Title
Texture Preference, Facial Attractiveness, and the Effect of Race on Lightness Perception

Author
Stephens, Kyle David

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Texture Preference, Facial Attractiveness, and the Effect of Race on Lightness Perception

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Kyle David Stephens

Dissertation Committee:
Professor Donald Hoffman, Chair
Professor Charlie Chubb
Professor Gregory Hickok

2015

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Professor Don Hoffman, who continually encouraged my scholarly development through his own excitement for groundbreaking research.

I would also like to express my gratitude to my committee members, Charlie Chubb and Greg Hickok, and to the members of my advancement committee, Barbara Sarnecka, Jennifer Trueblood and Penelope Maddy, who all provided insight and guidance during my time at UCI.

I extend special thanks to the non-UCI-affiliated researchers who have helped along the way: Stefan Thumfart and Edwin Lughofer for sharing information about their computational models, detailed in Chapter 1; Logan Trujillo, Jessica Jankowitsch, and Judith Langlois for kindly sharing the data used in Chapter 2, Logan in particular for helpful feedback and email correspondence; Amanda Montoya and Andrew Hayes for allowing me to beta test their mediation software, used in Chapter 2; Daniel Levin and Mahzarin Banaji for sharing stimuli, used in Chapter 3; and Gary Lupyan for email correspondence related to Chapter 3, in particular for guiding me to Firestone and Scholl's (then unpublished) 2015 manuscript.

I also thank Daniel Stehr for his help creating the stimuli used in Chapter 3; lab mates Darren Peshek, Justin Mark, and Brian Marion for helpful advice and feedback; and undergraduate research assistants Faizan Memon, Tatiana Dannenbaum, Jonathan Kramer-Feldman, Angela Babadjanian, Roberto Melendez, Ema Bidiwala, Lilit Barsegyan, Melanie Adamian, Emily Wong, and Hussein Aldahan for assistance running participants and being independent raters for work in Chapters 1 and 3. Additionally, I would like to thank Matt Inverso and Marcus Cole for help with pilot versions of the experiments in Chapter 3 and as independent raters.

I thank the members of the writing workshop at UCI for helping me hone my writing skills, Barbara Sarnecka in particular for creating and leading the workshop.

I thank the Department of Cognitive Science and the School of Social Science for their support, both financial and otherwise. I thank Clara Schultheiss and John Sommerhauser for being particularly good at their administrative duties and I thank Ruth Kim for running the UCI Human Subjects Lab.

CURRICULUM VITAE

Kyle D. Stephens

kdstephe@uci.edu

EDUCATION

University of California, Irvine
Department of Cognitive Sciences
Ph.D., Psychology — Fall, 2015
M.A., Psychology — Fall, 2014
Advisor: Donald D. Hoffman

Northeastern University
B.S., Chemical Engineering — Spring, 2009
B.S., Mathematics — Spring, 2009

PUBLICATIONS

Stephens, K.D. and Hoffman, D.D. (invited revision to *Perception*). On visual texture preference: Can an ecological theory explain why people like some textures more than others?

Stephens, K.D., Stehr, D.A., and Hoffman, D.D. (submitted). Changing the perceived race of a face does not change how light its skin looks.

Stephens, K.D. and Hoffman, D.D. (in preparation). Why Are Averaged Faces Attractive? A Mediation Model.

Achyuta, A.K.H, **Stephens, K.D.**, Pryce-Lewis, H.G., and Murthy, S.K. (2010). Mitigation of reactive human cell adhesion on poly(dimethylsiloxane) by immobilized trypsin. *Langmuir, 26(6)*, 4160-4167.

PRESENTATIONS

Stephens, K.D. (2014, February). The effect of race on the perceived lightness of faces. Experimental Social Science Graduate Student Workshop, University of California, Irvine.

Stephens, K.D. (2012, May). When does natural selection favor veridical perceptions? Cognitive Sciences Colloquium, University of California, Irvine.

Stephens, K.D. (2008, April). Vapor deposited polymer coatings for implantable neuroprosthetic devices. American Institute of Chemical Engineers Northeast Regional Student Conference.

Stephens, K.D. (2008, November). Vapor deposited polymer coatings for implantable neuroprosthetic devices. American Institute of Chemical Engineers National Student Conference.

ACADEMIC HONORS, AWARDS, & FELLOWSHIPS

University of California Associate Dean's fellowship	Jan., 2015
University of California graduate research fellowship	June, 2011
summa cum laude	June, 2009
Sears B. Condit Award: top 100 university-wide GPA's	April, 2009
Calvin S. Cronin Award: Excellence in written and oral communication	April, 2009
American Institute of Chemical Engineers National Student Design Competition honorable mention (2nd place)	Oct., 2009
American Institute of Chemical Engineers National Student Paper Competition honorable mention (4th place)	Nov., 2008
American Institute of Chemical Engineers Northeast Regional Student Paper Competition winner	April, 2008
IGERT Nanomedicine undergraduate research fellowship	Sept., 2007

SELECTED EMPLOYMENT

Consultant, VF Corporation **June-Aug., 2014, 2015**
Developed, coded, and analyzed experiments for research with human subjects. Provided consulting advice for the development of new design features.

Teaching Assistant, University of California, Irvine **Sept. 2010-May 2015**
Lead discussion sections, hosted office hours, graded papers, and sometimes lead lectures.

Research Assistant, Northeastern University **Sept. 2007-April 2009**
Synthesized polymer thin films via chemical vapor deposition for use as coatings of neuroprosthetic devices. Also cultured fibroblasts, glial cells and primary neurons for experimentation. Supervised by Shashi K. Murthy and Daniel D. Burkey.

Research Assistant, Massachusetts Institute of Technology **Jan. 2007-June 2007**
Full time position. Synthesized polymer thin films via chemical vapor deposition for use as coatings of optodes for drug screening. Supervised by Karen K. Gleason

TEACHING ASSISTIANTSHIPS

Psych 120P: Personality Theory	(V. Mann)	Spring, 2015
Psych 130A: Sensation & Percept	(V. Richards)	Fall, 2014
Psych 10C: Prob and Stats III	(M. Liljeholm)	Spring, 2014
Psych 10B: Prob and Stats II	(J. Trueblood)	Winter, 2014
Psych 112A: Experimental Psych	(B. Miller)	Fall, 2013
Psych 10C: Prob and Stats III	(J. Hagedorn)	Spring, 2013
Psych 10B: Prob and Stats II	(J. Trueblood)	Winter, 2013
Psych 112A: Experimental Psych	(K. Saberi)	Fall, 2012
Psych 7A: Intro to Psychology	(C. Lofgren)	Spring, 2012
LPS 30: Symbolic Logic	(K. Wehmeier)	Winter, 2012
Psych 140L: Learning Theory	(C. Chubb)	Fall, 2011
SocSci 3A: Computer Research	(J. Sobaugh)	Spring, 2011
Psych 9B: Psych Fundamental	(M. Steyvers)	Winter, 2011
Psych 7A: Intro to Psychology	(J. Hagedorn)	Fall, 2010

FREQUENTLY USED SOFTWARE

Matlab	Coding experiments, analyzing data, running simulations
R	Statistical analysis, analyzing data, creating figures
SPSS	Statistical analysis, mediation analysis
Python	Running simulations
Adobe Photoshop	Creating stimuli for experiments
Adobe Illustrator	Creating stimuli for experiments
FantaMorph	Morphing face images together (experimental stimuli)
G*Power	Conducting statistical power analysis

ABSTRACT OF THE DISSERTATION

Texture Preference, Facial Attractiveness, and the Effect of Race on Lightness Perception

By

Kyle David Stephens

Doctor of Philosophy in Psychology

University of California, Irvine, 2015

Professor Donald Hoffman, Chair

It's uncontroversial to point out that people have preferences and these preferences can vary considerably from one person to the next. My niece loves the color yellow; my grandmother hates the color yellow. But, it's more interesting to point at how consistent preferences can be. Almost everyone likes certain shades of blue. In the first part of this dissertation, I examine people's preferences in two domains.

First, I examine how well people's preferences can be predicted for a certain kind of images known as visual textures (Chapter 1). I find that people's preferences can be explained well by an ecological model according to which people like visual textures to the degree that they like the objects most-associated with those textures.

Next, I examine people's preferences for faces (Chapter 2). One of the most robust findings in face research is that people rate faces with average configurations as highly attractive. Despite the consistency of this finding, we still don't know why this should be true. In Chapter 2, I use a statistical mediation model to investigate why averaged faces are so attractive. I find that the result is not explained by any of the mediators tested and argue that averageness *per se* is attractive.

Finally, I pivot in the third chapter: Instead of examining the surprising consistency of high-level experience (preference ratings), I examine the potential idiosyncrasy of low-level experience. Low-level experiences – of things like color and lightness – are thought to be relatively consistent across people. But, recently, researchers have claimed that low-level perception can be influenced by idiosyncratic cognitive factors like beliefs or desires. In the third chapter, I investigate whether the perceived race of a face (a high-level cognitive construct) can influence how light its skin looks (low-level lightness perception), regardless of how light its skin actually is. Contrary to previous claims, I do not find support for this assertion.

Overall I find that, on the one hand, people are surprisingly consistent in their preferences, both for visual textures and for faces. On the other hand, I find no evidence that idiosyncratic cognitive factors can affect low-level perception.

INTRODUCTION

It's uncontroversial to point out that people have preferences and these preferences can vary considerably from one person to the next. My niece loves the color yellow; my grandmother hates the color yellow. But, it's more interesting to point at how consistent preferences can be. Almost everyone likes certain hues of blue (Yokosawa, Yano, Schloss, Prado-Leon, & Palmer 2010). In the first part of this dissertation, I examine people's preferences in two domains.

First, I examine how well people's preferences can be predicted for a certain kind of images known as visual textures (Chapter 1). Visual textures are images that are not uniform colors, but are also not scenes with decipherable objects; they're in between. I find that people's preferences can be explained well by the *ecological valence theory* according to which people like basic stimuli to the degree that they like the objects or entities most-associated with those stimuli (Palmer & Schloss, 2010). The ecological valence theory has already been used successfully to account for human color preferences (Palmer & Schloss, 2010; Taylor & Franklin, 2012) and human odor preferences (Schloss, Goldberger, Palmer, & Levitan, 2015). In Chapter 1, I extend the theory to visual texture preference and compare its performance to that of more traditional texture-preference models based on computational features. Overall, I find that the ecological model does reasonably well – explaining 63% of the variance in people's preference ratings – considering its low complexity.

Next, I examine people's preferences for faces (Chapter 2). One of the most robust findings in face research is that people rate faces with average configurations as highly attractive (see, e.g., Rubenstein, Langlois, & Roggman, 2002). Despite the consistency of this finding, we still don't know why this should be true. In Chapter 2, I use a statistical mediation model to investigate why averaged faces aree so attractive. I conclude that the effect is not mediated by skin quality, perceived youthfulness, age, sexual dimorphism, processing fluency, or general familiarity, and I argue that averageness *per se* is attractive. I discuss why we might expect such a result because of human sexual selection or as a by-product of the way our brains solve certain information processing tasks.

Finally, I pivot in the third chapter: Instead of examining the surprising consistency of high-level experience (preference ratings), I examine the potential idiosyncrasy of low-level experience. Low-level experiences – of things like color and lightness – are thought to be relatively consistent across people. But, recently, researchers have claimed that low-level perception can be influenced by idiosyncratic cognitive factors like beliefs or desires (see, e.g. Colins & Olson, 2014; Dunning & Balcetis, 2013). The best affirmative evidence for this comes from a study of lightness perception: Levin and Banaji (2006) showed that the race of a face (a high-level cognitive construct) can affect how light its skin looks (low-level perception) regardless of how light its skin actually is. In the third chapter, I test this claim with carefully controlled stimuli, taking special care to obscure the research hypothesis from participants. I find no support that the perceived race of a face affects how light its skin looks and argue that the original results were due to participant response bias.

Chapter 1

On visual texture preference: Can an ecological theory explain why people like some textures more than others?

Kyle D. Stephens and Donald D. Hoffman (invited revision to *Perception*).

Abstract

What visual textures do people like and why? Here we test whether Palmer and Schloss' (2010) *ecological valence theory* can predict people's preferences for visual texture. According to the theory, people should like visual textures associated with positive objects or entities and dislike visual textures associated with negative objects or entities. We compare the results for the ecological model to more traditional texture-preference models based on computational features – namely, the model of Thumfart et al. (2011) – and find that the ecological model performs reasonably well considering its lower complexity, explaining 63% of the variance in the human preference data.

Introduction

For members of the species *H. sapiens,* textures are ubiquitous. Every day we are surrounded by objects and surfaces that give us unique feelings when we touch them – fur feels soft, tree bark feels rough, and silk feels smooth. These different feelings arise from variations on the surfaces of objects which also lead to variations in the pattern of light that reaches our eyes – fur *looks* soft, tree bark *looks* rough, and silk *looks* smooth. We call this visual aspect of surface variation *visual texture*[1]. Artists and designers have long used visual texture as a tool to evoke emotions and set moods (see, e.g., Brodatz, 1966; Gatto, Porter, & Selleck, 1999). In the current paper, we investigate the aesthetics of visual texture more formally by empirically exploring the question: What visual textures do people like and why? We examine whether the ecological valence theory proposed by Palmer and Schloss (2010) for color preferences – and later extended to odor preferences (Schloss, Goldberger, Palmer, & Levitan, 2015) – can also be used to explain preferences for visual texture. But, first we review previous work on visual texture and the aesthetics of images.

