic questions about the links between emotion, cognition, motivation, and action (Frijda, 1986; Zajonc, 1998) but also about the connections between attitudes and behaviors (Allport, 1935; Petty, Fazio, & Briñol, 2009). This article addresses these questions by exploring a novel relationship between evaluation, action, and a core property of cognition—*fluency*, or its processing dynamics. Essentially, we propose that fluency gives neutral stimuli the ability to facilitate approach-related action, as elaborated next. Before that, we provide some background on emotion and motivation, and then discuss how fluency interacts with these processes in shaping action.

Starting with Darwin's (1872/1997) observation that emotions are tied to motivational states and motor tendencies, researchers have been trying to characterize these connections (Krieglmeyer, De Houwer, & Deutsch, 2013; Neumann, Förster, & Strack, 2003). Both classic and modern theories posit that affective and motivational processes are organized on an *approach-avoidance dimension* (Harmon-Jones, Harmon-Jones, & Price, 2013; Lang, Bradley, & Cuthbert, 1990). It is often suggested that these processes are linked to specific action-tendencies. For example, individuals are generally faster to pull a lever toward the body for positive stimuli and/or push a lever away from the body for negative stimuli (Krieglmeyer, Deutsch, De Houwer, & De Raedt, 2010). It is important that these approach-avoidance tendencies could be represented according to body-centered direction (toward vs. away; Chen & Bargh, 1999), specific muscle activation (bicep flexion vs. tricep extension; Cacioppo, Priester, & Berntson, 1993), object-centered directions (closer vs. farther spatial distance; Seibt, Neumann, Nussinson, & Strack, 2008), or self-relative

Evan W. Carr, Departments of Psychology and Cognitive Science, University of California, San Diego; Mark Rotteveel, Department of Psychology, University of Amsterdam, and Amsterdam Brain and Cognition (ABC) Center, Amsterdam, the Netherlands; Piotr Winkielman. Department of Psychology, University of California, San Diego, Behavioural Science Group, Warwick Business School, University of Warwick, and Department of Psychology, University of Social Sciences and Humanities.

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.

We would like to acknowledge very helpful comments from Dave Barner, Christopher Oveis, Liam Kavanagh, Galit Hofree, and Andy Arnold. Also, we are very grateful to the many research assistants who helped with these experiments, including Cassie Richards, Tan Nguyen, Ayesha Saxena, Meghan Jeffrey, Anders Tse, Timothy Yu, Samuel Wood, Nathan Olesen, Alexandria Bell, Roma Shah, Lorelei Himlin, Kevin Wright, Tom Karaffa, Alex Gray, Jerin Tan, Kayla Sheldon, Raseana Williams, Jason Herbert, Valerie Dapsis, and Melanie McLaughlin.

Correspondence concerning this article should be addressed to Evan W. Carr, Department of Psychology, University of California, San Diego, 9500 Gilman Drive 0109, La Jolla, CA 92093. E-mail: Ewcarr@ucsd.edu

movements (Markman & Brendl, 2005; but see Van Dantzig, Zeelenberg, & Pecher, 2009).

In recent years, new insights into the links between perception, cognition, emotion, and action came from examining *fluency*—or changes in processing speed and effort (Schwarz, 1998). Extensive evidence suggests that higher perceptual and conceptual fluency (usually manipulated through stimulus repetition, priming, clarity, contrast, duration, or typicality) enhances affective responses and evaluative judgments, ranging from basic preferences (Winkielman & Cacioppo, 2001), consumer choices (Novemsky, Dhar, Schwarz, & Simonson, 2007), assessments of attractiveness (Winkielman, Schwarz, Fazendeiro, & Reber, 2003), brand evaluations (Lee & Labroo, 2004), trustworthiness ratings (Winkielman, Olszanowski, & Gola, 2015), to purchases of novel stocks (Alter & Oppenheimer, 2006). According to the *hedonic fluency model*, these effects occur because easy processing elicits mild positive affect, which is then (mis)attributed to the target stimulus (Winkielman, Schwarz, Fazendeiro, & Reber, 2003). This positive affect presumably emerges because fluency reflects (or probabilistically indicates) lower conflict and cost in processing, greater coherence, as well as higher stimulus familiarity. Note that all this may occur early in stimulus processing, coloring the initial impression of a fluent stimulus with positivity.

Given these findings and theoretical assumptions, it is surprising that fluency's motivational consequences have thus far received limited attention. There are indications that mere exposure (which increases fluency, though also familiarity) augments neural electroencephalogram (EEG) indices of approach motivation (Harmon-Jones & Allen, 2001). However, no study has yet tested whether fluency influences motivationally related *action.* This is especially surprising given the availability and popularity of various methods to study action-tendencies associated with approach and avoidance, as discussed shortly.

There are several important reasons for exploring fluency's consequences on motivation-relevant action. First, it is not obvious that fluency should impact approach-avoidance action. After all, fluency effects on preference judgments and affective reactions are typically found with mild, inherently neutral stimuli. In contrast, effects on motivation-related action are typically found with strongly valenced stimuli (i.e., highly emotion-laden pictures, faces, or words). This raises a question of whether any fluency effects on valence of initially neutral stimuli are strong, enduring, or pervasive enough to impact more "basic" action. This question relates to the debates about the limits of weak (as opposed to strong) evaluative objects to spontaneously trigger action-related processes (e.g., Fazio, 2001). Second (and more generally), note that there is no obligatory or straightforward connection between affect and motivated-action. For instance, factors that increase stimulus liking do not always increase stimulus wanting (Aharon et al., 2001; Litt, Khan, & Shiv, 2010; Winkielman & Berridge, 2003). In another example, factors that decrease stimulus evaluation can sometimes increase approach behavior, as is the case with anger (Harmon-Jones, Harmon-Jones, & Price, 2013). Finally, if there are any fluency effects on action, they are probably subject to important boundary conditions. One key issue here is their possible automaticity. Fluency effects on approach-related action could be relatively unconditional, as has been previously argued for inherently valenced stimuli (e.g., Chen & Bargh, 1999). However, recent reviews suggest that approach-avoidance action effects even to strongly valenced stimuli are robustly observed *only* when such action is embedded in an affective classification tasks (see Phaf, Mohr, Rotteveel, & Wicherts, 2014, for a meta-analysis). As an example, positive valence facilitates approach action when participants classify emotional faces into affective categories, such as "positive" or "negative," but not into nonaffective categories, such as "male" or "female" (Rotteveel et al., 2015; Rotteveel & Phaf, 2004).

To address these questions, four experiments investigated how fluency impacts affect and approach-avoidance actions. The general logic of all the studies was to present novel and neutral stimuli (pseudowords) in the context of a "word judgment" task. Fluency was manipulated using a standard procedure of varying font readability. We assessed hedonic effects of fluency with self-report measures of liking and physiological reactivity—facial electromyography (fEMG). Critically, we also gauged fluency's consequences on motivated action using a method that taps into the perceiver's readiness to perform approach- and avoidance-related movements, as explained next.

There are important debates about the relative strengths of various approach-avoidance methods and paradigms (see Krieglmeyer & Deutsch, 2010). Here, we chose a robust method for which the available data suggest that approach motivation facilitates flexion movements (see Phaf, Mohr, Rotteveel, & Wicherts, 2014). However, as elaborated in the discussion, our hypothesis is committed to the fluency-approach link, rather than fluency-flexion link, with this particular paradigm providing a good way of measuring approach with flexion. More specifically, we used a vertical button-stand where participants pressed either a top or bottom response-button, which map onto approach (bicep contraction—resulting in arm flexion) and avoidance (tricep contraction—resulting in arm extension) movements, respectively (Figure 1; Rotteveel & Phaf, 2004). Note that this device keeps the spatial distance of each response type consistent, while providing two different dependent measures—*release time* (*RelT*, or time to initiate a response) and

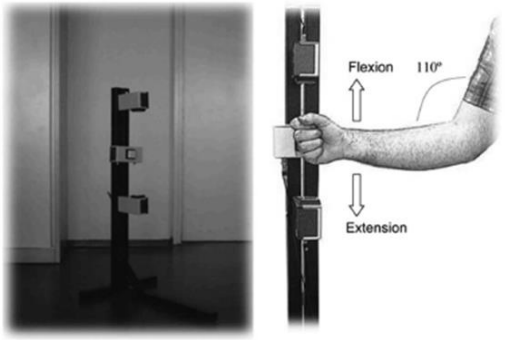


Figure 1. Experimental apparatus used for all four experiments in the approach-avoidance task (AAT). Three button boxes are affixed vertically on a metal stand. The middle box is the "home" button, where participants would rest the back of their hand and wait to respond. The two boxes above and below serve as response buttons, where participants would either flex or extend their arm to make a decision, respectively.

movement time (*MovT*, or time to actually move the arm). Affective influences are typically found in RelT and not MovT (Phaf & Rotteveel, 2009; Rotteveel & Phaf, 2004), perhaps because emotion is about action-preparation, rather than action-performance (Frijda, 2010). Given this, we will not focus on exploratory MovT analyses here, because we had no specific predictions (but see footnotes for each individual experiment).

We hypothesized that fluency (compared with disfluency) would elicit faster flexion responses, indicating approach (Experiments 1 and 3). Note that flexion responses may be particularly sensitive to processing facilitation, given earlier research suggesting that fluency manipulations tend to influence positive rather than negative affect (as discussed later). Furthermore, as mentioned, fluency-action effects should be limited to when movements are embedded in affective classification tasks (whether the stimulus is "good" or "bad"; Experiments 1 and 3) rather than *non*-affective classifications (whether stimulus is "living" or "nonliving"; Experiments 2 and 4).

Finally, in all four experiments, we measured affective consequences of fluency. We predicted that fluency will lead to higher reports of liking for the stimulus. We also predicted that high fluency would elicit physiologically detectable positive response (increased smiling activity and reduced frowning activity, via fEMG over the *zygomaticus major* and *corrugator supercilii*, respectively), as measured in Experiments 3 and 4.

Experiment 1

Experiment 1 had two goals. First, we wanted to test how fluency influences rapid flexion movements (via reaction times [RTs], using the vertical button-stand; Figure 1). Second, we wanted to assess how fluency influences overt evaluations (liking ratings and classification decisions).

Method

Participants. Thirty-two University of California, San Diego (UCSD) undergraduates participated for course credit. All participants were right-handed English speakers.

Materials and apparatus. Targets were 100 neutral pseudowords (i.e., pronounceable strings of letters that appear to be real words but have no actual meaning; all between 5 and 7 letters). These stimuli were selected out of a set of 200 pseudowords, generated using the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). To select our neutral targets, we conducted an online survey using a separate sample of 72 UCSD undergraduates, where they rated all 200 pseudowords using a 1 to 7 scale on both valence (1 = *very negative* to 7 = *very positive*) and arousal (1 = *not arousing at all* to 7 = *very arousing*). Results showed that all 200 pseudowords were rated as relatively neutral in valence ($M_{valence}$ = 3.71, $SD_{valence}$ = 0.27) and low to medium in arousal ($M_{arousal}$ = 2.83, $SD_{arousal}$ = 0.22). For the final targets, we selected the 100 most neutral pseudowords on both scales ($M_{valence}$ = 3.92, $SD_{valence}$ = 0.17; $M_{arousal}$ = 2.87, $SD_{arousal}$ = 0.23).