There are many models that predict perceptual properties of visual textures. For example, there are models that predict how rough (Ho, Landy, & Maloney, 2006; Tamura, Mori, & Yamawaki, 1978), glossy (Anderson & Kim, 2009; Kim & Anderson, 2010; Montoyishi, Nishida, Sharan, & Adelson, 2007), or complex (Amadasun & King, 1989; Tamura et al., 1978) they look, as well as how they are segmented (Ben-Shahar, 2006; Julesz, 1962, 1981; Landy & Bergen, 1991; Malik & Perona, 1990; Portilla & Simoncelli, 2000; Rosenholtz, 2000; Tyler, 2004; Victor, 1988) or classified (e.g., Dong, Tao, Li, Ma, & Pu, 2015; Guo, Zhang, & Zhang, 2010; Haralik, Shanmugam, & Dinstein, 1973; Randen & Husy, 1999; Varma & Zisserman, 2005) by human observers (for reviews, see Bergen, 1991; Landy & Graham, 2004; Rosenholtz, 2015).

[1] More formally, if we think of an image as a 2-D array of pixels and we think of each pixel as a realization of a random variable on intensity, then visual textures are those images that have variation in pixel intensity at the limit of resolution but whose spatial covariance is relatively homogenous (i.e., moving a small window around the image does not significantly alter the statistics within the window; Haindl & Filip, 2013; Julesz, 1962; Portilla & Simoncelli, 2000), although other constraints are often employed and precise definitions vary by application (Haindl & Filip, 2013; Sebe & Lew, 2001; Tuceryan & Jain, 1998). Also note that visual texture can arise from variations that are not tactile (e.g., a birch table looks different from a mahogany table but they might feel identical).

There are also many models that predict aesthetics of images, including how much observers like the images (e.g., Chen, Sobue, & Huang, 2008; Kawamoto & Soen, 1993; S. Kim, E. Kim, Jeong, & J. Kim, 2006; Lee & Park, 2011; Um, Eum & Lee, 2002; Zhang et al., 2011; for reviews, see Joshi et al., 2011; Palmer, Schloss, & Sammartino, 2013). But these models are typically applications-focused – they were not designed to predict preferences for visual textures generally, nor do they necessarily extend well to the general space of visual textures (Thumfart et al., 2011).

There are, however, two extant models that predict preferences for visual textures specifically and they are quite similar to one another.

Current Computational Models of Visual Texture Aesthetics

Thumfart et al. (2011) and Liu, Lughofer, and Zeng (2015) have both recently published parametric models that link aesthetic properties[2] of visual textures to computational features measured from those textures. Specifically, they both use a hierarchical feed-forward model with 3 layers to predict aesthetic properties. In both models, the first layer contains primary/physical aesthetic properties (e.g., Cold-Warm, Smooth-Rough, Dark-Light), the second layer contains more aggregate aesthetic properties (e.g., Artificial-Natural, Simple-Complex, Inelegant-Elegant), and the third layer contains emotional aesthetic properties (of primary interest: Dislike-Like). Each layer feeds prediction for the next higher layer, so that the properties in the first layer are predicted using only computational features, but the properties in the third layer are predicted using computational features plus the aesthetic properties from the previous two layers.

---

[2] "Aesthetic properties" are operationalized as antonym pairs such as Warm-Cold, Hard-Soft, or Artificial-Natural. Thus, preference is defined by the antonym pair Dislike-Like.

Both of these models have been able to predict human preference (i.e., Dislike-Like) for visual texture with high accuracy (Liu et al. claim to account for as much as 99% of the variance in human preference ratings). However, these models employ a “see what sticks” approach which emphasizes model prediction over model understanding. For instance, Thumfart et al. characterize each visual texture using a total of 188 computational features and 27 aesthetic properties, a veritable grab-bag from the literature on computer vison, image processing and aesthetics. To the credit of Thumfart et al. and Liu et al., both models are more interpretable than black-box models such as neural networks, and Thumfart et al. increase interpretability by using an optimization measure that punishes complexity (defined by number of regression terms).

Still, these models were not constrained by any overarching theory dictating which features should be included and which should not, and it’s unclear why certain features are so important for predicting preference while others are not. For instance, Thumfart et al. found that how “premium,” “sophisticated” and “woodlike” a texture looks were all important for predicting how much someone likes that texture, whereas how “elegant” and “natural” it looks were not. Likewise, they found that the skewness contained in 12 circular ring segments of the Fourier power spectrum of a texture is important for predicting how much someone likes that texture, whereas the kurtosis or standard deviation of those same 12 ring segments is not. While these results are interesting and useful, they are not theoretically motivated and would not have been predicted by any current theory.

In other words, these models are good at predicting *what* visual textures people like, but they do not give the best sense of *why*.

Ecological Valence Theory

In contrast to the "see what sticks" computational models described above, Palmer and Schloss' (2010) ecological valence theory is based on a simple premise: People's preferences for basic stimuli (e.g., colors, odors) are derived from how they feel about the objects or entities associated with those stimuli. So, for example, a student might have a particular affinity for the color blue because they like blueberries and/or attend the University of California, Irvine (whose official color is blue), whereas another student might have an affinity for the color red because they like cherries and/or attend the University of Southern California (whose official color is red). Indeed, there is evidence that people like the colors of their own social group more than members of a rival social group (Schloss & Palmer, 2014; Schloss, Poggesi & Palmer, 2011).

But, Palmer and Schloss take this basic premise a step further and claim that people's preferences are dictated by a summary statistic that accounts for the valence across *all* objects associated with a particular stimulus. That is, how much a person might like a given stimulus can be explained by how positive or negative people feel, on average, about all objects and entities associated with that stimulus. The motivation for this reasoning is evolutionary: Assuming that people have positive feelings about objects that lead to beneficial outcomes, then stimuli that are typically associated with those objects serve as indicators to the beneficial outcomes. Thus, such preferences should steer organisms to approach objects that lead to beneficial outcomes and avoid those that lead to harmful outcomes, an evolutionarily advantageous strategy (Palmer & Schloss, 2010; Schloss et al., 2015). This is the ecological valence theory.

Palmer, Schloss and colleagues have already found support for this theory for colors and odors: Average color preferences are well predicted by how people feel about all objects associated with the colors (Palmer & Schloss, 2010; Taylor & Franklin, 2012), and average odor preferences are well predicted by how people feel about all objects associated with the odors (Schloss et al., 2015). Furthermore, the theory predicts that preferences across cultures or social groups should vary only to the degree that these groups associate different objects with the same stimuli or have different feelings about the same objects, and there is empirical support that this is the case with colors (Schloss, Hawthorne-Madell, & Palmer, in press; Yokosawa, Yano, Schloss, Prado-Leon, & Palmer, 2010).

Thus far, Palmer, Schloss and colleagues have only tested the ecological valence theory with colors and odors, but the theory should be generally applicable to different kinds of stimuli in any modality (see, e.g., Schloss et al., 2015). The goal of the present research was to test how well the ecological valence theory accounts for preferences for visual texture: Can people's average preferences for visual textures be predicted by how positive or negative they feel about all objects associated with the textures?

Present Study

Testing the ecological valence hypothesis. To test whether preference for a given visual texture can be predicted by the average valence for all objects or entities associated with that visual texture, we adapted the procedure of Palmer and Schloss (2010) to visual textures. This procedure required data from four different groups of participants.

The first group gave the to-be-predicted preference ratings. These participants were simply asked to rate how much they liked each visual texture on a line-mark ratings scale coded from -100 (dislike) to $+100$ (like).

The next three groups of participants were needed to calculate the predicted preference[3] for each texture: one group typed out descriptions of objects they associated with each visual texture (the object-description group); another group rated their emotional response (positive/negative) to a condensed set of these object descriptions (the description-valence group); and the final group rated how well each visual texture matched all of the descriptions ascribed to that texture (the texture-description match group). The prediction for any given visual texture was then given by the sum of the average valence ratings for each object associated with the texture, weighted by how well those objects match the texture:

$$P_t = \frac{1}{n_t}\sum_{d=1}^{n_t} w_{td} v_d \tag{1.1}$$

where, P_t is the predicted preference for texture t (-100 to 100), n_t is the number of descriptions ascribed to texture t, w_{td} is the average match weighting between texture t and description d (0 to 1), and v_d is the average valence rating for description d (-100 to 100). Using this formulation, Palmer and Schloss (2010) accounted for 80% of the variance in average color preferences, and Schloss et al. (2015) account for 76% of the variance in average odor preferences.

Testing other hypotheses. One might (reasonably) question whether applying this formulation to visual textures is sensible. After all, isn't it generally obvious what object a (natural) visual texture corresponds to? In this case, preference for a visual texture would boil down to how people feel about the single associated object, an uninteresting result.

---

[3] Palmer and Schloss (2010) and Schloss et al. (2015) refer to their preference predictions as weighted affective valence estimates, or WAVEs.

Schloss et al. (2015) faced a similar problem when applying the ecological valence theory to odors, since they used odor pens that were designed to smell like a particular object (e.g., *honey, apple, leather*). But, they found that the associated objects were more ambiguous than one would think and that odor preference was better explained by the valence for all objects associated with the odor pens, rather than just the object the pen was designed to smell like. We hypothesized a similar outcome for visual textures.

To test this outcome, we examined two additional hypotheses: the *single-associate hypothesis* and the *namesake hypothesis* (see also, Schloss et al., 2015). The single-associate hypothesis states that preference for a visual texture is best predicted by the valence of the single object that the texture is most associated with. The namesake hypothesis states that preference for a visual texture is best predicted by the valence of the namesake object that produced the visual texture. So, for example, one of the visual textures used was a close-up of a strawberry (see Figure 1.1). The ecological valence theory predicts that preference for this texture is best explained by the weighted average valence ratings for all descriptions ascribed to this texture (e.g., 'beans', 'flowers', 'eggs', 'peacock tail', 'seeds', 'strawberry', 'insect eyes', etc.); the single-associate hypothesis predicts that preference for this texture is best explained by the average valence of the description 'eggs,' since this was the description most often associated with the texture; and the namesake hypothesis predicts that preference for this texture is best explained by the average valence of the description 'strawberry,' since this is the namesake object that the texture was produced from.

Comparing to the computational models. After testing the single-associate and namesake hypotheses, we also compared the results of the ecological valence model to the computational model of Thumfart et al. (2011). First, we compared the error of the ecological model to the error of Thumfart et al.'s model using an error term that punishes the model for complexity (see Equation 1.2 in results section). Then, we tested Thumfart et al.'s model more directly.

Thumfart et al. found that texture preference was best predicted by six factors: (1) how "premium" a texture looks, (2) how "sophisticated" a texture looks, (3) how "rough" a texture looks, (4) how "woodlike" a texture looks, (5) the skewness of the distribution of Fourier energy in concentric rings in the Fourier space of the texture (*cSkew*), and (6) the *strength* of the texture, a computational feature derived from the neighborhood graytone difference matrix. Thus, we had participants rate textures on line-mark rating scales for each of the antonym pairs described above (NotPremium-Premium, NotSophisticated-Sophisticated, Smooth-Rough, and NotWood-Wood), and we calculated their *cSkew* and their *strength*. We recruited two groups of participants: one group rated the textures on the scales for premium, sophisticated, and rough, and the other group rated the textures on the scale for woodlike.

Materials and Methods

Participants

There were six separate groups of participants, all of whom were undergraduate students at the University of California, Irvine and received course credit for participation.

36 participants (19 male) took part in the texture-preference rating task, rating how much they liked each visual texture. The mean age was 20.8 years (range 18-28 years).

32 participants (15 male) took part in the object-description task, describing objects associated with each visual texture. Once all of the descriptions were consolidated (see Procedure section), one of the authors (K. Stephens) rated how well he thought the

descriptions matched the visual textures they were ascribed to on a 0-100 line-mark scale. We ended up excluding one female participant because 38 of her descriptions were given a 0 match-rating by this author (this was 4.5 standard deviations above the mean number of 0 match-ratings for the rest of the group). After excluding this participant, the mean age was 21.6 years (range 18-37 years).

37 participants (18 males) took part in the description-valence rating task, rating how positive or negative they felt about each object description. The mean age was 20.7 years (range 18-32 years).

27 (14 male) participants took part in the texture-description match rating task, rating how well each description matched the texture it was ascribed to. We excluded one male participant because he did not finish the experiment (we lost connection to our Matlab license three-quarters of the way through). After excluding this participant, the mean age was 22.5 years (range 18-40 years).

24 (11 male) participants took part in the first perceptual-property rating task, rating how “premium”, “sophisticated”, and “rough” the textures looked. The mean age was 21.1 years (range 18-36 years).

Finally, 22 (6 male) participants took part in the second perceptual-property rating task, rating how “woodlike” the textures looked. We excluded one female participant because her mean reaction time was more than 2 standard deviations faster than the mean for the group. After excluding this participant, the mean age was 20.7 years (range 18-34 years).

Stimuli

Texture-preference rating, object description, and perceptual-property rating. For these tasks the stimuli consisted of 62 different visual textures. 52 of these textures were from Brodatz's (1966) album, and 10 were from shutterstock.com. We used the same Brodatz textures as those used by Rao and Lohse (1993, 1996), excluding D43, D75, D102, and D108 (see tables in the appendix for a complete list). Of the Shutterstock images, eight were chosen specifically to have either high valence (chocolate, fresh lettuce, fresh strawberry, sea/ocean) or low valence (mud, mold, rotten strawberry, infected skin); the remaining two images were Gaussian blurred versions of the chocolate and the mud images respectively. All images were grayscale and viewed at 256-by-256 pixels (see Figure 1.1 for the Shutterstock images; see tables in the appendix for a complete list of all visual textures. See Brodatz, 1966 or Rao and Lohse, 1993 for pictures of the Brodatz textures).

Texture-description match rating. Because of the large number of descriptions for each texture (even after consolidation; see Procedure section), it would have taken too long to show participants all texture-description pairs for all 62 textures in the texture-description match rating task. Thus, we only included 47 of the original 62 textures in this task. We included all 10 of the shutterstock.com images, but only 37 of the original 52 Brodatz textures. We chose which 15 Brodatz textures to exclude (D11, D25, D29, D31, D39, D40, D42, D55, D63, D80, D82, D83, D89, D94, D101) based on their similarity to other textures already included (see table A1 in the appendix for a complete list of all visual textures used in this experiment).

Description-valence rating. For the description-valence rating task, the stimuli consisted of black text presented in 28-point Arial font.