Participants were instructed to respond using a vertical button-stand (for this setup and procedure, see Figure 1 and Rotteveel & Phaf, 2004). All participants were right-handed and sat to the left of the button-stand. To trigger the start of a trial, participants pressed and held the "home" button (fixed in the middle of the tower, 10 cm between the top and bottom buttons—adjusted to each individual participant to maintain the 110° angle between the upper and lower arm) with the back of their right hand, while they were waiting to respond. As they pressed one of two response buttons with the top or bottom side of their hand, they did not turn their hand when responding (Figure 1) and either flexed or extended their arm (revealing their "approach" vs. "avoidance" responses, respectively).

All stimuli were presented on a 17-inch Dell flat-screen from a PC running Windows XP and E-Prime 2.0.

Design and procedure. The experiment was introduced as "a language study where [participants would make] timed judgments of different words." After eight practice trials, participants were told to "make fast, intuitive judgments of whether the different words [were] *good* or *bad*—even though they [were not] in English, and [they would] not know their true meaning" (recall that the "non-English" pseudowords had no actual meaning and were selected as being neutral on both valence and arousal).

Each participant progressed through four blocks of 25 randomized trials, totaling 100 trials in the approach-avoidance task (AAT). As shown in Figure 2, each trial began with a 3,000-ms fixation, followed by the 300-ms presentation of a pseudoword in either fluent text (easy-to-read, black Arial font that was bolded) or disfluent text (difficult-to-read, silver Script font that was italicized). After 300 ms of target presentation, a decision screen appeared that said "GOOD -or- BAD," where these labels were paired with either the top or bottom response buttons (response labels were randomized across trials). Participants would release the middle "home" button to hit either the top or bottom button (thereby flexing or extending their arm). After logging the AAT response, participants would then would rate how much they liked the pseudoword on a 1 (*not at all*) to 4 (*very much*) scale, using their left hand on the keyboard (Figure 2).

Note that previous research with this paradigm has used facial expressions as stimuli, where the buttons can be labeled as "positive" versus "negative" or "male" versus "female"—thus allowing the responses to be evaluated on accuracy (e.g., Rotteveel & Phaf, 2004). However, in the current experiments, there was no explicit instruction regarding the association between high or low perceptual fluency and the specific arm movements. Participants responded only according to the instruction regarding their own affective evaluation of the targets on each trial, pressing the upper or lower buttons (i.e., arm flexion and arm extension). As a result, responses could not be evaluated according to accuracy (because no response could be regarded as incorrect).

Throughout the experiment, no explicit references were made to anything related to approach-avoidance movements (the response device was simply referred to as a "button-stand"), arm flexion versus arm extension, or positive versus negative emotion. After the session concluded, each participant was given an exit interview and asked for a brief statement on what they thought the experiment was investigating. None of the participants for any of the experiments reported anything linking fluency (font readability) to the type of arm movement (flexion vs. extension).

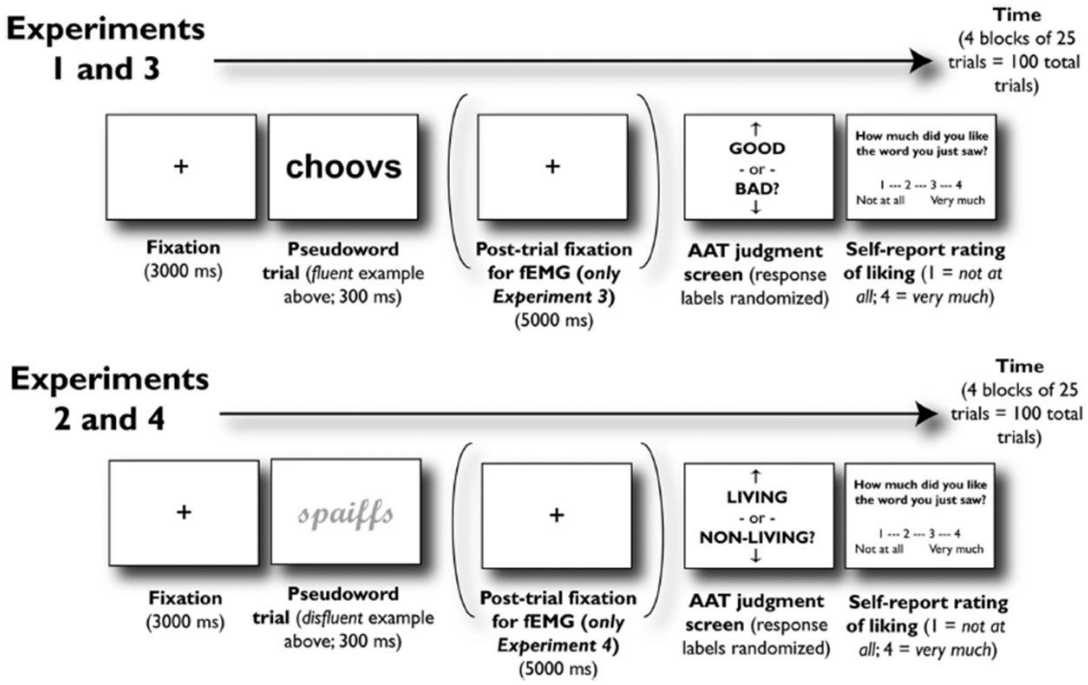


Figure 2. Design and procedure used for Experiments 1, 2, 3, and 4. AAT = approach-avoidance task.

Results

Analysis strategy. All repeated-measures analyses used mixed-effects modeling via maximum likelihood, because this method offers numerous analytical advantages—including more effective handling of unbalanced data with missing observations, reliance on fewer assumptions regarding covariance structures, and increased parsimony and flexibility between models (Bagiella, Sloan, & Heitjan, 2000). All models were built with the *lme4* (Bates, Maechler, Bolker, & Walker, 2015) and *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2014) packages in R (R Core Team, 2014), using a maximal random-effects structure appropriate for the data (Barr, Levy, Scheepers, & Tily, 2013). To obtain *p* value estimates for fixed-effects, we used Type III Satterthwaite approximations, which can sometimes result in decimal degrees of freedom, based on the number of observations (West, Welch, & Galecki, 2014).

RelTs. RTs were analyzed with a Fluency (2: fluent, disfluent) × Movement (2: flexion, extension) fixed-effects structure. To normalize the RelT distribution and reduce the impact of outliers, we removed all RelTs greater or less than 3 SDs from each subject's total average RelT (which constituted 3.44% of all trials), and all remaining RelTs were $\log_{10}$-transformed.

Recall that affective influences are typically found in RelT and not MovT (Phaf & Rotteveel, 2009; Rotteveel & Phaf, 2004), so we will not focus on exploratory MovT analyses here, because we had no specific predictions (see footnote 1 for Experiment 1 MovT results summary; MovTs were analyzed using the same methods as RelTs).

On RelTs, we found a Fluency × Movement interaction, $F(1, 3057.16) = 5.79$, $p = .02$ (independent from the actual "good" or "bad" classification made by the participants). Post hoc analysis of this interaction showed that participants initiated flexion movements quicker to fluent pseudowords, significant when compared with disfluent flexion, $b = .05$, $t = 4.16$, $p < .001$, $d_Z = 0.74$, and marginal when compared with fluent extension, $b = .02$, $t = 1.67$, $p = .09$, $d_Z = 0.30$. Note that we also observed a fluency main effect, $F(1, 29.94) = 11.25$, $p < .01$, $d_Z = 0.59$, such that overall, subjects initiated all movements more quickly in response to fluent pseudowords (Figure 3).

Liking judgments. Scale ratings for self-report liking (1 = *not at all*; 4 = *very much*) were analyzed in the same way as RelTs, using mixed-effects modeling according to a Fluency (2: fluent, disfluent) × Movement (2: flexion, extension) fixed-effects structure.

[1] With MovTs for Experiment 1, we found main effects for both fluency, $F(1, 67.46) = 8.76$, $p < .01$, and movement, $F(1, 32.23) = 48.92$, $p < .001$. Post hoc breakdowns of these effects showed that participants had quicker MovTs in response to disfluent pseudowords, and overall, they were faster to perform extension movements. Outliers (greater or less than 3 SDs from each individual subject's total average MovT) constituted 3.75% of all trials, which were removed before $\log_{10}$-transforming the remaining valid MovTs.

We only observed a main effect of fluency, $F(1, 32.04) = 15.21$, $p < .001$, $d_Z = 0.69$, where participants reported greater liking for fluent pseudowords ($M = 2.55$; $SD = .83$) than disfluent pseudowords ($M = 2.38$; $SD = .76$; Figure 3).

Classification decisions. We also evaluated whether or not fluency impacted participants' classification decisions ("good" or "bad") of the different pseudoword targets. To do this, we constructed a mixed-effects model on the binary decision outcome ("good" or "bad" classification), according to a Fluency (2: fluent, disfluent) × Movement (2: flexion, extension) fixed-effects structure.

We found a main effect of fluency, $F(1, 31.23) = 9.41$, $p < .01$, such that fluent pseudowords predicted a greater likelihood for subjects to log a "good" classification compared with disfluent pseudowords, after both flexion movements, $b = .09$, $t = 2.62$, $p = .01$, $d_Z = 0.46$, and extension movements, $b = .09$, $t = 2.72$, $p < .01$, $d_Z = 0.48$ (Table 1). To gauge whether or not participants were more likely to just classify the pseudowords as "good" or "bad" overall, we tested subjects' mean proportions against a 50% chance level (0 = "*bad*"; 1 = "*good*"). This showed no effect, $t(31) = 1.13$, nonsignificant (*ns*), demonstrating that participants generally were not more or less likely to classify pseudowords as "good" or "bad."

Experiment 2

Experiment 1 established the basic effect that perceptual fluency facilitates flexion (but not extension) RelTs, along with more positive classifications and evaluative ratings of liking. Experiment 2 aimed to examine the boundary conditions of the RelT effect. Recall that participants in Experiment 1 made movements to classify the different pseudowords as either "good" or "bad," which embeds the action in an affective categorization task. Therefore, the only change in Experiment 2 was that the AAT classification task was *non*-affective (i.e., classifying each pseudoword as "living" or "nonliving"). If action effects of fluency require embedding them in an affective context, then the RelT effects should disappear, just as they do in approach-avoidance studies that use facial expressions as stimuli (Rotteveel & Phaf, 2004). Critically, fluency effects on affect should remain and still influence the self-report ratings of liking. As such, we kept the later preference judgment the same, to assess any changes in the simple liking component of fluency (Figure 2).

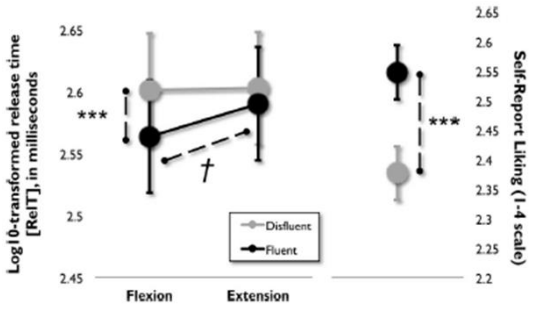


Figure 3. In Experiment 1, we observed a Fluency × Movement interaction for release times (RelTs). Participants initiated flexion (but not extension) movements quicker in response to fluent compared with disfluent pseudowords (left panel). Participants also reported greater liking for fluent compared with disfluent pseudowords (right panel). Error bars represent ± 1 *SEM* (*** $p < .001$; ** $p < .001$; * $p < .05$; † $p < .10$).

Table 1
Percentage of Classification Decisions According to Fluency (Fluent vs. Disfluent), Classification ("Good" vs. "Bad"), and Movement (Flexion vs. Extension) for Experiment 1

		Movement		
Fluency	Classification	Flexion	Extension	Total
Disfluent	"Good"	11.20	10.18	21.38
	"Bad"	15.00	13.92	28.92
Fluent	"Good"	13.09	12.74	25.83
	"Bad"	12.00	11.87	23.87
	Total	51.29	48.71	100.00

Note. As evidenced by a main effect of fluency on the binary classification decision ("good" or "bad"), participants made more "good" classifications when responding to fluent pseudowords, along with more "bad" classifications when responding to disfluent pseudowords. Overall, participants showed relatively even proportions for total "good" versus "bad" classifications. Note that the split between fluent and disfluent trials may not be exactly 50-50, given that a very small number of error trials did not record any AAT classification response (e.g., subjects releasing the "home" button before pseudoword onset).

Method

Participants. Thirty-five UCSD undergraduates participated for course credit. All participants were right-handed English speakers.

Materials and apparatus. We used the same stimuli and setup as Experiment 1 (Figures 1 and 2).

Design and procedure. To investigate boundary conditions, we only made one change to the Experiment 1 design. Instead of an affective classification ("good" or "bad"), Experiment 2 used a *non*-affective classification ("living" or "nonliving"). As before, participants were asked to make fast, intuitive decisions and were told that the words are not in English. All other task parameters were the same (Figures 1 and 2).

Results

RelTs. As with Experiment 1, after removing outliers greater or less than 3 *SD*s from each subject's total mean RelT (which constituted 4.51% of all trials), valid RelTs were then $\log_{10}$-transformed and analyzed using mixed-effects modeling, according to a Fluency (2: fluent, disfluent) × Movement (2: flexion, extension) fixed-effects structure. Once again, we will not focus on exploratory MovT analyses here, because we had no specific predictions (see footnote 2 for Experiment 2 MovT results summary; MovTs were analyzed using the same methods as RelTs).

[2] Our analysis of Experiment 2 MovTs only showed a main effect of movement, $F(1, 34.92) = 49.23$, $p < .001$, such that subjects were faster at performing extension movements. Outliers (greater or less than 3 *SD*s from each individual subject's total average MovT) constituted 4.86% of all trials, which were removed before $\log_{10}$-transforming the remaining valid MovTs.

In contrast to Experiment 1, we did *not* observe a Fluency × Movement interaction, $F(1, 67.24) = 1.92$, *ns* (independent from the "living" or "nonliving" classification). Our analysis only yielded a significant main effect of fluency, $F(1, 34.97) = 9.61$, $p < .01$, $d_Z = 0.52$, which revealed that subjects initiated movements more quickly in response to fluent pseudowords (Figure 4).

Liking judgments. Liking judgments were analyzed with the same methods as Experiment 1, using mixed-effects modeling. Here, we replicated the fluency main effect from Experiment 1, $F(1, 35.19) = 4.98$, $p = .03$, $d_Z = 0.38$, where participants reported greater liking for fluent ($M = 2.52$, $SD = .88$) compared with disfluent ($M = 2.35$, $SD = .86$) pseudowords (Figure 4).

Classification decisions. Classification decisions ("living" or "nonliving") were analyzed with the same methods as Experiment 1, using mixed-effects modeling. Here, we also did *not* observe any fluency or movement effects on the binary classification outcome, Fs < 1.36, *ns*. As with Experiment 1, we also assessed whether or not participants were more likely to just classify the pseudowords as "living" or "nonliving" overall, by testing subjects' mean proportions against a 50% chance level (0 = "*nonliving*"; 1 = "*living*"). This test was significant, $t(34) = 5.38$, $p < .001$, $d_Z = 0.91$, demonstrating that generally, participants were more likely to classify the pseudowords as "nonliving" over "living" (Table 2).

Table 2
Percentage of Classification Decisions According to Fluency (Fluent vs. Disfluent), Classification ("Living" vs. "Nonliving"), and Movement (Flexion vs. Extension) for Experiment 2

		Movement		
Fluency	Classification	Flexion	Extension	Total
Disfluent	"Living"	10.68	11.21	21.89
	"Nonliving"	13.84	14.46	28.30
Fluent	"Living"	8.85	11.04	19.89
	"Nonliving"	13.93	15.99	29.92
	Total	47.30	52.70	100.00

Note. No effects of fluency or movement were observed on the binary classification decision. Overall, participants were more likely to classify the pseudowords as "nonliving" over "living." Note that the split between fluent and disfluent trials may not be exactly 50-50, given that a very small number of error trials did not record any AAT classification response (e.g., subjects releasing the "home" button before pseudoword onset).

Experiment 3

To recap, Experiment 1 showed that higher fluency enhances flexion RelTs (but not extension). Fluency also increases "good" classifications, along with later liking judgments. Experiment 2 found that when movements are embedded in a nonaffective decision, fluency effects on action disappear, but remain in the later liking judgments.

Considering these findings, we had three main goals with Experiment 3. First, we aimed to replicate the action effect from Experiment 1, again embedding the movement in an affective classification. Second, we wanted to see whether fluency effects on movement generation are observed even if the classification action is not immediate, but delayed for several seconds. The presence of the effect despite the delay would argue for at least a temporary change in the evaluative perception of the stimulus. Third, we wanted to further explore the nature of the underlying affective response to fluency with a measure that is unobtrusive (noninvasive), valence-specific (able to distinguish between positive and negative reactions), continuous (high temporal resolution), and nonverbal (independent of self-reports). All of this is possible with fEMG, which is an ideal measure for these purposes (Tassinary, Cacioppo, & Vanman, 2007). We expected a rapid increase in *zygomaticus major* ("smiling muscle") activity and a rapid decrease in *corrugator supercilii* ("frowning muscle") activity to fluent stimuli, indicating that fluency-preference effects reflect a genuine affective change (Cannon, Hayes, & Tipper, 2010; Harmon-Jones & Allen, 2001; Topolinski, Likowski, Weyers, & Strack, 2009; Topolinksi & Strack, 2015; Winkielman & Cacioppo, 2001). It is important that this effect should occur before participants make any overt evaluation of the stimulus, highlighting the rapid and spontaneous unfolding of the affective process.

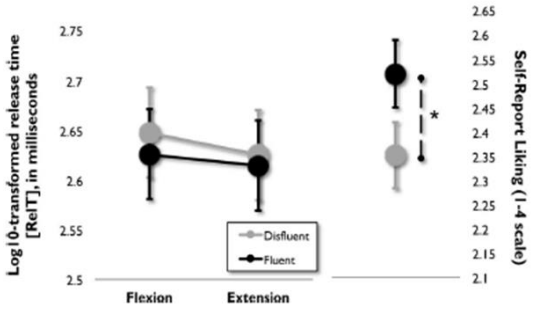


Figure 4. For Experiment 2, we did not observe any interactive effects for release times (RelTs; left panel). However, similar to Experiment 1, participants still reported greater overall liking for fluent compared with disfluent pseudowords (right panel). Error bars represent ± 1 *SEM* (*** $p < .001$; ** $p < .001$; * $p < .05$; † $p < .10$).

Method

Participants. Twenty-nine UCSD undergraduates participated for course credit. All participants were right-handed English speakers.

Materials and apparatus. We used the same stimuli and experimental setup as Experiments 1 and 2 (Figures 1 and 2).

Design and procedure. The task for Experiment 3 was the same as Experiment 1, but we also incorporated fEMG acquisition for 5,000 ms after stimulus presentation, to evaluate the spontaneous affective responses to fluency with a physiological measure (before any self-report) and whether fluency effects on movement generation are expressed in a *delayed* action (Figure 2).

After receiving task instructions, bipolar surface electrodes were placed unilaterally on the left side of the face over the *zygomaticus major* ("smiling muscle") and *corrugator supercilii* ("frowning muscle") to gauge fEMG responses, in accordance with past research (Tassinary, Cacioppo, & Vanman, 2007). All other methods were the same as Experiment 1 (Figures 1 and 2).

Signals were recorded with MP150CE/EMG2-T BioNomadix wireless data acquisition system and AcqKnowledge Version 4.1.1 software (Biopac Systems Inc., Santa Barbara, CA). All channel sampling rates were 2,000 Hz. Analyses (calculation, cleaning, and standardization) used MindWare EMG 2.52 (MindWare Technologies Ltd., Gahanna, OH) and coding scripts in MATLAB (Version R2014a; MathWorks Inc, Natick, MA).

Results

RelTs. We assessed RelTs in the same way as Experiments 1 and 2. After removing outliers (i.e., RelTs greater or less than 3 SDs from each subject's total average RelT; 1.66% of all trials), the remaining valid RelTs were $\log_{10}$-transformed and analyzed using mixed-effects modeling, according to a Fluency (2: fluent, disfluent) × Movement (2: flexion, extension) fixed-effects structure. As before, we will not focus on exploratory MovT analyses here, since we had no specific predictions (see footnote 3 for Experiment 3 MovT results summary; MovTs were analyzed using the same methods as RelTs).

We replicated the Fluency × Movement interaction from Experiment 1, $F(1, 2794.72) = 4.05$, $p = .04$, such that participants initiated flexion movements more quickly in response to fluent stimuli, compared with fluent extension, $b = 0.01$, $t = 2.38$, $p = .02$, $d_Z = 0.44$. Similar to Experiment 1, fluent flexion RelTs were faster than disfluent flexion RelTs, but this did not quite reach significance, $b = 0.01$, $t = 1.21$, *ns*, $d_Z = 0.22$ (Figure 5).

Liking judgments. Liking judgments were analyzed with the same methods as Experiments 1 and 2, using mixed-effects modeling. Critically, we replicated the fluency main effect from Experiments 1 and 2, $F(1, 29.03) = 28.19$, $p < .001$, $d_Z = 0.99$, where participants reported greater liking for fluent ($M = 2.57$, $SD = .76$) compared with disfluent ($M = 2.34$, $SD = .69$) pseudowords (Figure 5).