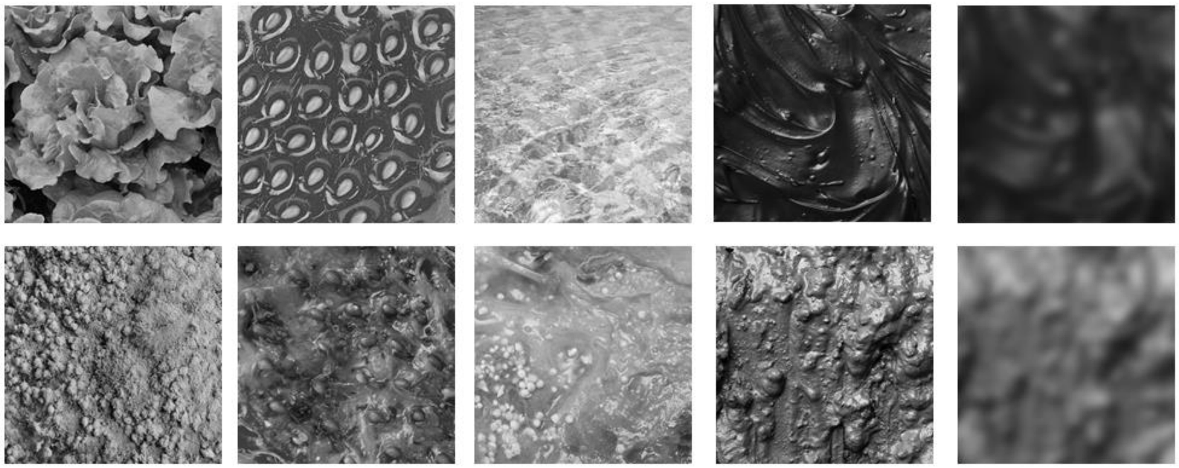


Figure 1.1. The ten visual textures from Shutterstock used for all tasks. The rightmost pictures are Gaussian-blurred versions of the preceding pictures. From top left to bottom right: lettuce, strawberry, tropical ocean, chocolate, Gaussian-blurred chocolate, mold, rotted strawberry, skin infection, mud, Gaussian-blurred mud.

Displays

Monitor for visual textures. For any task displaying visual textures (i.e., the texture-preference rating, object description, texture-object match rating, and perceptual-property rating tasks), stimuli were presented using an 27-inch Apple iMac display set at a resolution of 2560x1440 and a refresh rate of 60Hz. All stimuli were presented against a neutral gray background with luminance 28.8 cd/m^2 (as measured using Photo Research PR-670 spectroradiometer). The monitor was calibrated using an X-Rite ColorMunki Photo spectrophotometer. Participants were all run in the same room, one at a time.

Monitor for descriptions. For the task displaying only text (i.e., the description-valence rating task), stimuli were presented using Dell computers attached to 17-inch Dell LCD monitors set at a resolution of 1280x1024 and a refresh rate of 60Hz. The text stimuli were displayed against a white background. Participants were run in groups of one to five in a room containing 8 computers.

Line-mark rating scale. Five of the tasks – the texture-preference rating, description-valence rating, texture-description matching rating, and perceptual-property rating tasks – gathered ratings using a line-mark rating scale, located at the bottom of the display. This scale consisted of a dark-gray rectangular box demarcated into 10 equally spaced regions (5 below the center point and 5 above), with text above the scale indicating what rating was being performed (e.g., "How much do you like this texture?"), text at the end-points of the scale to indicate the extreme ratings (e.g., "not at all" vs. "very much"), and a black slider bar that participants could slide along the scale with the mouse to indicate their rating. Participants could either click and drag the slider bar or move the bar by pointing and clicking (i.e., without dragging). When displayed on the larger monitor (i.e., for the texture-preference rating, texture-object match rating, and perceptual-property rating tasks), the top edge of the question text was approximately 247 pixels below the bottom edge of the texture, the rating scale was 768 pixels wide by 72 pixels high, and the scale's top edge was approximately 41 pixels below the bottom edge of the question text. When displayed on the smaller monitor (i.e., for the description-preference rating task), the rating scale was made smaller – 384 pixels wide by 52 pixels high – in proportion to the smaller resolution of the screen. The black slider bar was always 6 pixels wide by the height of the scale (either 72 or 52 pixels), and all text was black, in 24-point Arial font. The scale and the question text were always horizontally centered. The code for all tasks using a line-mark rating scale was written in Matlab, using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007).

Description edit boxes. The object-description task gathered descriptions using edit boxes located below the textures. There were five edit boxes in total, where participants could click and enter text. The edit boxes were 200 pixels wide by 50 pixels high and were located approximately 137 pixels below the bottom edge of the texture. They were horizontally centered and evenly spaced approximately 123 pixels apart from one another. Below the edit boxes was a pushbutton labeled 'submit' where participants could click to input their responses and move on to the next trial. This pushbutton was 150 pixels wide by 50 pixels high, horizontally centered, and the top edge was approximately 96 pixels below the bottom edge of the edit boxes. The code for this task was written using Matlab's uicontrol functions.

Procedure

The present study included five between-subject tasks: (i) texture-preference rating, (ii) object description, (iii) description-valence rating, (iv) texture-description match rating, and (v) perceptual-property rating. For all tasks, there was a blank 500ms ISI in between trials, and participants sat approximately 60cm away from the monitor. At this viewing distance, the visual textures spanned approximately 5.7 degrees of visual angle.

Texture-preference rating task. On each trial, participants viewed one of the 62 visual textures, vertically and horizontally centered on a neutral gray background with the text "How much do you like this texture?" displayed below the texture. Participants indicated their preference using a line-mark rating scale labeled "not at all" on the left end (coded as -100) and "very much" on the right end (coded as +100). Participants used the mouse to indicate their preference. They pressed the spacebar to input their rating and move on to the next texture. The textures were displayed in random order and all participants gave ratings for all textures.

Object description task. Participants saw each texture individually against a neutral gray background and were instructed to type out as many objects as they could think of (up to five) that might describe or be associated with the given visual texture. They were instructed not to name adjectives (e.g., ‘heavy,’ ‘light,’ ‘furry’), abstract entities (e.g., ‘dreams,’ ‘hopes,’ ‘fears’), or objects that would not be known to other people (e.g., ‘my aunt’s blanket’). They were also told that the experimenters were interested in all associated objects, whether pleasant or unpleasant. They typed their responses into one of five edit boxes located below the texture on the screen and clicked a pushbutton labeled ‘submit’ to move on to the next texture once they were satisfied with their descriptions. They were not given a time-limit. The textures were displayed in random order and all participants gave descriptions for all 62 textures.

After combining exact repeats, we were left with 1,264 unique descriptions (from an initial total of 4,858). We compiled these descriptions into a single list of 206 items using a procedure similar to that of Palmer and Schloss (2010). First, we discarded items from the list if they: (i) were abstract concepts instead of objects (e.g., ‘anger’, ‘dream’, ‘joy’); (ii) were adjectives instead of objects (e.g., ‘big’, ‘bumpy’, ‘blurry’, ‘curvy’, ‘dark’, ‘heavy’); or (iii) were ambiguous or could describe many different things (e.g., ‘arrangement’, ‘game’, ‘layers’, ‘material’, ‘mixture’, ‘surface’, ‘painting’). We excluded 366 descriptions this way, leaving 898 items on the list. We then categorized the descriptions to reduce the number that needed to be rated in the description-preference rating and the texture-object match rating tasks. We combined descriptions into a single object category if they seemed to refer to the same object (e.g., ‘blanket’, ‘quilt’ and ‘piece of a blanket’; ‘beehive’, ‘bee’s nest’ and ‘bee cells’; ‘carpet’, ‘carpret [sic.]’ and ‘rug’) and, in certain cases, we combined descriptions into superordinate categories if there were several exemplars referring to the same type of object. In the latter case, exemplars were included in the category description (e.g., ‘bedding (covers, sheets, pillows)’; ‘bricks (brick wall, brick building, brick sidewalk)’; ‘reptile skin (snake, crocodile, lizard)’). Descriptions were further excluded if they were given by only one person for only one image and they didn’t fit into any other object categories. After this process, we ended up with the final list of 206 consolidated object descriptions.

Even after this consolidation, each texture still had, on average, 20.4 descriptions attributed to it. With this many texture-description pairings, the texture-description matching task would have been too long (even after pairing the textures down from 62 to 47), so two independent raters ran a preliminary version of the texture-description match rating task and we excluded any texture-descriptions pairings that were given by only one participant and assigned a 0 match rating by both independent raters. After this procedure, the textures had an average of 18.4 descriptions each.

Description-valence rating task. Following Palmer and Schloss (2010), participants were first given a list of 8 sample items (sunset, bananas, diarrhea, sidewalk, boogers/snot, wine (red), chalkboard, chocolate) to give them an idea of the range of descriptions they would see. They were then presented with each of the 206 consolidated descriptions in black text, horizontally and vertically centered against a white background. The procedure was the same as for the texture-preference rating task, but the instruction text was different. Below each description was the text, “What’s your emotional reaction to the object(s)/thing described above?” and participants indicated their rating using a line-mark rating scale labeled “negative” on the left end (coded as -100) and “positive” on the right end (coded as +100). The descriptions were displayed in random order and all participants gave ratings for all descriptions.

Texture-description match rating task. Because of the large number of texture-description pairings, we used only 47 of the original 62 visual textures in this task to keep it from being too long (see discussion in the Stimuli section for more details). On each trial, participants viewed one of the 47 textures along with one of the descriptions ascribed to that texture against a neutral gray background. The description was displayed in black, 28-point Arial font, horizontally centered, approximately 110 pixels above the top edge of the texture. Below each texture-description pairing was a line-mark rating scale with the text "How well does this description match the image?" The scale was labeled "very poorly" on the left end (coded as 0) and "very well" on the right end (coded as 1). Participants saw all possible parings of texture images and descriptions ascribed to those textures. We split the participants for this experiment into two groups: One group saw the same description for all textures given that description before moving on to the next description, and the other group saw the same texture for all descriptions given to that texture before moving on to the next texture. The order of the descriptions and the textures was randomized.

Perceptual-property rating tasks. Participants viewed each of the 62 visual textures, one at a time, against a neutral gray background with the text "This texture looks:" displayed below the texture. They indicated their rating for the requested perceptual property using a line-mark rating scale coded from -100 to +100.

For the first perceptual-property rating task, participants rated the textures on three different antonym pairs: NotPremium-Premium, NotSophisticated-Sophisticated, and Smooth-Rough. The experiment was blocked so that participants rated all textures on one perceptual property before moving on to the next, and they were given a short break between blocks. The order of the perceptual properties was randomized as was the order of the textures within a block.

The second perceptual-property rating experiment was identical to the first, except that participants rated the textures on only one antonym pair: NotWood-Wood.

Results and Discussion

We first describe participants' texture preference ratings for the 47 visual textures used in all tasks. We then test the ecological valence hypothesis for these textures, namely that texture preferences can be explained by the combined valence of all objects associated with the textures. Next, we test this account against the single-associate and namesake hypotheses. Finally, we fit the model of Thumfart et al. (2011) and compare the results of this model to those of the ecological valence model.

Texture preferences

We used Cronbach's coefficient alpha (Cronbach, 1951; see also, Ritter, 2010) to assess interrater reliability for texture-preference ratings. Reliability was high (0.83), so we averaged across participants to get mean preference ratings. Figure 1.2 shows the mean preference ratings for the 47 visual textures used in all experiments (see table A2 in the appendix for preference ratings for the other 15 visual textures used in this experiment). The most preferred textures included clouds, lettuce, lace, and tropical sea. The least preferred textures were the two blurred images, mud, and rotted strawberry.

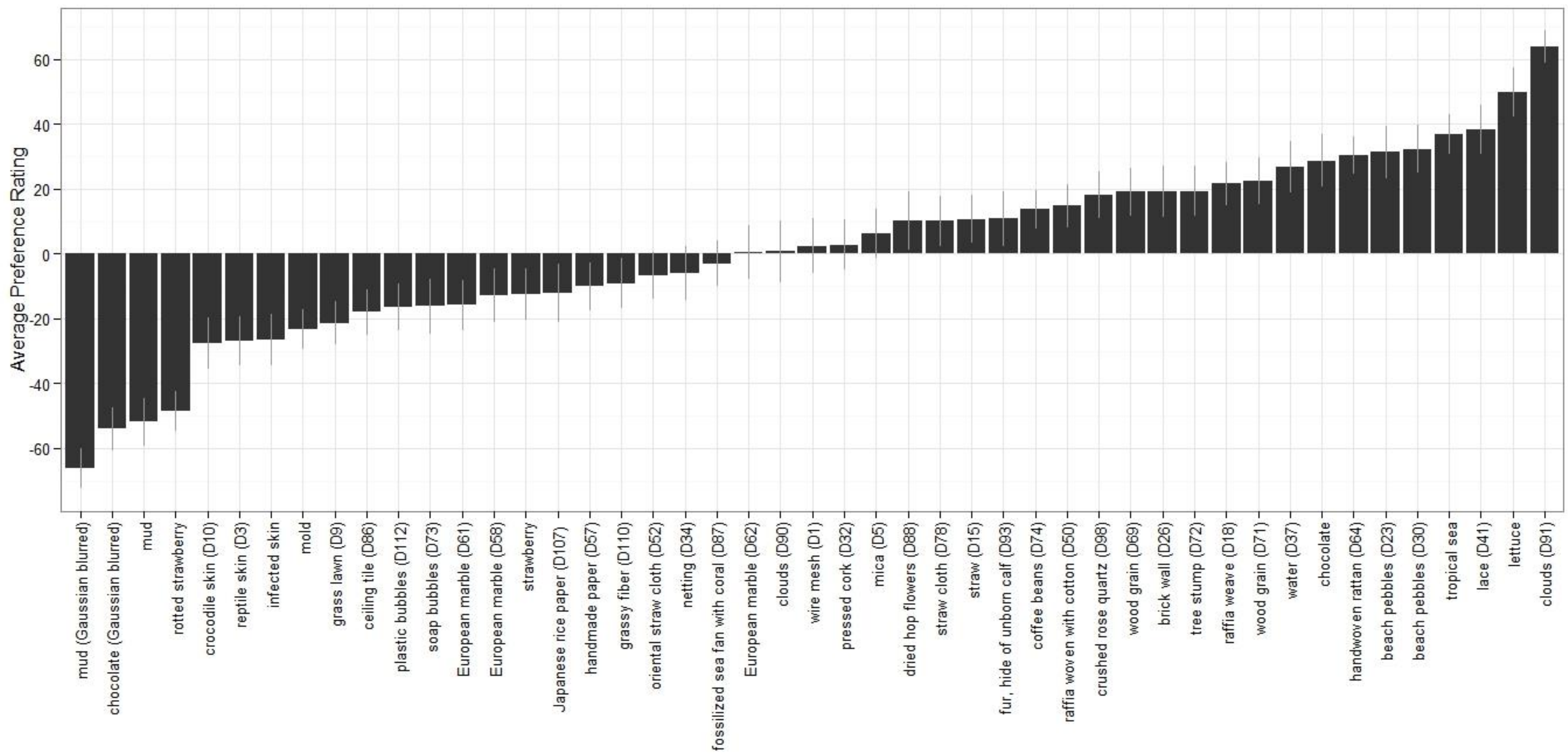


Figure 1.2. Average texture-preference ratings (***y***-axis) for each of 47 visual textures (***x***-axis). For Brodatz textures, the Brodatz's (1966) identifier is given in parentheses, and the labels are his descriptions. For the other textures, the labels come from the Shutterstock search terms that produced the images. Error bars represent the standard errors of the means.