Classification decisions. Classification decisions ("good" vs. "bad") were analyzed with the same methods as Experiments 1 and 2, using mixed-effects modeling on the binary AAT outcome.

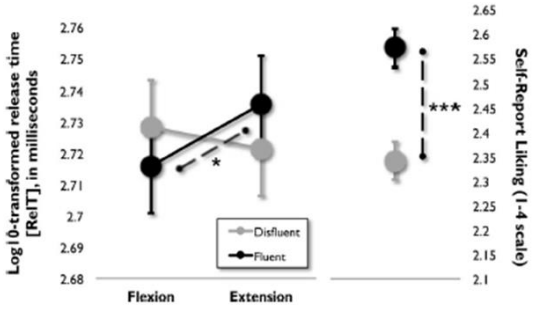


Figure 5. In Experiment 3, we replicated the Fluency × Movement interaction for release times (RelTs) from Experiment 1, where participants initiated flexion movements more quickly in response to fluent pseudowords (left panel). Also, as was the case with Experiments 1 and 2, participants reported greater overall liking for fluent compared with disfluent pseudowords (right panel). Error bars represent ± 1 *SEM* (*** $p < .001$; ** $p < .001$; * $p < .05$; † $p < .10$).

Here, we replicated the fluency main effect, $F(1, 29.02) = 26.66$, $p < .001$, from Experiment 1, such that more fluent stimuli predicted a greater likelihood for subjects to log a "good" classification, after both flexion movements, $b = .14$, $t = 4.26$, $p < .001$, $d_Z = 0.79$, and extension movements, $b = .15$, $t = 4.45$, $p < .001$, $d_Z = 0.83$ (Table 3). Also, as before, we tested subjects' mean proportions against a 50% chance level, to gauge whether or not they were more likely to just classify the pseudowords as "good" or "bad" overall (0 = "*bad*"; 1 = "*good*"). Similar to Experiment 1, this comparison showed no effect, $t(28) = 1.27$, *ns*, demonstrating that participants generally were not more or less likely to classify pseudowords as "good" or "bad."

fEMG. We also examined participants' fEMG data with mixed-effects modeling (using similar methods as RelT, MovT, and liking data), according to a Fluency (2: fluent, disfluent) × Muscle (2: corrugator, zygomaticus) × Time (10: 500 ms to 5,000 ms [in 500 ms groups]) fixed-effects structure.[4]

During each trial, the 3,000-ms fixation period was used as a baseline to which the subsequent 5,000-ms response period was compared. First, all EMG signals were filtered, rectified, and evaluated for movement artifacts. Second, to ensure that the data were properly cleaned and filtered, means and *SD*s were first calculated on each participant's raw dataset, and all trial points outside the ± 3 *SD* range from the mean signal for that participant were removed. Third, a new mean and *SD* for each participant was calculated based on the remaining valid trials, and all signals were *z*-scored. Finally, the median of the *z*-scored baseline for each participant was subtracted from each 500 ms *z*-scored trial point, and this process yielded a time-course of baseline-corrected, *z*-scored signals across 500-ms intervals for all trials. In sum, the final data (Figure 6) used for our analysis represent the *z*-scored change in muscle activity from baseline (standardized by each individual participant).

As expected, we found a Fluency × Muscle interaction, $F(1, 49.64) = 10.25$, $p < .01$, showing increased zygomaticus activity to fluent compared with disfluent pseudowords, $b = .04$, $t = 2.95$, $p < .01$, $d_Z = 0.59$. Participants also increased corrugator activity to disfluent compared with fluent pseudowords, but this effect did not reach significance, $b = .02$, $t = 1.57$, *ns*, $d_Z = 0.31$.

Note that we also observed a Muscle × Time interaction, $F(9, 225.00) = 9.66$, $p < .001$, which showed that the corrugator peaked very early, at the first 500-ms time point, which probably reflects an orienting response. The zygomaticus peaked later, at approximately 2,000 ms after stimulus onset, but still substantially before any overt liking judgments (Figure 6).

Experiment 4

As a quick review, in three experiments, we found that greater fluency leads to faster initiation of arm flexion during affective classification of initially neutral stimuli (Experiments 1 and 3). These

[3] We did not observe any significant main effects or interactions on Experiment 3 MovTs. Outliers (greater or less than 3 *SD*s from each individual subject's total mean MovT) constituted 2.14% of all trials, which were removed before $\log_{10}$-transforming the remaining valid MovTs.

[4] Note that for all Experiment 3 fEMG analyses, we had $n = 25$ (instead of $n = 29$ with the behavioral data), because of computer errors in saving four participants' physiology files.

Table 3
Percentage of Classification Decisions According to Fluency (Fluent vs. Disfluent), Classification ("Good" vs. "Bad"), and Movement (Flexion vs. Extension) for Experiment 3

		Movement		
Fluency	Classification	Flexion	Extension	Total
Disfluent	"Good"	9.52	10.28	19.80
	"Bad"	14.72	15.48	30.20
Fluent	"Good"	12.83	14.17	27.00
	"Bad"	10.83	12.17	23.00
	Total	47.90	52.10	100.00

Note. Similar to Experiment 1, a main effect of fluency on the binary classification decision ("good" vs. "bad") showed that participants made more "good" classifications when responding to fluent pseudowords, along with more "bad" classifications when responding to disfluent pseudowords. Overall, participants showed relatively even proportions for total "good" versus "bad" classifications. Note that the split between fluent and disfluent trials may not be exactly 50-50, given that a very small number of error trials did not record any AAT classification response (e.g., subjects releasing the "home" button before pseudoword onset).

effects were accompanied by increased liking ratings and higher proportion of positive classifications for the fluent stimuli. Experiment 3 also found that these effects are accompanied by a low-level hedonic response, as shown by increased zygomaticus "smiling" activity to fluent pseudowords. It is interesting that when the same stimuli are placed in a *non*-affective decision context (i.e., "living" or "nonliving" classifications), these RelT effects disappear (Experiment 2).

With Experiment 4, we aimed to use fEMG to answer an important open question regarding the relationship between the hedonic response (physiology) and action-tendency (RelTs). Recall that in Experiment 3, we observed significantly increased smiling (zygomaticus) and a tendency toward reduced frowning (corrugator) to fluent pseudowords when the stimuli were embedded in an affective decision (i.e., "good" or "bad" classifications). This leaves unclear the role of the classification decision context for the relationship between fluency-elicited affect and action. Specifically, Experiment 2 showed that the fluency effect on approach action dissipates when the decision is *non*-affective (i.e., "living" or "nonliving" classifications). However, does this occur simply because no genuine hedonic response is elicited in that case? Or, perhaps more interestingly, an underlying hedonic state is elicited by high fluency, but it does not extend to modify the action?

We answered this question very simply in Experiment 4, by changing the affective classification task from fEMG Experiment 3 ("good" or "bad" classifications) to the same *non*-affective classification task from Experiment 2 ("living" or "nonliving" classifications). This simple change makes the results quite informative. Let us assume we *do* observe similar fEMG effects in Experiment 4 as Experiment 3 (i.e., increased zygomaticus and reduced corrugator activation to fluent pseudowords) *without* a RelT effect. This would suggest that RelT effects disappear for nonaffective classifications because the elicited hedonic reaction is not translated to a motor response (the fluency-based affect is somehow "gated" from the action-tendency). It would also mean that any RelT effects cannot simply be because of the global changes in the affective state itself. Conversely, let us assume we *do not* detect similar fEMG effects in Experiment 4 as Experiment 3 (i.e., no differences in zygomaticus or corrugator between fluent and disfluent pseudowords). This would suggest that RelT effects disappear for nonaffective classifications because fluency fails to elicit positive affect in the first place (nonaffective action context "switches off" early affective responding, which causes any downstream RelT effects to go away).

Method

Participants. Thirty-seven UCSD undergraduates participated for course credit. All participants were right-handed English speakers.

Materials and apparatus. We used the same stimuli and experimental setup as Experiments 1, 2, and 3 (Figures 1 and 2).

Design and procedure. To investigate the fEMG responses as a function of decision context, we only made one change to the Experiment 3 fEMG design. Instead of an affective classification ("good" or "bad"), Experiment 4 used the same *non*-affective

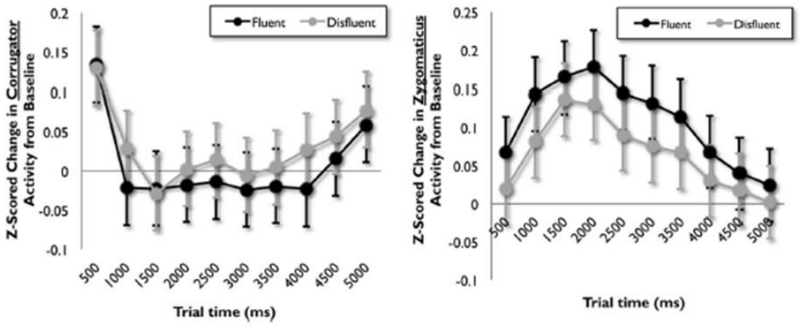


Figure 6. During Experiment 3, we measured facial electromyography (fEMG) over the corrugator (left panel) and zygomaticus (right panel) muscles. A Fluency × Muscle interaction indicated that participants recruited more zygomaticus (smiling) activation in response to fluent pseudowords. Subjects also recruited more corrugator (frowning) activation in response to disfluent pseudowords, but this effect did not reach significance. Error bars represent ± 1 *SEM*.

classification as Experiment 2 ("living" or "nonliving"). All other task parameters were the same as Experiment 3 (Figures 1 and 2).

Results

RelTs. We evaluated RelTs in the same way as Experiments 1, 2, and 3. After removing outliers (i.e., RelTs greater or less than 3 SDs from each subject's total mean RelT; 1.57% of all trials), the remaining valid RelTs were $\log_{10}$-transformed and analyzed using mixed-effects modeling, according to a Fluency (2: fluent, disfluent) × Movement (2: flexion, extension) fixed-effects structure. As was the case with previous experiments, we will not focus on exploratory MovT analyses here, because we had no specific predictions (see footnote 5 for Experiment 4 MovT results summary; MovTs were analyzed using the same methods as RelTs).

Here, we observed similar results to Experiment 2. We did *not* detect any evidence for a Fluency × Movement interaction, $F(1, 72.62) < .01$, *ns*. Our analysis only yielded a main effect of Movement, $F(1, 73.05) = 4.82$, $p = .03$, $d_Z = 0.36$, which showed that subjects initiated flexion movements faster than extension movements (Figure 7).

Liking judgments. Liking judgments were analyzed with the same methods as Experiments 1, 2, and 3, using mixed-effects modeling. Crucially, we replicated the fluency main effect from the three previous experiments, $F(1, 37.10) = 13.15$, $p < .001$, $d_Z = 0.60$, where subjects reported greater liking for fluent ($M = 2.53$, $SD = .75$) compared with disfluent ($M = 2.36$, $SD = .74$) pseudowords (Figure 7).