Testing the ecological account of texture preferences

We used Cronbach's coefficient alpha to assess interrater reliability separately for the groups that rated description-valence and description-texture matching. Reliability was high for these groups (0.96, 0.99), so we averaged across participants to obtain mean valence-ratings for each description (v_d) and mean match-ratings[4] for each texture-description pairing (w_{td}). Using these ratings along with the consolidated descriptions from the object description task, we were able to predict preference ratings for the each texture using Equation 1.1.

Figure 1.3 shows the strong positive correlation ($r = 0.79, p < .001$) between the measured texture preferences, averaged across participants, and the preferences as predicted by Equation 1.1. This indicates that people like textures that remind them of positive objects and dislike textures that remind them of negative objects, and it supports the ecological valence hypothesis. It is also worth noting that the prediction equation accounted for less variance if we either dropped the normalizing term $1/n_t$ ($r = 0.66$), or set all of the non-zero weights (w_{td}'s) to 1 ($r = .72$).

[4] Recall that we split the participants in the texture-description match rating task into two groups: both groups saw all pairings of textures and descriptions ascribed to those textures, but one group saw the same description for all textures given that description before moving on to the next description, and the other group saw the same texture for all descriptions given to that texture before moving on to the next texture. The latter group went significantly slower (mean difference in reaction time per trial $= 2.05\ sec$, $t(24) = 10.36,\ p < .001$), but, as Cronbach's alpha indicates, the groups' ratings were in strong agreement, and using data from either group led to the same correlation coefficient.

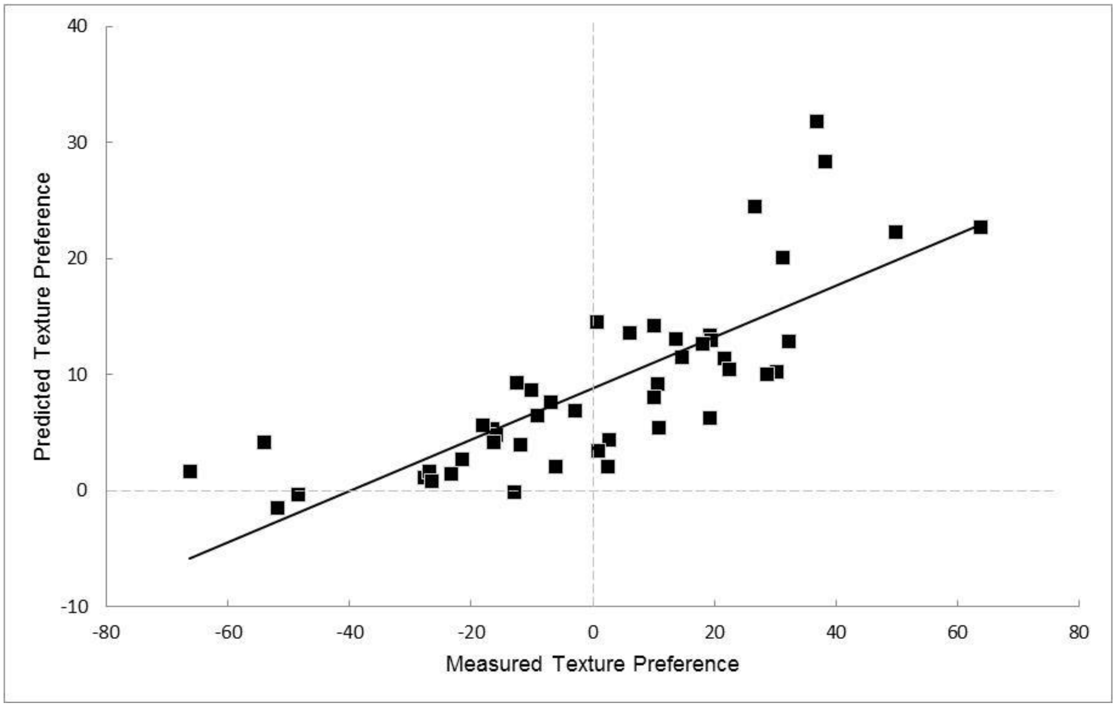


Figure 1.3. For a set of 47 visual textures, shows the correlation ($r = 0.79, p < .001$) between measured preferences, averaged across participants, and the preferences predicted by Equation 1.1. Numbers represent ratings on a line-mark rating scale, where -100 indicates that the texture is liked "not at all" and +100 indicates that it is liked "very much."

The main problem with the predicted ratings is that their range is compressed compared to the measured ratings and that most of the predicted ratings are positive while many of the measured ratings are negative. Schloss and colleagues reported similar issues when predicting color preferences (Palmer & Schloss, 2010) and odor preferences (Schloss et al., 2015). To explain why the model underpredicts negative ratings, they hypothesized that participants in the object-description group either underreport negative objects because they are too shy to report gross or disgusting things, or they are just biased toward generating/thinking about positive objects. To explain why the ratings for the prediction are compressed relative to the measured ratings, we posit that people likely have weaker feelings about text describing objects than they have about actual images.

Despite these failings, the model does reasonably well: The predictions account for 62.5% of the variance in measured preference with a single predictor.

Evaluating the single-associate hypothesis

We tested the single-associate hypothesis by first identifying the object named most frequently for each visual texture (see table A1 in the Appendix). We then calculated the correlation between average preference ratings for the 47 visual textures and average valence ratings for the most frequently named objects. Although there was a significant positive correlation ($r = 0.47, p < .001$), it was weaker than the relation between measured preferences and preferences predicted by Equation 1.1 (correlations compared using the Fisher r-to- z transformation: $z = 2.63, p = .01$). Therefore, texture preferences were better predicted by a summary statistic of valence for all objects associated with a texture than by valence for the single object most frequently associated with a texture.

It is also worth noting that, although the valence of the most frequent description did not predict texture preference better than the valence of all associated descriptions, texture preference was negatively correlated with the number of descriptions ($r = -0.67, p < .001$), so that the textures with fewer descriptions were better liked. Taylor and Franklin (2012) found this to be the case with colors as well, but Schloss et al. (2015) did not find this with odors. On average, each texture had 18.4 descriptions associated with it (range = 6-34).

Evaluating the namesake hypothesis

We tested the namesake hypothesis by first identifying the "namesake" object for each visual texture, that is, the object from which the texture image was created. We then calculated the correlation between average preference ratings for the visual textures and average valence ratings for the namesake objects.

To identify the namesake objects, we chose the consolidated description that was closest either to Brodatz' (1966) description (for the Brodatz textures) or to the Shutterstock search term that produced the image (for the Shutterstock textures). The matches were not always exact, but were reasonably close. For example, Brodatz described texture D37 as "water," but we used "ocean/sea" as the namesake object. Similarly, one of the Shutterstock textures was produced from the search term "rotten strawberry," but we used "rotten/spoiled food" as the namesake object (see table A1 in the appendix for more details). We excluded texture D32 because nothing close to Brodatz' description for the texture (pressed cork/corkboard) was ever given as an object description and thus D32's namesake was never given a valence rating. We also excluded the two Gaussian blurred visual textures because the namesake objects for these textures would be too difficult to visually determine. Thus, overall, the correlation was based on ratings for 44 of the visual textures.

Although there was a significant positive correlation ($r = 0.60, p < .001$), it was weaker than the relation between measured preferences and preferences predicted by Equation 1.1 (with the 44 textures: $r = 0.82, p < .001$; correlations compared using the Fisher r-to- z transformation: $z = 2.10, p = .04$). Thus texture preferences were better predicted by the average valence over all objects associated with that texture than by the valence of the namesake object that produced the texture.

Comparing to the computational models of Thumfart et al. and Liu et al.

While the ecological model accounted for 62.5% of the variance in texture preference ratings, the accuracy of this prediction was not nearly as good as that provided by the computational models of Thumfart et al. (2011) and Liu et al. (2015), the latter of which claimed to account for 99% of the variance in human preference ratings. However, the ecological model utilized only a single predictor, whereas the models of Thumfart et al. and Liu et al. utilized 6 and 5 predictors respectively. Since Thumfart et al. optimized their model using an error measure that punished the model for complexity[5], we were able to compare the results of the ecological model to the results of Thumfart et al.'s model, taking into account the lower complexity of the ecological model. The punished error measure was a convex combination of the following form (see Thumfart et al., 2011, Equation 2):

$$Error_{punished} = \alpha \times Accuracy + (1 - \alpha) \times Complexity \tag{1.2}$$

with α in the range [0,1]. Thus, the parameter α controls the ratio of accuracy versus complexity, with lower complexity emphasized for decreasing values of α. Using this formulation, we found that the punished error for the ecological model matched that of Thumfart et al.'s model when $\alpha \approx 0.35$. Thus, although the ecological model is less accurate than Thumfart et al.'s model, it has comparable performance when the modeler values low complexity over high accuracy. We were unable to make such a comparison with Liu et al.'s model since they did not work with a punished error term.

We also tested Thumfart et al.'s model more directly by measuring the six factors in their final texture-preference prediction equation: (1) how "premium" a texture looks, (2)

---

[5] For linear models like the ones we are discussing here, complexity refers to the number of predictors in the model. e.g., a complexity of 6 means the model has 6 predictors.

how “sophisticated” a texture looks, (3) how “rough” a texture looks, (4) how “woodlike” a texture looks, (5) the skewness of the distribution of Fourier energy in concentric rings in the Fourier space of a texture (*cSkew*), and (6) the *strength* of the texture, a computational feature derived from the neighborhood graytone difference matrix. We gathered ratings for “premium,” “sophisticated,” “rough” and “woodlike” in the perceptual-property ratings tasks, and we measured the computational features *cSkew* and *strength* following Thumfart et al. (2011) and Amadasun and King (1988) respectively. We used Cronbach’s coefficient alpha to assess interrater reliability separately for ratings of “premium”, “sophisticated”, “rough”, and “woodlike,” Reliability was generally high for all rating scales (0.76, 0.81, 0.75, and 0.93 respectively), so we averaged across participants and used the average ratings as predictors for each texture.

A multiple linear regression revealed that “premium” was the only significant predictor (Full Model: $F(6,40) = 12.2, p < .001$, multiple-$R^2 = 0.65$; Premium: $t(40) = 7.56, p < .001$). Interestingly though, how “premium” a texture looks accounted for 62.3% of the variance in texture preference ratings ($r = 0.79, p < .001$), almost identical to the variance accounted for by the ecological valence model. Furthermore, “premium” seemed to track better with preference than the ecological predictions since it was not compressed relative to preference, nor was it inflated in the positive direction (see Figure 1.2). We hypothesize that this is simply because participants in our experiment viewed Dislike-Like as synonymous with NotPremium-Premium.

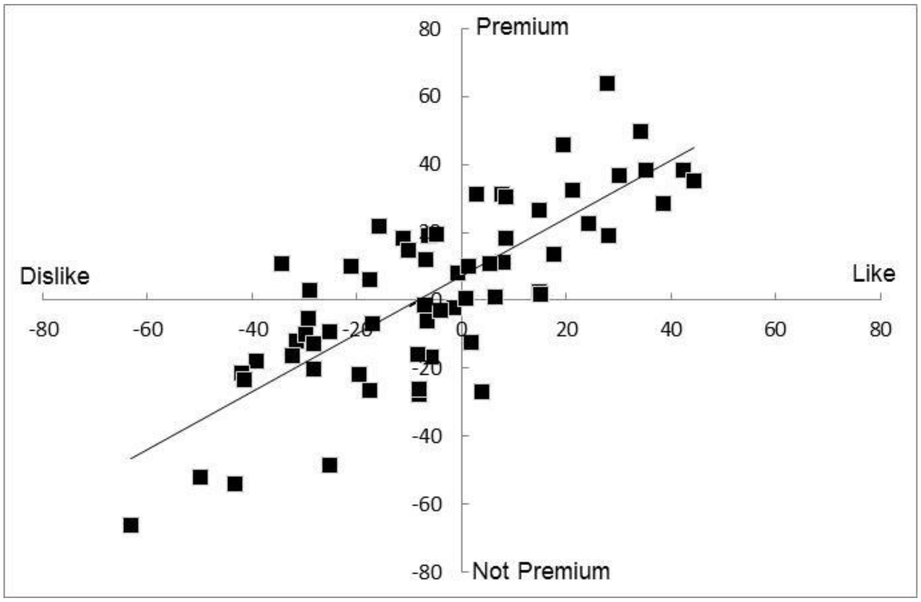


Figure 1.4. For a set of 47 visual textures, shows the correlation ($r = 0.79, p < .001$) between measured preferences and measured premiumness, averaged across participants. Numbers represent ratings on a line-mark rating scale, where -100 indicates that the texture is liked "not at all" or is "not premium" and +100 indicates that it is liked "very much" or is "premium."

We should also note that this is not necessarily a critique on Thumfart et al.'s model since they did not intend the model to be tested in this way. In personal communication, S. Thumfart indicated that they intended only to investigate the feature set as a whole (i.e., all 188 computational features + 27 aesthetic antonym pairs) and that any particular equation from the paper may not necessarily be applicable to other samples of textures.

Finally, we did not explicitly test the model of Liu et al. (2015), but we note that they found the aesthetic property Mussy-Harmonious to be an important predictor of preference, and we hypothesize that, as with NotPremium-Premium, Mussy-Harmonious may serve as a synonym for Dislike-Like (indeed, "preference" and "harmony" are often conflated in color research; see Palmer et al., 2013 for discussion).

General Discussion

The goal of the present study was to test whether the ecological valence theory – already used to explain color preferences (Palmer & Schloss, 2010) and odor preferences (Schloss et al., 2015) – could be extend to account for people's preferences for visual textures. The theory posits that preference for a given stimulus is determined by the combined valence of all objects associated with that stimulus, and that those preferences steer organisms to approach beneficial outcomes and avoid harmful ones (Palmer & Schloss, 2010; Schloss et al., 2015). We found some support for the theory for visual textures: The combined valence of all objects associated with a set of visual textures accounted for 62.5% of the variance in texture preference ratings, which was more than was explained by the valence of the single most-frequently associated object for each texture or by the valence of the single namesake object that produced each texture.

While this prediction accuracy is reasonable, the model did better explaining average preferences for colors (where it accounted for 80% of the variance; Palmer & Schloss, 2010) and odors (where it accounted for 76% of the variance; Schloss et al., 2015). At first blush this may seem curious, especially for colors: If the model states that people's preferences for a stimulus are determined by what objects are associated with that stimulus, then the model should perform better for visual textures than for colors because it should be more obvious what objects are associated with visual textures than with colors. However, the objects associated with each visual texture were not as obvious as one would think: Each texture had an average of 18 descriptions associated with it.

Furthermore, visual textures are not homogeneous fields like colors, and other factors such as composition, style or context likely play a role in determining preferences for visual texture (for review, see Palmer et al., 2013).

The prediction accuracy of the ecological valence model also pales in comparison to the computational models of Thumfart et al. (2011) and Liu et al. (2015) – the latter of which claims to account for 99% of the variance in texture preference ratings – although we found that the ecological valence model performs similarly to the Thumfart et al. model when one uses an optimization measure that emphasizes lower complexity (the ecological model has only 1 predictor; the Thumfart et al. model has 6). One problem with the computational models is their use of some aesthetic properties that participants may interpret as synonymous with preference (e.g., NotPremium-Premium for Thumfart et al., Mussy-Harmonious for Liu et al.). Still, for practical applications (e.g., in marketing or advertising), a computational model like Thumfart et al.'s or Liu et al.'s is preferable not only because of the increase in prediction accuracy, but also because the end result of such a model is able to predict texture preference for novel images based only on computational features measured from the images (i.e., with no human intervention). While the ecological valence model has low complexity in a statistical sense – it only uses one predictor – it has high complexity in a pragmatic sense – computing that single predictor requires a lot of experimentation.

The real benefit of the ecological valence theory is as a theoretical tool. The computational model of Liu et al. (2015) may be able to account for 99% of the variance in human texture-preference, but it utilizes predictors that have no theoretical significance. For example three of the predictors in their equation for texture preference come from a set of 20 features which were generated from a larger set of 106 computational features whose dimensionality was reduced using stochastic neighbor embedding. Because of the nature of this reduction technique, these features don't even have a clear physical interpretation (unlike the starting 106 computational features) let alone a compelling theoretical one. The ecological valence theory, on the other hand, stems from a basic, evolutionarily-inspired premise: An observer's preferences for low-level stimuli should be driven by the real-world objects those stimuli are most associated with, and how harmful or beneficial those objects are to the observer. This premise leads to an entire theory of preferences (i.e., the ecological valence theory) and generates testable predictions.

The main limitation of the present study was the range of visual textures used. We used only 63 visual textures (only 47 for prediction), most of which were Brodatz textures, and all of which were naturalistic and grayscale. This set of visual textures does not span the range of possibility, and it could be the case that the ecological valence model performs poorly for abstract, non-natural textures. Furthermore, it is known that adding color to a visual texture can affect a person's emotional response to that texture (e.g., Lucassen, Gevers, & Gijsenij, 2010). Any future study should utilize a more complete set of textures, perhaps sampling from a range of texture databases (as in, e.g., Thumfart, Heidl, Scharinger, & Eitzinger, 2009).

In sum, we conclude that – despite the limitations and although there are likely other mechanisms at play – it is reasonable to think of a person's preference for a visual texture as a summary statistic of how they feel about all the objects associated with that texture.

***

We have seen that people are consistent in their preferences for natural visual textures, and these preferences can be predicted well by an ecological model. We turn our attention now to preferences for faces. People consistently rate faces with averaged configurations to be highly attractive. In the following chapter we use a statistical mediation model to investigate why this should be the case.

Chapter 2

Why Are Averaged Faces Attractive? A Mediation Model

Kyle D. Stephens and Donald D. Hoffman (in preparation).

Abstract

It is a curious fact that taking the mathematical average of a set of faces produces a composite face that is highly attractive (often, more attractive than any of the starting faces). It is also a curious fact that researchers still don't know exactly why this should be the case. Current theory suggests that averaged faces are attractive because either: (1) averaged faces are more easily/fluently processed than individual faces, (2) averaged faces seem more generally familiar/less distinctive than individual faces, or (3) averageness *per se* is attractive because it indicates mate fitness or is a byproduct of the way our brains solve certain information processing tasks. In the current paper, we use data from Trujillo, Jankowitch and Langlois (2014) in a multiple mediator model to tease out the mechanism by which averageness affects attractiveness. We find that – at least with this dataset – the effect is not mediated by how fluently the faces are processed or by how distinctive the faces look. This lends support to theories claiming that averageness in itself is attractive.

Introduction

Consider a population of faces, any population (say, academics). Now, consider a face whose configuration of features is close to the average configuration for this population. It is an astonishing fact that this "average" face will be highly attractive, regardless of the underlying population. Francis Galton (1878) was the first to notice this – in an attempt to fashion a prototype for the 'criminal face,' he composited several portraits of criminals together and was surprised to discover that, not only did the resulting composite not look 'criminal,' it was *beautiful*. "The result is a very striking face, thoroughly ideal and artistic, and singularly beautiful. It is, indeed, most notable how beautiful all composites are" (Galton, 1907, p. 240).

Since Galton, several groups of researchers have confirmed this finding: When researchers morph faces together to create an averaged composite, the resulting face is highly attractive, often more attractive than any of the individual faces (Langlois & Roggman, 1990; Langlois, Roggman, & Musselman, 1994; Rhodes, Sumich, & Byatt, 1999; Rhodes, Harwood, Yoshikawa, Nishitani, & McLean, 2002). The more faces researchers add to the composite, the more attractive the composite gets (Langlois & Roggman, 1990; Langlois, et al., 1994; Rhodes et al., 1999, 2002), and faces can be made more or less attractive by increasing or decreasing their similarity to an averaged composite face (Rhodes et al., 1999; Rhodes & Tremewan, 1996). Researchers have also found that averageness predicts attractiveness in computational/morphometric models (Bronstad, Langlois, & Russell, 2008; Komori, Kawamura, & Ishihara, 2009a, 2009b; Valenzano, Mennucci, Tartarelli, & Cellerino, 2006; but see DeBruine, Jones, Unger, Little, & Feinbert, 2007; Said & Todorov, 2011). Averaged faces are found attractive across cultures (Apicella, Little, & Marlowe, 2007; Rhodes et al., 2002; Rhodes, et al., 2001a), and even in infants

(Rubenstein, Langlois, & Kalakanis, 1999). In other words, the attractiveness of averaged faces is well-established.[6] But, are average faces attractive *because* of their averageness?

We know that experimental control of averageness – by, for example, changing the number of faces included in a computer-averaged composite – systematically alters attractiveness. This is demonstrated in the simple path diagram in Figure 2.1 and suggests that averageness *precedes* attractiveness.

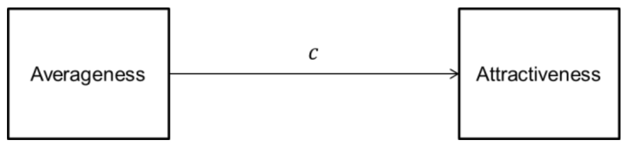


Figure 2.1. A path diagram demonstrating the *total effect* of averageness on attractiveness. Attractiveness is expected to change by c units overall given a 1 unit increase in averageness.

But, it is still unclear whether averageness *directly* influences attractiveness. It could be the case that averageness affects attractiveness through some mediating variable like skin quality or symmetry. Figure 2.2 shows a path diagram including the wide array of mediating variables that have been proposed in the literature.

In what follows, we will first argue that general familiarity and processing fluency are the only suggested mediators that haven't yet been accounted for. Then, we will use existing data (from Trujillo, Jankowitsch, & Langlois, 2014) and the statistical mediation techniques of Hayes and colleagues (Hayes, 2013; Montoya & Hayes, submitted) to test whether fluency or general familiarity can account for the influence of averageness on attractiveness for faces.

[6] However, we should note that this only seems to be true for faces whose identity is not previously known (see Halberstadt, Pecher, Zeelenberg, Wai, & Winkielman, 2013).

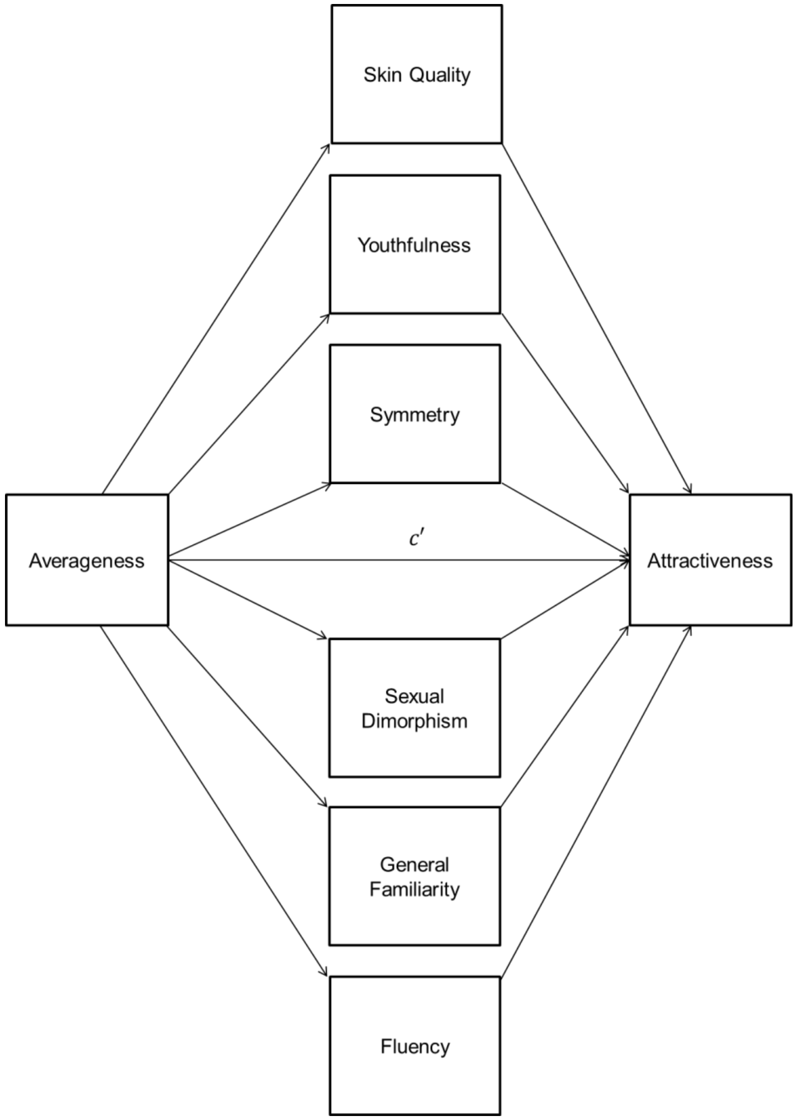


Figure 2.2. A path diagram for the effect of averageness on attractiveness, including all of the mediating variables that have been proposed in the literature. c' represents the *direct effect* of averageness on attractiveness given these mediators. That is, attractiveness is expected to change by c' units given a 1 unit increase in averageness with skin quality, youthfulness, symmetry, sexual dimorphism, general familiarity, and fluency held constant. The diagram has been drawn as strictly parallel mediation, but in principle the mediators could also affect each other (e.g., fluency could influence general familiarity).

Rounding up the mediators

Skin quality. One natural hypothesis is that averaged composite faces are attractive because the compositing process produces smooth, uniform skin-tones, free from blemishes. This fact alone could account for the boost in attractiveness. While it is true that skin quality is predictive of facial attractiveness (see, e.g., Jones, Little, Burt, & Perrett, 2004), this does not account for the attractiveness of computer-averaged faces.

Langlois and colleagues (Landglois & Roggman, 1990; Langlois et al., 1994; Rubenstein et al., 1999) smoothed and blurred individual faces to match any smoothing or blurring that occurs in averaged faces as a result of the averaging process, and found that this did not increase ratings of attractiveness. Langlois et al. (1994) found that averaging together different portraits from the same individual (which should also smoothen the skin) did not produce more attractive composites, unlike averaging together portraits from different individuals. Rhodes and Tremewan (1996) found that even line-drawn faces – which do not contain blurring or smoothing artifacts when composited – are more attractive in an average configuration. All of this suggests that skin quality does not mediate the effect of averageness on attractiveness.