Classification decisions. Similar to previous experiments, we analyzed classification decisions ("living" or "nonliving") using mixed-effects modeling. Here, we did observe a main effect of fluency, $F(1, 3663.10) = 7.57$, $p < .01$, $d_Z = 0.45$, which showed that participants were more likely to categorize fluent pseudowords as "nonliving," compared with disfluent pseudowords. Furthermore, as with the previous three experiments, we also evaluated whether or not participants were more likely to just classify the pseudowords as "living" or "nonliving" overall, by testing subjects' mean proportions against a 50% chance level (0 = "nonliving"; 1 = "living"). Similar to Experiment 2, this test was significant, $t(36) = 7.33$, $p < .001$, $d_Z = 1.21$, demonstrating that generally, participants were more likely to classify the pseudowords as "nonliving" over "living" (Table 4).

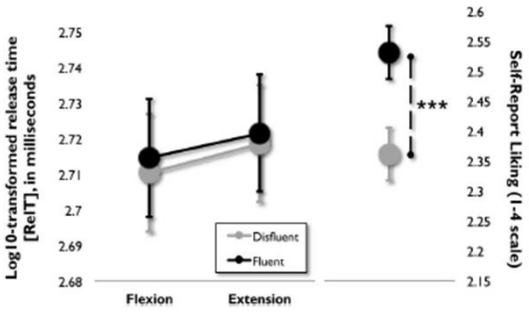


Figure 7. Similar to Experiment 2, we did not observe any interactive effects for release times (RelTs) in Experiment 4 (left panel). Also, as was the case with the previous three experiments, participants reported greater overall liking for fluent compared with disfluent pseudowords. Error bars represent ± 1 *SEM* (*** $p < .001$; ** $p < .001$; * $p < .05$; † $p < .10$).

fEMG. We assessed subjects' fEMG data with the same preprocessing steps and analysis methods as Experiment 3, using mixed-effects modeling according to a Fluency (2: fluent, disfluent) × Muscle (2: corrugator, zygomaticus) × Time (10: 500 ms to 5,000 ms [in 500-ms groups]) fixed-effects structure.

Intriguingly, even though we did not detect any RelT effects (similar to Experiment 2), we *did* observe a similar pattern of fEMG effects as Experiment 3. Specifically, we saw a Fluency × Muscle interaction, $F(1, 72.32) = 4.55$, $p = .04$, where subjects exhibited less corrugator (frowning) muscle reactivity to fluent compared with disfluent pseudowords, $b = .04$, $t = 2.43$, $p = .02$, $d_Z = 0.40$. We also observed a similar pattern of zygomaticus activation as Experiment 3, with increased smiling to fluent compared with disfluent pseudowords, but this effect did not reach the level of significance, $b = .01$, $t = .56$, *ns*, $d_Z = 0.09$ (Figure 8).

Note also that the temporal dynamics of muscle activation was similar to Experiment 3, resulting in a similar Muscle x Time interaction, $F(9, 332.87) = 5.25$, $p < .001$. As shown in Figure 8, the corrugator peaked early at approximately 500 ms (probably an orienting response), while the zygomaticus peaked later at around 1,500 ms. Crucially though, fluency condition influenced the physiological measure of affect early in the trial.

General Discussion

In summary, the current research discovered a link between perceptual fluency and approach action. Experiments 1 and 3 found faster RelTs to initiate arm flexion to fluent stimuli. This effect occurred with actions that were rapid (Experiment 1) or delayed (Experiment 3) and was primarily observed in faster initiation of arm flexion to fluent stimuli, but not arm extension to disfluent stimuli. It is important that the fluency-flexion link required embedding the action within affective classification decisions and was absent in Experiments 2 and 4—which used the same action for nonaffective classification decisions. This is rather interesting, given that in all four experiments, fluency robustly enhanced self-reported liking ratings. We also found that fluency elicits early, spontaneous affective responses that are detectable on the physiological level using fEMG (i.e., increased smiling activity and reduced frowning activity; Experiments 3 and 4). These fEMG effects occurred across *both* affective and nonaffective decision contexts, before participants made any overt liking judgments. Finally, all these effects emerged without participants reporting a connection between fluency, movement, and affect.

Our findings suggest that, within a context of affective decision, perceptual fluency influences action-tendencies. This influence occurred particularly with respect to action-readiness, as suggested by the impact on RelT (Frijda, 2010)—which was particularly pronounced for flexion (i.e., associated with approach in our paradigm; Rotteveel & Phaf, 2004). To our knowledge, this is the

[5] We did not observe any significant main effects or interactions on Experiment 4 MovTs. Outliers (greater or less than 3 SDs from each individual subject's total average MovT) constituted 2.03% of all trials, which were removed before $\log_{10}$-transforming the remaining valid MovTs.

Table 4
Percentage of Classification Decisions According to Fluency (Fluent vs. Disfluent), Classification ("Living" vs. "Nonliving"), and Movement (Flexion vs. Extension) for Experiment 4

		Movement		
Fluency	Classification	Flexion	Extension	Total
Disfluent	"Living"	10.38	9.65	20.03
	"Nonliving"	15.35	14.62	29.97
Fluent	"Living"	7.73	10.19	17.92
	"Nonliving"	14.81	17.27	32.08
	Total	48.27	51.73	100.00

Note. We observed a main effect of fluency on the binary classification decision ("living or nonliving"), which demonstrated that subjects made more "nonliving" classifications when responding to fluent pseudowords. Overall, participants were more likely to classify the pseudowords as "nonliving" over "living." Note that the split between fluent and disfluent trials may not be exactly 50-50, given that a very small number of error trials did not record any AAT classification response (e.g., subjects releasing the "home" button before pseudoword onset).

first evidence for a link between fluency and approach behavior. Theoretically, this finding suggests an important revision to current fluency models, highlighting its previously neglected consequences for motivation-relevant action (Winkielman et al., 2003). This is particularly interesting given that our pseudoword stimuli were initially neutral and low in arousal. As such, our findings suggest that, at least under certain conditions, perceptual fluency can modify the valence of the stimulus with enough strength and duration (Experiment 3) to make it function like an intrinsically valenced stimulus (such as an emotional word or an emotional facial expression). Keep in mind, however, that our study did not directly compare different types of stimuli and manipulations; thus, future research may assess the relative size of the action effect elicited by fluency and other valence sources.

Our findings are consistent with earlier reports that the mere-exposure effect is related to neural indices of individual differences in approach motivation (Harmon-Jones & Allen, 2001). Because the mere-exposure effect involves both fluency and familiarity, future studies may explore the relative role of these closely related factors in the respective neural effects and consequences for action. Interestingly, because our study used completely unfamiliar pseudowords, it suggests that relatively "pure" enhancements of stimulus fluency (without familiarity) are sufficient to influence approach action.

It is worth acknowledging that our predictions, methods, and discussions of the results assumed that arm flexion indicates approach. This is consistent with other work using this particular button-tower paradigm (see Phaf et al., 2014). However, note that our main claim is about a link between fluency and approach, which in this specific paradigm manifests as a link between fluency and arm flexion. Future studies may examine whether similar results would be obtained with different approach-avoidance paradigms (for a review, see Krieglmeyer et al., 2013) or even with different framings of the same movement (e.g., framing extension as approach, as in reaching out to pet a cute animal; see Seibt, Neumann, Nussinson, & Strack, 2008).

A related theoretical issue is the potential role of response labeling in these effects—or the idea that action is facilitated by the match between stimulus valence and the evaluative meaning of the response labels that are used in the task instructions (Eder & Rothermund, 2008). Basically, an action to fluent (and thus, more positive) stimuli could be initiated faster when the action is paired with a more positive label, such as a "top button" (involving arm flexion) as opposed to a "bottom button" (involving arm extension). However, this account would need to explain several features of our data. First, this alternative "affective response mapping" account is well suited to explain the connection between clearly valenced stimuli (positive and negative pictures or words) and clearly valenced response labels. However, in our paradigm, the stimuli (pseudowords) do not have any intrinsic valence—their new and temporary value derives from fluency. As such, the affective-mapping account would need to clarify how such newly acquired value gets mapped onto affective responses. Second, it is not clear on this alternative account why the fluency-flexion effect only occurred with actions embedded in evaluative decisions ("good" vs. "bad"), but not with actions involving nonaffective

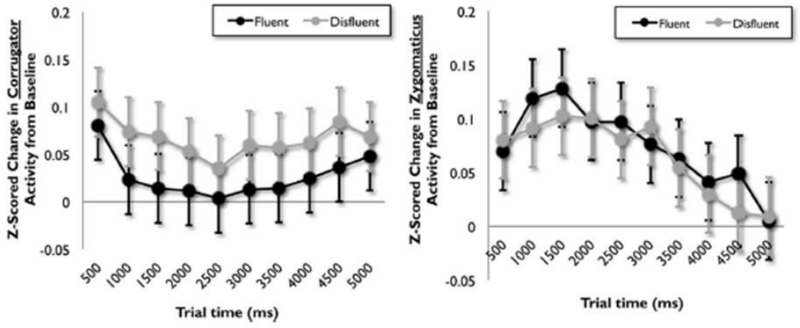


Figure 8. As with Experiment 3, we measured facial electromyography (fEMG) over the corrugator (left panel) and zygomaticus (right panel) muscles in Experiment 4. A Fluency × Muscle interaction indicated that participants recruited more corrugator (frowning) activation in response to disfluent pseudowords. While subjects also recruited more zygomaticus (smiling) activation in response to fluent pseudowords, this effect did not quite reach the level of significance. Error bars represent ± 1 *SEM*.

classification ("living" vs. "non-living"). After all, in all experiments, the response buttons were labeled "top" and "bottom," and fluent stimuli were still positive—as observed with positive fEMG responses and higher liking judgments (Experiment 4). Third, and perhaps most problematically for the affective-response mapping account, our facilitation effects were preferentially obtained on arm flexion to fluent stimuli, but not arm extension to disfluent stimuli. Nevertheless, future research should test our preferred motivational-direction versus the alternative response-labeling explanations of these action effects by, for example, reassigning button labels to the opposite response direction (i.e., labeling the "top" and "bottom" buttons as the "bottom-facing button" and "top-facing button," respectively). Critically, however, note that our core conclusions about the ability of fluency to influence action embedded in an affective context do not hinge on this particular debate.

As mentioned, in Experiments 2 and 4, when the classification decision was embedded in *non*-affective decisions ("living" or "non-living"), fluency had no effects on the speed of flexion or extension action. Fluency also only clearly influenced affective decisions ("good" vs. "bad"), since it increased the likelihood of classifying a pseudoword as "good" over "bad" in Experiments 1 and 3 (with no consistent effects on nonaffective decisions ["living" vs. "non-living"] in Experiments 2 and 4). It is important that the absence of the fluency effect on nonaffective decisions is unlikely to be driven by a floor effect, given the relatively similar proportions of "good/bad" decisions and "living/nonliving" decisions across experiments (Tables 1–4). Furthermore, fluency effects on liking ratings and physiological indices of positive affect were obtained across *both* affective and nonaffective contexts of the initial classification decision.

These differences in the impact of fluency, as a function of action context and measure, are interesting for a number of reasons. First, they demonstrate the power of context in shaping the emergence of valence-action links. This is consistent with previous reports showing that action effects using even strong, intrinsically valenced stimuli (e.g., happy and angry faces) require embedding the action in an affective decision (Rotteveel & Phaf, 2004). Second, they are consistent with proposals that even genuinely "liked" stimuli require a particular context to facilitate motivated behaviors, presumably indicative of "wanting" (Winkielman & Berridge, 2003). This is especially clear when considering the RelT and fEMG findings from Experiments 3 and 4. We found similar patterns of zygomaticus and corrugator fEMG, regardless of the initial classification task, indicating genuine liking for the fluent stimulus. Yet, the effects on action initiation (RelTs) only emerged during the affective classification task. Overall, these findings suggest that fluency instantiates a low-level hedonic response across multiple contexts, but this affective response is only selectively translated to action-tendency based on the relevance to the task at-hand (i.e., ones that require emotionally based judgments). Still, we hesitate to make any claims about an unconditional link between fluency and valence, since all four experiments involved a self-report of liking at the end. Future experiments may test whether fluency effects on fEMG emerge even in the absence of any consideration for the affective nature of the stimulus. This is indeed a difficult question, as nonaffective judgments or classification tasks can serve as potential distractors from the affective nature of stimuli and can disturb even stronger effects (e.g., Pessoa, McKenna, Gutierrez, & Ungerleider, 2002).

Regardless, in the current paradigm, the psychophysiological findings (fEMG; Experiments 3 and 4) suggest that the basic hedonic response to fluency arises quickly, prior to any overt judgment. This response usually has primarily a positive component, as revealed via increased zygomaticus (smiling) activity to fluent stimuli, for instance (Winkielman & Cacioppo, 2001). It is interesting that this positive skew was evident in the RelT results during Experiment 3 (i.e., fluency was associated with faster approach RelTs, but *dis*fluency was not connected to quicker extension RelTs). At the same time, Experiment 4 (which found no RelT effects) observed more robust fEMG responses on the corrugator (frowning), with fluent pseudowords significantly *reducing* corrugator reactivity (zygomaticus reactivity in response to fluent vs. disfluent pseudowords was not quite significant, but in the predicted direction). While many studies have demonstrated increased zygomaticus (smiling) activity to fluent stimuli (Winkielman & Cacioppo, 2001), other studies with fluency manipulated via semantic coherence have reported effects on the corrugator, which likely arise from reduced negative affect and relaxed mental effort (Topolinski, Likowski, Weyers, & Strack, 2009). Again, future studies may further explore these differences.

Finally, it is worth highlighting some general connections of the current work to other research on the links between fluency, action, and affect. Several labs have shown that easier actions (either because of practice, priming, or compatibility) elicit more positive responses, even to unrelated stimuli (Beilock & Holt, 2007; Brouillet, Ferrier, Grosselin, & Brouillet, 2011; Cannon et al., 2010). Our work is clearly different in that we manipulated the ease of perception (not the ease of action) and measured the relative speed of specific actions related to approach (flexion) versus avoidance (extension). Considering this, both lines of research point to the affective nature of the links between fluency of perception and fluency of action. More generally, these areas of research emphasize the fundamental connection between the dynamics of mental processes and the embodied, action-based nature of emotion (Winkielman, Niedenthal, Wielgosz, Eelen, & Kavanagh, 2015).

In conclusion, we provided the first evidence of a link between processing fluency and approach-related action. These results reveal that high fluency is linked to faster initiation of flexion movements—a novel demonstration that smoothness of perceptual processing shapes motor responding within the environment.

References

Aharon, I., Etcoff, N., Ariely, D., Chabris, C. F., O'Connor, E., & Breiter, H. C. (2001). Beautiful faces have variable reward value: FMRI and behavioral evidence. *Neuron, 32,* 537–551. http://dx.doi.org/10.1016/S0896-6273(01)00491-3

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

Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 9369–9372. http://dx.doi.org/10.1073/pnas.0601071103

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67,* 1–48.

Beilock, S. L., & Holt, L. E. (2007). Embodied preference judgments: Can likeability be driven by the motor system? *Psychological Science, 18,* 51–57. http://dx.doi.org/10.1111/j.1467-9280.2007.01848.x

Brouillet, T., Ferrier, L. P., Grosselin, A., & Brouillet, D. (2011). Action compatibility effects are hedonically marked and have incidental consequences on affective judgment. *Emotion, 11,* 1202–1205. http://dx.doi.org/10.1037/a0024742

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes. II: Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology, 65,* 5–17. http://dx.doi.org/10.1037/0022-3514.65.1.5

Cannon, P. R., Hayes, A. E., & Tipper, S. P. (2010). Sensorimotor fluency influences affect: Evidence from electromyography. *Cognition and Emotion, 24,* 681–691. http://dx.doi.org/10.1080/02699930902927698

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin, 25,* 215–224. http://dx.doi.org/10.1177/0146167299025002007

Darwin, C. (1997). *The expression of the emotions in man and animals.* New York, NY: Oxford University Press. (Original work published 1872)

Eder, A. B., & Rothermund, K. (2008). When do motor behaviors (mis-)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology: General, 137,* 262–281. http://dx.doi.org/10.1037/0096-3445.137.2.262

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion, 15,* 115–141. http://dx.doi.org/10.1080/02699930125908

Frijda, N. H. (1986). *The emotions.* New York, NY: Cambridge University Press.

Frijda, N. H. (2010). Impulsive action and motivation. *Biological Psychology, 84,* 570–579. http://dx.doi.org/10.1016/j.biopsycho.2010.01.005

Harmon-Jones, E., & Allen, J. J. B. (2001). The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality and Social Psychology Bulletin, 27,* 889–898. http://dx.doi.org/10.1177/0146167201277011

Harmon-Jones, E., Harmon-Jones, C., & Price, T. F. (2013). What is approach motivation? *Emotion Review, 5,* 291–295. http://dx.doi.org/10.1177/1754073913477509

Krieglmeyer, R., De Houwer, J., & Deutsch, R. (2013). On the nature of automatically triggered approach-avoidance responses. *Emotion Review, 5,* 280–284. http://dx.doi.org/10.1177/1754073913477501

Krieglmeyer, R., & Deutsch, R. (2010). Comparing measures of approach–avoidance behaviour: The manikin task vs. two versions of the joystick task. *Cognition and Emotion, 24,* 810–828. http://dx.doi.org/10.1080/02699930903047298

Krieglmeyer, R., Deutsch, R., De Houwer, J., & De Raedt, R. (2010). Being moved: Valence activates approach-avoidance behavior independently of evaluation and approach-avoidance intentions. *Psychological Science, 21,* 607–613. http://dx.doi.org/10.1177/0956797610365131

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). LmerTest: Tests for random and fixed effects for linear mixed effect models. R Package, Version 2.0–3.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review, 97,* 377–395. http://dx.doi.org/10.1037/0033-295X.97.3.377

Lee, A. Y., & Labroo, A. A. (2004). The effect of conceptual and perceptual fluency on brand evaluation. *Journal of Marketing Research, 41,* 151–165. http://dx.doi.org/10.1509/jmkr.41.2.151.28665

Litt, A., Khan, U., & Shiv, B. (2010). Lusting while loathing: Parallel counterdriving of wanting and liking. *Psychological Science, 21,* 118–125. http://dx.doi.org/10.1177/0956797609355633

Markman, A. B., & Brendl, C. M. (2005). Constraining theories of embodied cognition. *Psychological Science, 16,* 6–10. http://dx.doi.org/10.1111/j.0956-7976.2005.00772.x

Neumann, R., Förster, J., & Strack, F. (2003). Motor compatibility: The bidirectional link between behavior and evaluation. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 7–49). Mahwah, NJ: Erlbaum.

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of Marketing Research, 44,* 347–356. http://dx.doi.org/10.1509/jmkr.44.3.347

Pessoa, L., McKenna, M., Gutierrez, E., & Ungerleider, L. G. (2002). Neural processing of emotional faces requires attention. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 11458–11463. http://dx.doi.org/10.1073/pnas.172403899

Petty, R. E., Fazio, R. H., & Briñol, P. (Eds.). (2009). *Attitudes: Insights from the new implicit measures.* New York, NY: Psychology Press.

Phaf, R. H., Mohr, S. E., Rotteveel, M., & Wicherts, J. M. (2014). Approach, avoidance, and affect: A meta-analysis of approach-avoidance tendencies in manual reaction time tasks. *Frontiers in Psychology, 5,* 378. http://dx.doi.org/10.3389/fpsyg.2014.00378

Phaf, R. H., & Rotteveel, M. (2009). Looking at the bright side: The affective monitoring of direction. *Emotion, 9,* 729–733. http://dx.doi.org/10.1037/a0016308

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology: Section A, 55,* 1339–1362.

Rotteveel, M., Gierholz, A., Koch, G., van Aalst, C., Pinto, Y., Matzke, D., . . . Wagenmakers, E. J. (2015). On the automatic link between affect and tendencies to approach and avoid: Chen and Bargh (1999) revisited. *Frontiers in Psychology, 6,* 335.

Rotteveel, M., & Phaf, R. H. (2004). Automatic affective evaluation does not automatically predispose for arm flexion and extension. *Emotion, 4,* 156–172.

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review, 2,* 87–99. http://dx.doi.org/10.1207/s15327957pspr0202_2

Seibt, B., Neumann, R., Nussinson, R., & Strack, F. (2008). Movement direction or change in distance? Self- and object-related approach–avoidance motions. *Journal of Experimental Social Psychology, 44,* 713–720. http://dx.doi.org/10.1016/j.jesp.2007.04.013

Tassinary, L. G., Cacioppo, J. T., & Vanman, E. J. (2007). The skeletomotor system: Surface electromyography. In J. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 267–302). New York, NY: Cambridge University Press.

Topolinski, S., Likowski, K. U., Weyers, P., & Strack, F. (2009). The face of fluency: Semantic coherence automatically elicits a specific pattern of facial muscle reactions. *Cognition and Emotion, 23,* 260–271. http://dx.doi.org/10.1080/02699930801994112

Topolinski, S., & Strack, F. (2015). Corrugator activity confirms immediate negative affect in surprise. *Frontiers in Psychology, 6,* 134. http://dx.doi.org/10.3389/fpsyg.2015.00134

Van Dantzig, S., Zeelenberg, R., & Pecher, D. (2009). Unconstraining theories of embodied cognition. *Journal of Experimental Social Psychology, 45,* 345–351. http://dx.doi.org/10.1016/j.jesp.2008.11.001

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

Winkielman, P., & Berridge, K. C. (2003). Irrational wanting and subrational liking: How rudimentary motivational and affective processes shape preferences and choices. *Political Psychology, 24,* 657–680. http://dx.doi.org/10.1046/j.1467-9221.2003.00346.x

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology, 81,* 989–1000. http://dx.doi.org/10.1037/0022-3514.81.6.989

Winkielman, P., Niedenthal, P., Wielgosz, J., Eelen, J., & Kavanagh, L. C. (2015). Embodiment of cognition and emotion. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA handbook of personality and social psychology, Vol. 1. Attitudes and social cognition* (pp. 151–175). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14341-004

Winkielman, P., Olszanowski, M., & Gola, M. (2015). Faces in between: Evaluative responses to faces reflect the interplay of features and task-dependent fluency. *Emotion, 15,* 232–242. http://dx.doi.org/10.1037/emo0000036

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: Lawrence Erlbaum.