Youthfulness. An alternative explanation is that the averaging process makes the faces look more youthful and hence more attractive (see Alley & Cunningham, 1991). Again, while it is true that ratings of attractiveness decline with age, especially for women (see, e.g., Alley, 1988; Deutsch, Zalenski, & Clark, 1986; Henss, 1991; Mathes, Brennan, Haugen, & Rice, 1985), this does not account for the attractiveness of computer-averaged faces because the faces used in averaging studies are typically *all* youthful. Furthermore, Langlois et al. (1994) found that, for the set of faces used in their original 1990 averaging study (Langlois & Roggman, 1990), attractiveness was not correlated with perceived age. In other words, youthfulness does not mediate the effect of averageness on attractiveness.

Symmetry. Perhaps the most tantalizing explanation for why averaged faces are attractive is that averaged faces also happen to be highly symmetrical and symmetric faces are attractive. This explanation is tantalizing because averaged faces do tend to have a high degree of bilateral symmetry (though they may still contain directional asymmetries; see Rhodes et al., 1999), and symmetry is linked to attractiveness in faces – both for naturally occurring faces (Grammer & Thornhill, 1994; Zebrowitz, Voinescu, & Collins, 1996), and for faces that are artificially (but carefully) made to be symmetric (Perrett et al., 1999; Rhodes, Proffitt, Grady, & Sumich, 1998; although effect sizes for symmetry tend to be smaller than the effect size for averageness or sexual dimorphism, see Brondstad, Langlois, & Russell, 2008; Rhodes, 2006).

However, while there is evidence that averageness and symmetry may independently influence attractiveness[7] (Komori et al., 2009a; Rhodes et al., 1999; Valentine, Darling, & Donnelly, 2004), the prevailing evidence suggests that symmetry does not significantly mediate the effect of averageness on attractiveness. In particular, Rhodes et al. (1999) found that averageness accounts for a significant amount of the variance in attractiveness even when symmetry is partialed out, and that manipulating the averageness of perfectly symmetrical faces still influences their attractiveness. Furthermore, Valentine et al. (2004) found that increasing the averageness of a face in profile view – where symmetry does not play a role – still increases its attractiveness. Taken together, these results suggest that symmetry, though itself important for attractiveness, does not significantly mediate the effect of averageness on attractiveness.

---

[7] For an argument that symmetric faces are attractive only because increasing symmetry increases averageness, see Enquist, Ghirlanda, Lundqvist, & Wachtmeister (2002)

Sexual dimorphism. Another possibility is that averaged faces are attractive because averaging makes the faces look more masculine or feminine which in turn makes them more attractive (note that computer-averaging likely makes all faces look more feminine since the averaging process tends to diminish masculine traits like coarse skin textures and square jaws; see Little & Hancock, 2002; Rhodes, 2006). While it's true that sexual dimorphism is an important component of facial attractiveness, particularly for female faces (for review, see Rhodes, 2006), it's not true that sexual dimorphism mediates the effect of averageness on attractiveness. Komori et al. (2009b) found that averageness and sexual dimorphism influence attractiveness independently. Lee, Mitchem, Writght, Martin, and Keller (2015) found that measures of sexual dimorphism (both objective and subjective) did not correlate with facial averageness and that controlling for the effect of sexual dimorphism did not significantly influence the correlation between averageness and attractiveness. Thus, sexual dimorphism does not mediate the effect of averageness on attractiveness.

General familiarity. Perhaps averaged faces are attractive because they seem more generally familiar. In this context, "general familiarity" refers to a subjective feeling of familiarity that researchers have defined as the degree to which a face resembles other faces in memory (Peskin & Newell, 2004; Vokey & Read, 1992). This has been distinguished from "episodic familiarity" for a face which is induced by prior exposure to that face (Peskin & Newell, 2004; Vokey & Read, 1992). A face that has never been seen before can still seem generally familiar; such faces are also referred to as looking “typical” or “not distinctive” (Peskin & Newell, 2004).[8]

Several studies have found that attractiveness is negatively correlated with how distinctive faces look and positively correlated with how typical they look or how generally familiar they seem (Langlois et al., 1994; Light, Hollander, & Kayra-Stuart, 1981; Peskin & Newell, 2004; Trujillo et al., 2014; Rhodes & Tremewan, 1996; Vokey & Read, 1992). But, there are conflicting reports about whether general familiarity mediates the effect of facial averageness on attractiveness: Halberstadt, Rhodes and Catty (2003) reported that the correlation between averageness and attractiveness remained when general familiarity was controlled, whereas Rhodes, Halberstadt, Jeffery, and Palermo (2005, p. 214-215) reported that the correlation between averageness and attractiveness was eliminated when general familiarity was controlled (although the later data-set was never formally published).

In the current paper, we will use Trujillo et al.’s (2014) ratings of attractiveness and distinctiveness – which, for present purposes, can be considered the converse of general familiarity (but see General Discussion) – for averaged and non-averaged faces to test whether distinctiveness mediates the effect of averageness on attractiveness.

Fluency. Finally, since averaged faces are close to the central tendency of facial configuration, they might simply be easier to process which makes them more pleasing (Langlois & Roggman, 1990, Trujillo et al., 2014).

---

[8] There is a lot of confusion in terminology in the literature. For instance, Rhodes et al. (1999) obtained ratings of how typical a face looks and refered to these as ratings of “averageness” (which they connoted as the converse of distinctiveness). Here, we reserve the term “averageness” for mathematical averages of faces obtained via computer manipulation; we refer to any ratings of how typical a face looks to ratings of general familiarity or distinctiveness.

In one regard, this is an uncontroversial position. It has long been known that, for many categories of objects, people prefer exemplars close to the central tendency for that category (so-called *prototypes*; see Hampton, 1997, 2006; Reed, 1972; Rosch, 1978; Rosch, Mervis, Gray, Johnson, & Boyes,-Braem, 1976). For example, people prefer prototypical (i.e., mathematically average) arrangements of dot patterns (Bomba & Siqueland, 1983; Winkielman, Halberstadt, Fazendeiro, & Catty, 2006), and they rate averaged configurations of birds, fish, cars, and, of course, faces as more attractive (Halberstadt & Rhodes, 2003; Langlois & Roggman, 1990).

But, only recently has processing fluency been offered as an explanation for why people should prefer prototypes. According to this account, people prefer prototypical stimuli not because of their averageness *per se* but because prototypical stimuli are processed fluently and fluent processing leads to positive affect (Harmon-Jones and Allen, 2001; Principe & Langlois, 2011, 2012; Winkielman and Caciopo, 2001; Winkielman et al., 2006). Presumably, this is because fluency indicates successful processing (Winkielman & Cacioppo, 2001; Winkielman et al., 2006).

There is some evidence for this fluent processing account for faces. In reaction time studies, attractive faces are correctly classified by gender (Hoss, Ramsey, Griffin, & Langlois, 2005) and as “human” vs. “chimpanzee” (Trujillo et al., 2014) faster than unattractive faces. In electroencephalographic (EEG) studies, the event-related potential evoked by faces – the N170 – is smaller when participants view averaged or individual attractive faces than when they view individual unattractive faces (Halit, de Haan, & Johnson, 2000; Trujillo et al., 2014). Taken together, this suggests that averaged and attractive faces are both processed more fluently than unattractive faces.

However, it's unclear from this data whether averaged faces are attractive *because* they are processed fluently, or whether averageness just also happens to be correlated with fluency. In the current paper, we will use Trujillo et al.'s (2014) fluency measurements and attractiveness ratings for averaged and non-averaged faces to test whether processing fluency mediates the effect of averageness on attractiveness.

The mediation model

Figure 2.1 represents the effect of averageness on attractiveness without considering any mediating variables. In this diagram, c represents the *total effect* of averageness on attractiveness, i.e., attractiveness is expected to change by c units overall given a 1 unit increase in averageness.[9]

We are interested in the process by which this effect operates which we can investigate using statistical mediation analysis (cf. Hayes, 2013). Using simple principles of linear modeling, mediation analysis can be used to quantify and test the pathways of influence from a causal variable to an outcome variable. The basic idea is that the presumed causal variable may not influence the outcome variable directly, but rather through one or more *mediator* variables: The causal variable affects the mediator variables, which in turn causally influence the outcome variable.

In the present paper, the causal variable is facial averageness and the outcome variable is facial attractiveness (see Figure 2.1). The discussion from the previous section suggests the effect of averageness on attractiveness might be mediated by one of the pathways depicted in Figure 2.3. In this diagram, the pathways $a_1 b_1$, $a_2 b_2$, and $a_1 a_3 b_2$

[9] Since, for the dataset analyzed here, averageness is a categorical variable (i.e., each face is either an individual face or an averaged composite of 32 individual faces – see Materials and Methods), a "1 unit increase in averageness" means going from a non-composite face (coded, for example, as averageness $= 1$) to a composite face (coded, for example, as averageness $= 2$). See Montoya and Hayes (submitted) for further discussion.

represent the *indirect effects* of averageness on attractiveness, and c' represents the *direct effect*. Thus, for example, attractiveness is expected to change by $a_1 b_1$ units via fluency given a 1 unit increase in averageness with distinctiveness held constant. Likewise, attractiveness is expected to change by c' units given a 1 unit increase in averageness with fluency and distinctiveness held constant. The pathway $a_1 a_3 b_2$ represents serial mediation in which the effect of averageness on attractiveness operates through more than one mediator, namely, averageness affects processing fluency, which in turn causally influences distinctiveness, which in turn causally influences attractiveness. We included this serial pathway because processing fluency is correlated with general familiarity but generally considered to be more fundamental (Winkielman & Cacioppo, 2001).

In analytical terms, the path diagrams (Figures 2.1 and 2.3) represent a set of linear equations and the values along the pathways are the coefficients of these equations, all of which can be estimated using ordinary least squares regressions (Hayes, 2013; Montoya & Hayes, submitted). It is worth noting that the values along the pathways in Figures 2.1 and 2.3 are all related by the equation: $c = c' + a_1 b_1 + a_2 b_2 + a_1 a_3 b_2$. For the specific regression equations that these path diagrams correspond to, see Equations 7 and 10-12 in Montoya and Hayes (submitted).

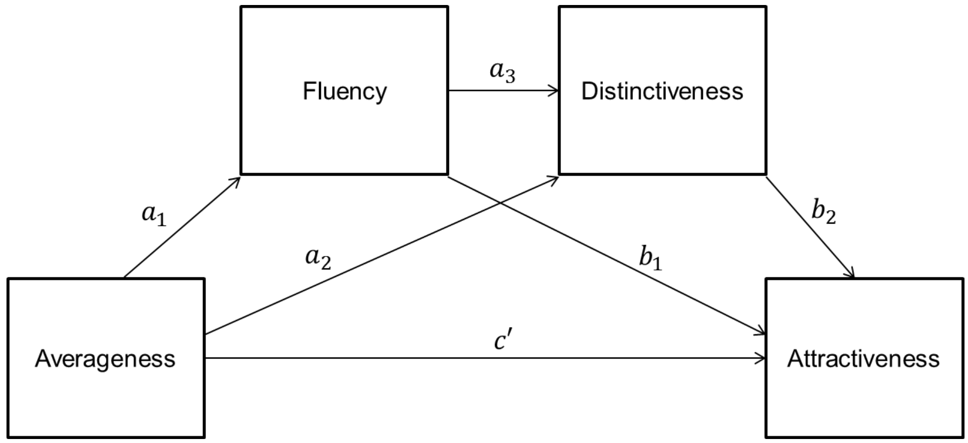


Figure 2.3. A path diagram for the effect of averageness on attractiveness with two mediators, fluency and distinctiveness. This is a serial mediation model.

Materials and Methods

The data from Trujillo et al. (2014)

We used data from Trujillo et al. (2014). The stimuli included three groups of faces (all of which were young-adult Caucasian females): high-attractive individual faces, low-attractive individual faces, and averaged composite faces. The groups of high- and low-attractive individual faces were created on the basis of a priori attractiveness ratings from 55 participants. The averaged faces were created by mathematically averaging the pixels between corresponding points on 32 individual faces.

The fluency data was obtained from another group of 55 participants (7 of which ended up being dropped due to technical problems with EEG) who participated in a species-categorization task. On each trial of this task, the participants viewed either one of the high-attractive individual faces, one of the low-attractive individual faces, one of the averaged composite faces, or a chimpanzee face and were instructed to indicate whether the face was "human" or "chimpanzee" as quickly and as accurately as possible. We used participant reaction time in this categorization task as our measure of processing fluency, noting that lower reaction-times lead to increased fluency.

After the categorization task, these participants also rated all three groups of faces for facial attractiveness and distinctiveness (defined as the ease of spotting a face in a crowd). We used these ratings as our measures of attractiveness and distinctiveness. See Trujillo et al. (2014) for more details on the subject make-up and specific procedures.

Model analysis

Considering the data from Trujillo et al. (2014) along with the path diagram in Figure 2.3, we first note that participants saw either non-averaged individual faces or averaged composite faces – that is, averageness was manipulated and not measured. Next, we note that all participants contributed measurements for both averaged and non-averaged faces on all measured variables (fluency, distinctiveness, and attractiveness). Thus, this data, along with the path diagram in Figure 2.3 is consonant with a two-condition within-participant serial multiple mediator model, as described by Montoya and Hayes (submitted; see their Figure 6). They provide an SPSS macro to analyze such a model – MEMORE (MEdeiation and MOderation analysis for REpeated measures designs) – and we used this macro to analyze the path diagram in Figure 2.3. Roughly, this involves comparing the measurements for fluency, distinctiveness and attractiveness for faces that are *not* averaged-composites against those that *are* averaged-composites (for more details, see Montoya and Hayes, submitted). We did this twice, once with low-attractive individual faces and once with high-attractive individual faces.

Results and Discussion

Results

The results of the mediation analysis are shown in Figures 2.4 and 2.5.