Zajonc, R. B. (1998). Emotions. In D. T. Gilbert, S. T. Fiske & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 591–632). Boston, MA: McGraw-Hill.

Received February 13, 2015
Revision received September 18, 2015
Accepted November 6, 2015 ■

Chapter 4 is, in full, a reprint of the material as it appears in *Emotion*. Carr, Evan W.; Rotteveel, Mark; Winkielman, Piotr, 2016. The dissertation author was the primary investigator and author of this paper.

GENERAL DISCUSSION

The preceding chapters have provided an in-depth examination of how basic social perception, judgment, and action are impacted by processing dynamics (or the ease, speed, and coherence of processing). Two key factors were investigated — *familiarity* (prior stimulus experience) and *fluency* (ease of stimulus processing). Across 13 experiments, both familiarity and fluency altered classic psychological effects (Chapters 1 and 3) and rudimentary reactions to neutral and emotional stimuli (Chapters 2 and 4). Critically, the current work shows that processing dynamics can infiltrate relatively rapid and low-level social responses.

Let us review the main findings from each chapter (recall that Chapters 1 and 2 focused on familiarity, while Chapters 3 and 4 focused on fluency). Starting with Chapter 1, four experiments investigated predictions made my modern memory theories for how mere exposure impacts the attractiveness of facial blends (*beauty-in-averageness [BiA] effect*). Even though we replicated the classic BiA effect when the individuals were weakly learned (i.e., morphs rated as more attractive than their constituent individuals; Experiment 1), we show the first evidence for an *ugliness-in-averageness (UiA) effect* when the individuals are made highly familiar (i.e., morphs rated as less attractive than their constituent individuals; Experiments 2, 3, and 4). Importantly, both the BiA and UiA effects in our studies were mediated by familiarity. The BiA results (Experiment 1) follow from previous findings suggesting that attractiveness of average faces is associated with their implicit familiarity (Peskin & Newell, 2004; Rhodes, Halberstadt, & Brajkovich, 2001). Memory models would predict such an effect, since weak learning on individual faces should lead to a prioritization of the blend as a better match to the "gist" or prototypical representation of the faces (Principe & Langlois, 2012). However, the UiA results (Experiments 2, 3, and 4) show that the attractiveness of facial blends is contingent on the familiarity of their constituent individuals (where familiar individuals are rated as more attractive than their blends). From the memory-based account, this presumably occurs because strong

individual learning strengthens the representations for familiar faces, thus making blends seem less similar (and familiar) to relevant memory traces. In turn, we found that the UiA effect is driven by a relative reduction in familiarity for morphs of trained individuals (compared to the trained individuals themselves). Note that we specifically dissociated this account from another perspective of cognitive "mismatch," where the morphs of two familiar individuals appear especially unattractive because of cognitive conflict between two established categories or "attractors" (Arnal & Giraud, 2012; Dreisbach & Fisher, 2015; Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005; Neta, Kelley, & Whalen, 2013). This clash would trigger negative affect, which would then generalize to the morph. This account clearly predicts that the dislike should be eliminated if the conflict is removed, which can be achieved by replacing one of the conflicting components. Our data supported the familiarity (memory-based) account in Experiment 3, since a UiA effect still emerged when using "cross-set" morphs (composed of one trained individual and one untrained individual, as opposed to the "within-set" morphs containing two trained individuals in Experiments 2 and 4). Finally, the UiA effect still emerges with a purely perceptual learning task, suggesting that these processes occur due to low-level visual familiarity (Experiment 4).

Chapter 1 not only delivers novel evidence for a new phenomenon (along with further support for the familiarity-positivity link; Garcia-Marques, Mackie, Claypool, & Garcia-Marques, 2004; Monin, 2003; Titchener, 1915), but it also opens many new questions. First, one interesting avenue for future research would have to do with the role of affective, motivational, and judgmental contexts in the UiA effect. The "warm glow" of familiarity can 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 judgement in-question, since different dimensions show varying sensitivity to mere exposure and prototypicality manipulations

(DeBruine, 2005; Rhodes, Halberstadt, & Brajkovich, 2001; Rhodes, Halberstadt, Jeffery, & Palermo, 2005). Second, future research should also examine timing differences in the UiA effect, perhaps through the addition of physiological measures (e.g., EMG or EEG). Judgment and physiology can sometimes dissociate as stimulus processing moves from perceptual (early) to conceptual (late) stages (Bradley & Lang, 2007; Von Helversen, Gendolla, Winkielman, & Schmidt, 2008). Finally, it would be useful to explore other social manipulations alongside the UiA effect, including the role of valenced expressions (e.g., smiling and frowning faces) or group status (e.g., race or gender; Bernstein, Young, & Hugenberg, 2007; Malpass & Kravitz, 1969; Hugenberg & Bodenhausen, 2004).

Chapter 2 extended these familiarity effects to the perception of happiness in others' facial expressions. Up to this point, it was still unclear whether mere exposure effects could impact perception or if this was limited to only influencing higher-level judgments. With two experiments that involved speeded perceptual judgments (Experiment 1), rapid forced-choice classifications (Experiment 2), and deliberative estimates of happiness (Experiment 2), we found that participants judged familiar individuals' expressions as happier — particularly when the expressions were neutral or positive. The latter suggests that this effect involves the selective enhancement of positive stimulus features (rather than the reduction of negative stimulus features). Note that these findings cannot be explained by simple response biases, given the selectivity of the familiarity-positivity effect to only certain levels of emotion in the faces. Also, we observed the same pattern of findings across multiple tasks involving rapid perception (Experiment 1) and judgment (Experiment 2). Thus, familiarity seems to imbue facial stimuli with intrinsic positivity, which can be misattributed in subsequent perception of certain facial expressions.

One of the most valuable contributions from Chapter 2 comes with how it helps to dissociate prominent mere exposure models on the link between familiarity and valence. Our

results support *hedonic skew* frameworks, which posit that familiarity expresses on positive affect (but not negative affect) and gets expressed via positive features (but not negative features; Harmon-Jones & Allen, 2001; Winkielman & Cacioppo, 2001). In turn, our findings are also inconsistent with models for *amplification* of pre-existing features (also called nonspecific activation; Albrecht & Carbon, 2014; Mandler, Nakamura, & Van Zandt, 1987), selective decrease in negative affect (or a *negative skew*; Lee, 2001; Zajonc, 2001), or a *generalized positivity shift* (Monin, 2003; Tichener, 1915). It is especially intriguing that these targeted effects from familiarity can emerge at earlier stages of processing. Our interpretation of these findings is that familiarity works selectively on positive affect and is more easily attributed to positive features, but this work should be replicated with different kinds of emotion morphs (especially given some views that anger is "special" in its social processing; e.g., Pinkham, Griffin, Baron, Sasson, & Gur, 2010). Another future direction would be with the application of other perceptual tasks (e.g., visual search with trained and untrained faces), since top-down effects of cognition on perception are still being fervently debated (Balcetis, 2016; Firestone & Scholl, 2015).

Moving on to the fluency findings, Chapter 3 was similar to Chapter 1 in that it focused on an effect that has been widely shown and replicated (but can be transformed when processing dynamics are varied). More specifically, Chapter 3 investigated the seemingly obligatory discomfort that arises from perceiving "mixed" agents (or those that contain both human and non-human features, like androids; Ishiguro, 2007). We showed that the relative dislike for mixed agents is not inherent or compulsory, but rather it can be modified by contextual factors. Instead, such dislike is generated when people classify the agents into human versus non-human categories — resulting in the experience of categorization disfluency (which triggers negative affect and generalizes to agent evaluations). Categorization fluency (or the effort needed to determine category membership; Halberstadt & Winkielman, 2013) is different than perceptual

fluency, which is usually manipulated via low-level "surface" stimulus features (e.g., contrast, readability, duration, etc.; e.g., Carr, Rotteveel, & Winkielman, 2016; Reber, Winkielman, & Schwarz, 1998). Consequently, categorization fluency is ultimately task-dependent, and processing difficulty instead depends on which (un)ambiguous feature dimensions are highlighted by the current task (Owen, Halberstadt, Carr, & Winkielman, 2016).

Crucially though, the findings from Chapter 3 provide an important qualification to previous claims about an "unbridgeable" boundary between human and non-human entities. Mixed agents were devalued more so when participants first classified them on the ambiguous human-likeness dimension, and categorization difficulty (longer RTs) mediated these effects (Experiments 1 and 2). On views of essentialism, spontaneous negative responses to mixed agents arise due to an aversive combination of human and non-human "natural kinds" (see Prentice & Miller, 2007). Other perceptual frameworks propose similar "mismatches" to occur, with conflicting cues in visual, auditory, and motion processing (e.g., Mitchell et al., 2011). In turn, our results from Chapter 3 argue against such theories that do not allow for contextual sensitivity in the "automaticity" of negative responses to mixed agents, or the ability for flexible construal of essential human and non-human categories. However, the findings from Chapter 3 do not rule out bottom-up effects in processing mixed agents. In the current studies, participants still took somewhat longer to respond to androids even during the alternative control tasks (Experiments 1 and 2). Therefore, it is important to keep in mind that some versions of bottom-up perceptual theories can be considered compatible with fluency frameworks. For instance, incongruity may be detected early, which would lead to low-level difficulty in stimulus processing. Moreover, other theories hold that disfluency is not the actual proximal driver of devaluation through effort, but rather the implications of such disfluency is the most important factor. Disfluency can signal a gap in knowledge (e.g., Kruglanski, 2013), a shifted sense of meaning (Proulx & Inzlicht, 2012), or prediction error (Saygin, Chaminade, Ishiguro, Driver, &

Frith, 2012). As a result, Chapter 3 should be viewed as evidence that negative responses to mixed agents involve an interaction between bottom-up and top-down processes (rather than being solely top-down).