For the effect of averageness on attractiveness when the non-averaged faces are low-attractive (Figure 2.4), we found that $c = 3.677$, $a_1 = -3.893$, $a_2 = -1.356$, $a_3 = -.009$, $b_1 = .004$, $b_2 = -.032$, and $c' = 3.649$. The indirect effect of averageness through fluency alone is $a_1 b_1 = -.014$, with a 95% bootstrap confidence interval of $(-.192, .139)$. The indirect effect of averageness through distinctiveness alone is $a_2 b_2 = .043$, with a 95% bootstrap confidence interval of $(-.209, .344)$. The serial indirect effect through both fluency and distinctiveness is $a_1 a_3 b_2 = -.001$, with a 95% bootstrap confidence interval of $(-.030, .033)$. These indirect effects sum to the total indirect effect of $.028$, with a 95% bootstrap confidence interval of $(-.260, .323)$. Note that none of the indirect effects are significant and the difference between the total effect, c, and the direct effect, c' is minimal (only $.018$).

For the effect of averageness on attractiveness when the non-averaged faces are high-attractive (Figure 2.5), we found that $c = .531$, $a_1 = .291$, $a_2 = -.748$, $a_3 = -.008$, $b_1 = .025$, $b_2 = 088$, and $c' = .590$. The indirect effect of averageness through fluency alone is $a_1 b_1 = .007$, with a 95% bootstrap confidence interval of $(-.074, .106)$. The indirect effect of averageness through distinctiveness alone is $a_2 b_2 = -.066$, with a 95% bootstrap confidence interval of $(-.355, .102)$. The serial indirect effect through both fluency and distinctiveness is $a_1 a_3 b_2 = .000$, with a 95% bootstrap confidence interval of $(-.011, .012)$.

These indirect effects sum to the total indirect effect of $-.059$, with a 95% bootstrap confidence interval of $(-.354, .122)$. Again, note that none of the indirect effects are significant and the difference between the total effect, c, and the direct effect, c' is minimal (only $-.059$).

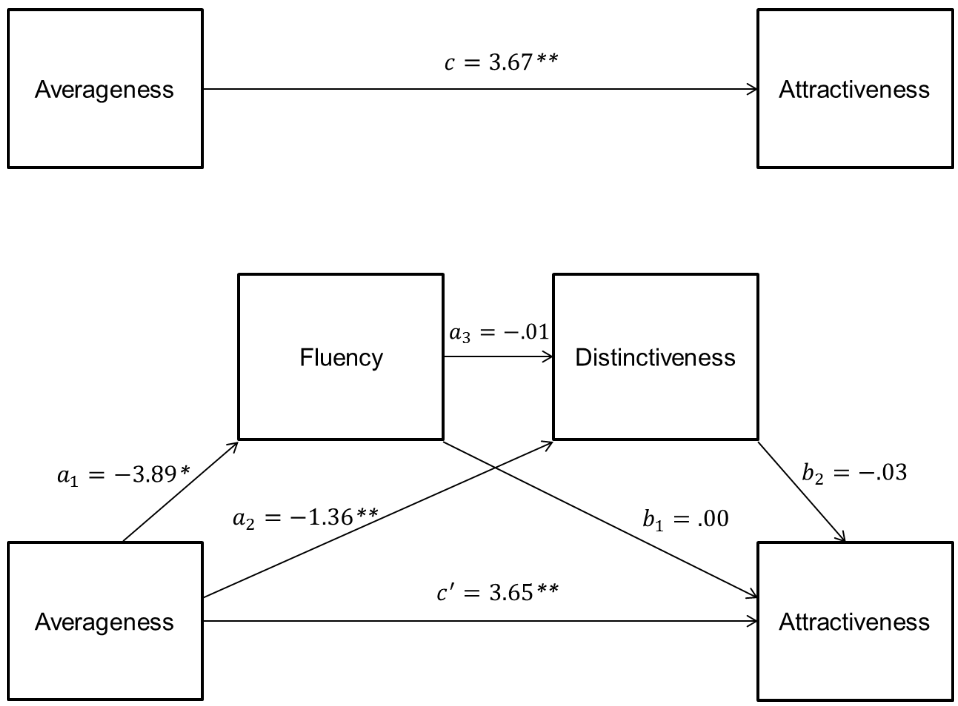
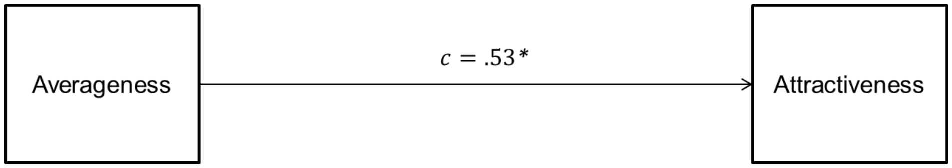


Figure 2.4. Model parameter values for averaged faces vs. low-attractive individual faces. The numbers indicate how many units the consequent variable is expected to change given a 1 unit increase in the antecedent variable. * indicates a significant effect at $p = .005$. ** indicates a significant effect at $p = .001$. Lack of a * indicates lack of significance at $p = .05$.



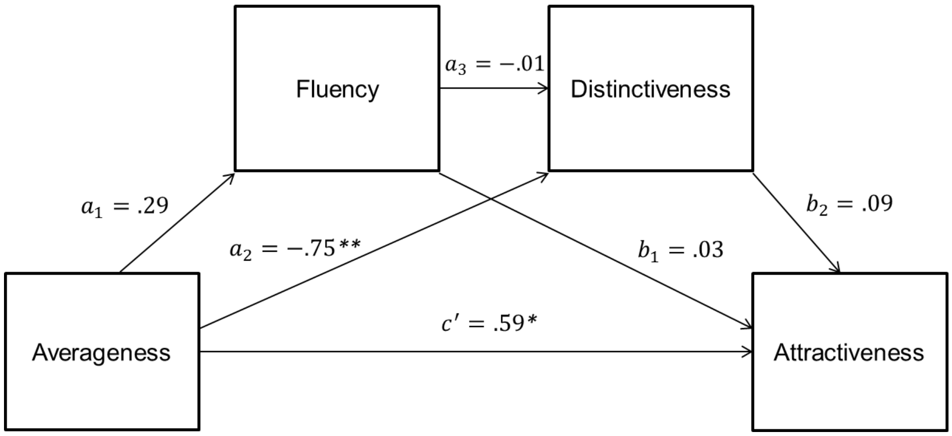


Figure 2.5. Model parameter values for averaged faces vs. high-attractive individual faces. The numbers indicate how many units the consequent variable is expected to change given a 1 unit increase in the antecedent variable. * indicates a significant effect at $p = .005$. ** indicates a significant effect at $p = .001$. Lack of a * indicates lack of significance at $p = .05$.

Discussion

In the results section we noted that none of the indirect effects were significant. Perhaps this is easier to see by looking at the individual effects for each model: While averageness significantly decreases distinctiveness and significantly decreases processing speed (i.e., increases fluency; also, this is only true when the averaged faces are compared to low-attractive individual faces), neither fluency nor distinctiveness has a significant effect on attractiveness once the direct effect of averageness has been accounted for. This can also be seen by noting how similar c is to c' in both figures: The total effect of averageness on attractiveness is almost identical to the direct effect of averageness on attractiveness once fluency and distinctiveness have been accounted for. In other words, the effect of averageness on attractiveness does not seem to be mediated by fluency or distinctiveness.

Looking at the models together, note that the effects are moderated by how attractive the individual faces are to begin with. In particular, changes in distinctiveness and attractiveness are more prominent when the averaged faces are compared to low-attractive individual faces than when they are compared to high-attractive individual faces. Also, while face averages are processed more fluently than low-attractive individuals, they are not processed significantly differently from high-attractive individuals. None of this is surprising given everything we already know about averageness and attractiveness: averaged faces are simply more similar to attractive individual faces than they are to unattractive individual faces.

General Discussion

It is now well-known that averaged configurations of faces are attractive. The interesting question is *why*. In the introduction, we argued that the effect is not mediated by skin quality, youthfulness, symmetry, or sexual dimorphism. In the subsequent mediation model, we showed that the effect is not mediated by general familiarity/distinctiveness or fluency. There must be something else about averaged faces that makes them attractive. But what?

Problem with the measures

One possibility is that our measures of distinctiveness and/or fluency were not adequate and better measures may yield significant indirect effects. For instance, Morris, Wickham and colleagues have argued (and found empirically) that distinctiveness as measured by the ease of spotting a face in a crowd (the measure used in the present study) is different from general familiarity as measured by how much a face looks like a typical face, and that only the latter is linearly related to attractiveness (Morris & Wickham, 2001; Wickham & Morris, 2003; Wickham, Morris, & Fritz, 2000). However, we used data from Trujillo et al. (2014) who found that distinctiveness as defined by the ease of spotting a face in a crowd *is* linearly related to attractiveness (strong, negative correlation), so this criticism may not apply to this particular dataset. Alternatively, reaction time on a species-categorization task (like the one used in the present study) may not be the best way to measure processing fluency. Perhaps fluency would mediate the effect of averageness on attractiveness with an alternative measure, like reaction time on a gender-based categorization task.

Averageness in its own right

Alternatively, perhaps we haven't found a significant mediator because there isn't one; averageness is *per se* attractive. There is already a wealth of evidence that the visual system has a prior expectation for the geometry and other features of a face – based on a moving window of faces that we have actually seen – and that the visual system automatically compares incoming faces against this prior expectation (Giese & Leopold, 2005; Leopold, O'Toole, Vetter, & Blanz, 2001; Loffler, Yourganove, Wilkinson, & Wilson, 2005; Malpass & Hughes, 1986; Rhodes, et al., 2004; Valentine, 1991). The visual system could easily compute averageness *per se* from a 2D image of a face by estimating how close the face is to this stored prior.

However, the fact that the visual system could easily compute the averageness of a face doesn't explain why averageness should be attractive. Two essential arguments have surfaced to explain why averaged faces might be attractive in their own right: One based on evolution and sexual selection, the other based on information-processing biases inherent in the nervous system.

Sexual selection. According to this argument, facial averageness is a signal for mate quality so that a preference for averaged faces evolved as an adaptation for finding good mates (for recent reviews, see Little, Jones & DeBruine, 2011; Rhodes, 2006). Facial averageness could either signal direct benefits for a potential mate (e.g., enhanced resource attainment or reduced risk of contagion) or indirect benefits which would not benefit a mate directly but would be passed on to offspring (e.g., heritable resistance to disease).

Direct benefits. Facial averageness has been argued to signal direct benefits because natural environments militate against extreme phenotypic traits in a population and tend to stabilize on mean or modal trait values (for discussion, see Symons, 1979). Thus, according to this line of thought, average facial morphology should be associated with above-average performance on tasks such as chewing and breathing, leading to a more robust, healthier mate (Symons, 1979).

Empirically, averaged faces look healthier (Rhodes, Yoshikawa, Palermo et al., 2007; Rhodes et al., 2001b), but no one has yet tested the connection between mathematical averageness and medical health as measured from actual medical records[10] (but, see Rhodes et al., 2001b for the connection between actual health and ratings of distinctiveness).

It is also worth noting that simply avoiding unusualness is in itself a viable mate selection strategy – even if averaged traits aren't necessarily the most fit – since unusual

---

[10] It should also be noted that attractiveness itself is only weekly related to actual, medical health (for review, see Rhodes, 2006), but this may be the result of good nutrition and modern medicine and not reflective of the evolutionary utility of attractiveness for our ancestors (Daly & Wilson, 1999; Thonhill & Gangestad, 1996; 1999).

mutations are deleterious more often than they are beneficial (Koeslag 1990; Koeslag & Koeslag, 1994). Computational models show that such a strategy can lead to evolutionarily stable populations with average trait values (Koeslag, 1990; Koeslag & Koeslag, 1994). A few authors have recently proposed that this mate-selection strategy could explain why people find averaged faces attractive (Iyengar, Kulkarni, & Vidya, 2015; Unnikrishnan, 2012), but there is no empirical data concerning how face preferences evolve or whether face preferences are even heritable.

Indirect benefits. Facial averageness has also been argued to signal indirect benefits via genetic quality, or "good genes" (see Andersson, 1994; Roberts & Little, 2008). The line of reasoning is as follows: For continuously distributed, heritable traits, trait values near the average for a population tend to be heterozygous (Markow & Gottesman, 1993; Soule & Cuzin-Roudy, 1982) which is beneficial because heterozygosity (genetic diversity) is associated with developmental stability, outbreeding, and resistance to parasites (Coltman & Slate, 2003; Mitton & Grant, 1984; Thornhill & Gangestad, 1993). Since facial averageness is continuously distributed and heritable (Lee et al., 2015), it serves as a viable cue to "good genes" and there should be evolutionary pressure to pick up on such a cue (Roberst & Little, 2008; Thornhill & Gangestad, 1993, 1999). Empirical support for this line of thinking is mixed.

On the one hand, there is empirical evidence connecting heterozygosity, averageness, and attractiveness for faces: Heterozygosity in the major histocompatibility complex (MHC) – a complex of genes that code for proteins involved in immune response – is positively associated with both facial averageness (Lie, Rhodes, & Simmons, 2008) and facial attractiveness (Roberts, Little et al., 2005).

On the other hand, in a study of twins and their siblings, Let et al. (2015) failed to find evidence that the genes that affect averageness also affect attractiveness. They used twin data to examine the genetic vs. environmental determinants of facial attractiveness and facial averageness. They found that, while the genetic component explained a significant amount of the variation for both attractiveness and averageness for faces, the genetic component for the covariation between attractiveness and averageness was not significant. This suggests that, while both facial attractiveness and facial averageness are heritable, the genes that affect one are not necessarily the same as the genes the affect the other. However, the authors note that this lack of finding may be due to a lack of power. More research needs to be done on the genetic determinants of facial averageness and the degree to which these genes can be considered “good” for mate quality.