Note that Chapter 3 also reveals an important theoretical extension of previous fluency models. There were no devaluation effects in Experiment 3, where the human-classification condition was replaced with a *color*-classification task of the same "difficulty structure" (i.e., 50/50 blue-green judgments made android images selectively disfluent). This demonstrates that devaluation effects do not result from any categorization difficulty per se, but rather only selectively emerges based on the (un)ambiguous feature in-question and the judgment being rendered. We propose that devaluation effects may only follow from disfluency that occurs in response to an agent's key central features (i.e., *integral* human-likeness) rather than ancillary cues with similar ambiguity (i.e., *incidental* colored backgrounds; also see Bodenhausen, 1993). Even though human-classification and color-classification led to similar experiences of disfluent processing, the interpretation of those metacognitive experiences is likely what drove differences in the evaluation of mixed agents (see Schwarz, 2010, for a review). If one experiences disfluency on an integral feature of the agent (e.g., human-likeness), this will likely have downstream negative consequences on judgment. However, similar disfluency on an incidental feature (e.g., color background) would not be deemed relevant, and thus gated from the evaluation.

While the findings from Chapters 1-3 are valuable in showing effects on perception and judgment, they did not examine consequences on action. Chapter 4 addressed this gap by demonstrating how fluency facilitates rapid approach action-tendencies. Using an approach-avoidance task (AAT) with a vertical button tower, we found faster approach movements (RTs to initiate arm flexion) to perceptually fluent stimuli when participants had to classify them with an affective judgment (i.e., "good or bad"; Experiments 1 and 3). Interestingly, this fluency-action

effect dissipated in a non-affective judgment context (i.e., "living or non-living"; Experiments 2 and 4), even though fluent stimuli elicited genuine hedonic physiological responses (i.e., increased smiling and reduced frowning via fEMG; Experiments 3 and 4). To our knowledge, this is the first evidence for a link between fluency and approach behavior, suggesting an important revision to current fluency models which do not mention consequences for motivation-relevant action (Winkielman, Schwarz, Fazendeiro, & Reber, 2003). Our findings show that perceptual fluency can modify the valence of a stimulus with enough strength and duration to make it function like an intrinsically valenced stimulus (such as an emotional word or an emotional facial expression), even when our pseudoword stimuli were initially neutral and low in arousal. Keep in mind that this is also distinct from the mere exposure manipulations used in Chapters 1 and 2, which likely involve contributions from both fluency and familiarity. Interestingly, Chapter 4 reveals that relatively "pure" enhancements of stimulus fluency (without familiarity) are sufficient to influence approach action.

Among the most intriguing results from Chapter 4 were the fluency differences as a function of action context (i.e., affective "good or bad" judgment vs. non-affective "living or non-living" judgment) and measure (i.e., AAT release times [RelTs] vs. fEMG). In short, we only found RelT effects during the affective task (Experiments 1 and 3), but we observed fEMG effects in both the affective and non-affective tasks (Experiments 3 and 4). This follows from previous work on approach-avoidance with valenced stimuli (e.g., happy and angry faces), which require embedding the action in affective decision contexts (Rotteveel & Phaf, 2004). In turn, our findings suggest that fluency instantiates a low-level hedonic response across multiple contexts (fEMG effects in both Experiments 3 and 4), but this affective response is only selectively translated to action-tendency based on task relevance (RelT effects only in Experiment 3). This corresponds to proposals that even though a stimulus can be genuinely liked, the facilitation of motivation-related action only occurs in specific contexts that elicit wanting (Winkielman &

Berridge, 2003). Further, the fEMG findings suggest that the basic hedonic response to fluency arises quickly, before any explicit judgment.

Despite these insights from Chapter 4, there are many avenues for future research. First, the link between fluency and valence may not be unconditional, since participants were required to give a liking rating in all four experiments. Future studies could test whether similar fluency effects on fEMG emerge without any consideration for the affective nature of the stimulus. Second, note that a "positive skew" was evident in both the RelT and fEMG results during Experiment 3 (i.e., fluency was associated with faster approach RelTs and increased smiling, but *dis*fluency was not connected to quicker extension RelTs or increased frowning). On Experiment 4, we found no RelT effects, but fluent pseudowords did significantly *reduce* corrugator reactivity (zygomaticus reactivity in response to fluent vs. disfluent pseudowords was in the predicted direction but not quite significant). While physiological responses to fluency do tend to be skewed positive (e.g., increased smiling to fluent stimuli; Winkielman & Cacioppo, 2001), future experiments should investigate when and why different actions or muscles get activated. For example, other fluency studies have reported reduced frowning effects on the corrugator, presumably because of reduced negative affect and relaxed mental effort (Topolinski, Likowski, Weyers, & Strack, 2009). Finally, future studies should also further probe the motivational nature of these fluency-action effects. In Chapter 4, we equate arm flexion RelTs with approach (following previous work using this AAT paradigm; Phaf, Mohr, Rotteveel, & Wicherts, 2014), but since our main claim is about a link between fluency and approach, others may examine whether similar results can be obtained with different approach-avoidance paradigms (see Krieglmeyer et al., 2013) or different framings of the same movement (e.g., framing extension as approach; see Seibt, Neumann, Nussinson, & Strack, 2008).

In conclusion, this dissertation has revealed the effects of two key factors in processing dynamics — familiarity and fluency — on shaping socially relevant perception, judgment, and

action. These effects were robust enough to augment, attenuate, or even reverse supposedly "automatic" phenomena in social cognition (Chapters 1 and 3). Moreover, we showed changes to low-level responses originating in perception and action, which are often thought to be shielded from variations in familiarity and fluency (Chapters 2 and 4). Taken together, these experiments demonstrate that processing dynamics elicit a subtle subjective experience which is flexibly incorporated into our interaction with the social environment.

References

Albrecht, S., & Carbon, C. C. (2014). The Fluency Amplification Model: Fluent stimuli show more intense but not evidently more positive evaluations. *Acta Psychologica, 148*, 195-203.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390-398.

Balcetis, E. (2016). Approach and avoidance as organizing structures for motivated distance perception. *Emotion Review*, *8*(2), 115-128.

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*(8), 706-712.

Bodenhausen, G. V. (1993). Emotions, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping* (pp. 13-37). San Diego, CA: Academic Press.

Bradley, M. M., & Lang, P. J. (2007). Emotion and motivation. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 581-607). Cambridge, UK: Cambridge University Press.

Carr, E.W., Rotteveel, M., & Winkielman, P. (2016). Easy moves: Perceptual fluency facilitates approach-related action. *Emotion*, *16*(4), 540-552.

DeBruine, L. M. (2005). Trustworthy but not lust-worthy: Context-specific effects of facial resemblance. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1566), 919-922.

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.

Dreisbach, G. & Fischer, R. (2015). Conflicts as aversive signals for control adaptation. *Current Directions in Psychological Science, 24*, 255-260.

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behavioral and Brain Sciences*, 1-72.

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*(6), 636-644.

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.

Halberstadt, J. & Winkielman, P. (2013). When good blends go bad: How fluency can explain when we like and dislike ambiguity. In C. Unkelbach & R. Greisfelder. *The experience of*

thinking: How feelings from mental processes influence cognition and behavior (pp. 133–151). New York, NY: Psychology Press.

Harmon-Jones, E., & Allen, J. J. (2001). The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality and Social Psychology Bulletin*, *27*(7), 889-898.

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*(5), 1191.

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*(5754), 1680-1683.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization the role of prejudice and facial affect in race categorization. *Psychological Science*, *15*(5), 342-345.

Ishiguro, H. (2007). Android science. In *Robotics Research* (pp. 118-127). Springer Berlin Heidelberg.

Krieglmeyer, R., De Houwer, J., & Deutsch, R. (2013). On the nature of automatically triggered approach-avoidance responses. *Emotion Review, 5*, 280–284.

Kruglanski, A. W. (2013). *Lay epistemics and human knowledge: Cognitive and motivational bases*. Springer Science & Business Media.

Lee, A. Y. (2001). The mere exposure effect: An uncertainty reduction explanation revisited. *Personality and Social Psychology Bulletin*, *27*(10), 1255-1266.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*(4), 330.

Mandler, G., Nakamura, Y., & Van Zandt, B. J. (1987). Nonspecific effects of exposure on stimuli that cannot be recognized. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 646.

Mitchell, W. J., Szerszen Sr, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, *2*(1), 10.

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology*, *85*, 1035-1048.

Neta, M., Kelley, W. M., & Whalen, P. J. (2013). Neural responses to ambiguity involve domain-general and domain-specific emotion processing systems. *Journal of Cognitive Neuroscience*, *25*(4), 547-557.

Owen, H. E., Halberstadt, J., Carr, E. W., & Winkielman, P. (2016). Johnny Depp, reconsidered: How category-relative processing fluency determines the appeal of gender ambiguity. *PLoS ONE, 11*(2), e0146328.

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

Phaf, R. H., Mohr, S. E., Rotteveel, M., & Wicherts, J. M. (2014). Approach, avoidance, and affect: A meta-analysis of approach-avoidance tendencies in manual reaction time tasks. *Frontiers in Psychology, 5,* 378.

Pinkham, A. E., Griffin, M., Baron, R., Sasson, N. J., & Gur, R. C. (2010). The face in the crowd effect: Anger superiority when using real faces and multiple identities. *Emotion*, *10*(1), 141.

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. *Current Directions in Psychological Science, 16*(4), 202-206.

Principe, C. P. & Langlois, J. H. (2012). Shifting the prototype: Experience with faces influences affective and attractiveness preferences. *Social Cognition, 30*(1), 109-120.

Proulx, T., & Inzlicht, M. (2012). The five "A" s of meaning maintenance: Finding meaning in the theories of sense-making. *Psychological Inquiry*, *23*(4), 317-335.

Reber, R., Winkielman P., & Schwarz N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science, 9,* 45-48.

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition, 19*(1), 57-70.

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.

Rotteveel, M., & Phaf, R. H. (2004). Automatic affective evaluation does not automatically predispose for arm flexion and extension. *Emotion, 4,* 156–172.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience, 7*(4), 413-422.

Schwarz, N. (2010). Meaning in context: Metacognitive experiences. In B. Mesquita, L. F. Barrett, & E. R. Smith (eds.), *The mind in context* (pp. 105-125). New York, NY: Guilford Press.

Seibt, B., Neumann, R., Nussinson, R., & Strack, F. (2008). Movement direction or change in distance? Self- and object-related approach–avoidance motions. *Journal of Experimental Social Psychology, 44,* 713–720.

Titchener, E.B. (1915). *A beginner's psychology*. New York, NY: Macmillan.

Topolinski, S., Likowski, K. U., Weyers, P., & Strack, F. (2009). The face of fluency: Semantic coherence automatically elicits a specific pattern of facial muscle reactions. *Cognition and Emotion, 23,* 260–271.

Von Helversen, B., Gendolla, G. H., Winkielman, P., & Schmidt, R. E. (2008). Exploring the hardship of ease: Subjective and objective effort in the ease-of-processing paradigm. *Motivation and Emotion*, *32*(1), 1-10.

Winkielman, P., & Berridge, K. C. (2003). Irrational wanting and subrational liking: How rudimentary motivational and affective processes shape preferences and choices. *Political Psychology, 24,* 657–680.

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology*, *81*(6), 989-1000.

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.

Zajonc, R.B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science, 10*, 224–228.