Information processing. Contrary to the sexual selection account, researchers have argued that a preference for the average is the optimal way to solve certain signal processing tasks, so the attractiveness of averaged faces evolved as a byproduct of how our brains process information (for reviews, see Enquist et al., 2002; Rhodes, 2006).

For instance, Enquist et al. (2002) proposed a simple model where observers are faced with only two face-processing tasks: (1) to discriminate between sexes and (2) to respond similarly to variation within a sex (e.g., adult female faces are all different but should all be recognized as adult females). Solving the first task leads to a preference for sexual dimorphism (i.e., feminized females and masculinized males) and solving the second task leads to a preference for averageness within a sex. Thus, according to this model, averaged faces (within a sex) should be generally attractive, but not necessarily most attractive, and the attractiveness of an averaged face could be improved by making it look more feminine or masculine. There is empirical evidence that this might be the case, at least for female faces (Ghirlanda, Jansson, & Enquist, 2002; Perrett et al., 1998; Rhodes, Hickford, & Jeffery, 2000; for a counterargument, see Rubenstein, Langlois, & Roggman, 2002. See also, DeBruine et al., 2007). There is also evidence that a simple model like this can explain preferences across species. For example, chickens peck more furiously at attractive faces (as judged by human raters) after being trained to discriminate faces by gender (Ghirlanda et al., 2002). Furthermore, if a preference for averageness is a by-product of the way we process information, then we would expect it occur for more categories of objects than just faces, and it does. In addition to faces, people prefer averaged configurations of dot patterns (Bomba & Siqueland, 1983; Winkielman, Halberstadt, Fazendeiro, & Catty, 2006), birds, fish and cars (Halberstadt & Rhodes, 2003). However, it should be noted that averaged faces are attractive even when viewed by someone, like Galton, who is not under the demands of an experimenter, trying to perform an artificial task.

Taken together, all of this suggests that a by-products account is a plausible explanation for why averaged faces are attractive, but more empirical work needs to be done.

Concluding remarks

In this paper we have considered six different mediating variables to explain why averaged faces are attractive – skin quality, youthfulness, symmetry, sexual dimorphism, fluency and distinctiveness (see Figure 2.2) – and argued or demonstrated that none of them actually mediate the effect of averageness on attractiveness. This suggests that there is just something about averageness that makes faces attractive. It could be that averageness indicates mate quality, or perhaps a preference for averageness is a by-product of how our brains process information. More research needs to be done to figure out which alternative is most likely, but we are closer to solving what makes averaged faces attractive.

***

We have seen that people's preferences for visual textures and for faces are surprisingly consistent. We found that visual texture preferences can be predicted well by an ecological model and we investigated why averaged faces are consistently judged to be attractive, concluding that averageness *per se* is attractive. We turn our attention now to lower-level perception, specifically lightness perception. Lightness perception is thought to be consistent across human observers – and thus not influenced by cognitive factors such as beliefs or desires – because all humans have a similar vision system. In the following chapter, we investigate whether a high-level cognitive factor (perceived race of a face) can influence lightness perception (i.e., the skin-tone of the face).

Chapter 3

Changing the perceived race of a face does not change how light its skin looks

Kyle D. Stephens, Daniel A. Stehr, and Donald D. Hoffman (submitted).

Abstract

Can the race of a face alter how light its skin looks, regardless of how light its skin actually is? The answer to this question has important implications for any theory of perception because it could provide evidence that a high-level cognitive construct (race) can alter low-level visual experience (lightness perception).

But reports that deal with this question are conflicting, with some researchers claiming that race affects lightness perception (Levin & Banaji, 2006; MacLin & Malpass, 2003, 2001), and others challenging this claim (Firestone & Scholl, 2015).

In the current paper, we test whether race can affect lightness perception by combining the approaches of Levin and Banaji (2006) and MacLin and Malpass (2003, 2001). First, we create a set of morphed African-American/Caucasian faces that are judged to be Caucasian when wearing a Caucasian hairstyle but African-American when wearing an African-American hairstyle. Next, we use an adjustment procedure to show that, even though changing the hairstyle changes the perceived race of the faces, it does not change how light the skin-tone is judged. Finally, we argue that previous results are likely due to response bias and not a change in lightness perception. Overall, we conclude that the perceived race of a face does not affect how light it looks.

Introduction

Can cognitive states such as beliefs, desires or expectations directly affect our perceptual experience of the world? (Bruner & Goodman, 1947; Colins & Olson, 2014; Dunning & Balcetis, 2013). Or is perception encapsulated from the influence of such cognitive states? (Fodor, 1988; Marr, 1982; Pylyshyn, 1999). This question has been the subject of many perceptual experiments.

For instance, Bruner and Goodman (1947) showed that poorer children estimate coins to be larger than richer children, presumably because their greater desire for money alters their perception of size. Carter and Schooler (1949) countered that this effect is only maintained when the children estimate coin-size from memory and argued that poorer children misjudge coin-size because they are less familiar with money, not because the coins actually look larger.

Bhalla and Proffitt (1999) showed that students judge the slant of a hill to be steeper when wearing a heavy backpack, presumably because their greater expectation of effort alters their perception of slant. Durgin et al. (2009) countered that this effect goes away when students are given a deceptive cover story that justifies the backpack's presence and argued that backpack-wearing students misjudge hills because they are motivated to comply with the anticipated results of the experiment, not because the hills actually look steeper (see also Durgin, Klein, Spiegel, Stawser, & Williams, 2012; Nichols & Maner, 2008; Orne, 1962).

Many studies claim to demonstrate cognitive influences on perception (for reviews, see Colins & Olson, 2014; Dunning & Balcetis, 2013; Proffitt, 2006; Witt, 2011; Zadra &

Clore, 2011; Zeimbekis & Raftopoulos, 2015) and many studies challenge these claims (e.g., Firestone, 2013; Firestone & Scholl, 2014, 2015; Francis, 2012; Shaffer, McManama, Swank, & Durgin, 2013).

Here we investigate one instance of this larger debate: Can the race of a face (a high-level cognitive construct) alter how light its skin looks (a low-level perception), regardless of how light its skin actually is?[11] Levin and Banaji (2006) provided evidence that this might be true: They showed that observers rate a prototypical African-American face as darker than a prototypical Caucasian face even when the faces have the same average luminance and contrast.[12] This study has been singled out in the literature because, unlike other studies purporting to show cognitive influences on perception, it provides a convincing visual demonstration that any observer can "see for themselves" (Firestone & Scholl, 2015, see Figure 3.1). Thus, Levin and Banaji's result has been cited as one of the strongest examples of a cognitive influence on perception, both by psychologists (Colins & Olson, 2014; Firestone & Scholl, 2015) and by philosophers (Macpherson, 2012).

However, Firestone and Scholl (2015) showed that Levin and Banaji's stimuli are confounded. They heavily blurred the faces and found that, although observers could no longer determine race, they still rated the blurred African-American face as darker than the blurred Caucasian face (Figure 3.1). Indeed, even observers who categorized the African-American face as Caucasian still judged it to be darker. In other words, they argued that the

---

[11] In the literature on lightness perception, "how light something actually is" refers to its *luminance*, the objective intensity of light radiating from it (measured in cd/m2), whereas "how light something appears to be" refers to its *lightness*, the technical term for its subjectively perceived shade (see Adelson, 2000).
[12] For a digital image, average luminance and contrast are measured by the mean and the standard deviation of the image's gray-level histogram respectively.

African-American face looks darker than the Caucasian face because of a low-level stimulus confound, not because the faces differ in race.

Figure 3.1. The top row shows the original, luminance-matched Black/African-American and White/Caucasian faces used by Levin and Banaji (2006). The bottom row shows the blurred versions of the faces used by Firestone and Scholl (2015). The African-American face (left) was rated as darker in both cases, although participants could not determine the race of the blurred faces

However, Levin and Banaji (2006) obtained similar results with line-drawn faces that do not vary in local luminance (and are thus not confounded), and MacLin and Malpass (2003, 2001) showed that people label a racially blended face as darker when it's perceived to be African-American than when it's perceived to be Hispanic. Thus, the question remains: Can the perceived race of a face alter how light its skin looks? Or, are all such results due to stimulus confounds or response biases? We investigate these questions in the current paper.

First, we argue that the most compelling results thus far are those of MacLin and Malpass (2003, 2001) with blended faces, but that these results need to be tested using better methodology. In particular, MacLin and Malpass measured lightness judgments on a response scale with text anchor points (*light* and *dark*), rather than with the direct gray-level sample matches typical of lightness perception research; they created their stimuli using a composite program which produces face sketches that lack photographic detail; and they made no attempt to obscure their research hypotheses from participants, which may encourage response bias (Orne, 1962).

To improve the methodology, we combine the approaches of Levin and Banaji (2006) and MacLin and colleagues (MacLin & MacLin, 2011; MacLin & Malpass, 2003, 2001), and also take special care to obscure the purpose of the experiment to control for response bias (Durgin et al., 2009; Nichols & Maner, 2008; Orne, 1962). Following MacLin and MacLin (2011), we create a set of morphed Afro/Caucasian faces that are judged to be Caucasian when wearing a Caucasian hairstyle but African-American when wearing an African-American hairstyle. Then, we use the adjustment procedure of Levin and Banaji (2006) to test whether changing the perceived race of the faces (via their hairstyle) also changes how light their skin looks. Before going into the details, we will first discuss our motivation by examining the other evidence that race affects lightness perception.

Other evidence that race affects lightness perception

Line-drawn faces. Firestone and Scholl (2015) showed that Levin and Banaji's initial stimuli are confounded (Figure 3.1), but they did not venture to guess what low-level confound, specifically, might be causing the lightness distortion. They just noted that, although the average luminance for the prototypical Caucasian and African-American faces is the same, the local luminance changes are quite different.

But Levin and Banaji (2006) themselves noted a potential problem with local luminance differences:

[I]t is possible that subjects adjusted the faces differently because they focus their attention on different parts of a face when they believe it represents one race instead of another. For example, if subjects believe a face to be Black, they may focus on the eyes (a darker region), and, when they believe it is White, they may focus on the nose (a lighter region). If so, they might adjust the sample patch to be relatively dark for the Black face because they are matching it with the face's eyes. Conversely, White faces would be overbrightened because subjects are focused on the relatively light nose. Thus, attentional focus could cause our effect instead of lightness perception. (p. 506)

To counteract this problem, they ran an additional experiment with line-drawings of faces that do not vary in local luminance (Figure 3.2). They were especially careful and used line-drawings that had both white and black outlines. Using an adjustment procedure where subjects adjusted gray patches to match the perceived shade of the faces, they reported that the African-American faces were matched to darker samples on average than the Caucasian faces.

Firestone and Scholl (2015) write off these results as not perceptually convincing – "the Black and White line-drawing faces just don't *look* differentially light" – and note that "the lack of an associated 'demo' leaves these results open to alternative explanations" (p. 13).

Figure 3.2. Line-drawn faces used by Levin and Banaji (2006). Each left face is prototypically Black/African-American; each right face is prototypically White/Caucasian. They used both black and white outlines to control for the effect of line color. They found the African-American face was rated darker than the Caucasian face overall in both cases, but these results were not consistent across subjects.

A more forceful argument against these results is that they are not consistent. Levin and Banaji (2006) report that only 27 out of 45 subjects chose darker samples for the line-drawn African-American faces than for the line-drawn Caucasian face ($\chi^2(1, N = 45) = 1.80$, $p = 0.180$). Thus, although the African-American face got matched to darker samples on average, quite a few participants decided that the Caucasian face actually looks darker than the African-American face (18 out of 45 to be exact), contrary to the proposed lightness distortion. To their credit, Levin and Banaji (2006) acknowledge the problem – "A look at the stimuli [Figure 3.2] suggests that the race-specifying information they contained was subtle... Accordingly, we may have traded a fair amount of validity for control" (p. 507-508). Overall, this does not provide strong evidence that the perceived race of a face, in general, affects how light it looks.

Blended faces. In an alternative demonstration, MacLin and Malpass (2003, 2001) showed that people label a racially blended face as darker when it's perceived to be African-American than when it's perceived to be Hispanic. They created blended faces using a facial composite program such that facial features overlapped across Afro/Hispanic racial lines. When they added a racial marker characteristic of either an African-American or a Hispanic face – in this case hairstyle – they found that the blended faces were perceived as being either African-American or Hispanic, consistent with the racial marker, and that the faces were labeled as darker when wearing the African-American hairstyle than when wearing the Hispanic hairstyle.

In light of the confounds discovered in Levin and Banaji's stimuli, this is a compelling result. In this case, changing the perceived race of an *identical* face changes how observers rate its skin tone. This does not leave much room for confounds – since the faces are identical save for the hairstyle, the only potential confound is with the hairstyle. It could be that a darker hairstyle makes a face look darker due to attentional focus (see Levin and Banaji block quote in previous section). However, MacLin and Malpass note that the Hispanic hairstyles were darker overall, so this hypothesis is inconsistent with the fact that the faces were labeled as lighter when wearing a Hispanic hairstyle. Alternatively, a darker hairstyle might make a face look lighter due to contrast. However, MacLin and Malpass (2003) found that the lightness judgments did not change when the faces were displayed on black vs. white backgrounds, suggesting that contrast does not significantly influence ratings of skin tone. Thus, these results seem to provide strong, non-confounded evidence that the perceived race of a face can affect how light it looks.

But, as noted above, there are some problems with MacLin and Malpass' (2001, 2003) experiments.[13] In particular, (1) they measured lightness judgments on a response scale with text anchor points (*light* and *dark*), rather than with direct gray-level sample

---

[13] 3It's worth pointing out that Levin and Banaji (2006) cited MacLin and Malpass' (2001, 2003) result and noted some of the same problems.

matches typical of lightness perception research, (2) they created their stimuli using a composite program which produces face sketches that lack photographic detail, and (3) they made no attempt to obscure their research hypotheses from participants which may encourage response bias (Orne, 1962). The purpose of the current research is to replicate MacLin and Malpass' results in a way that addresses these problems.

Overview of experiments

Before testing MacLin and Malpass' results, we first replicate Levin and Banaji's results with their original stimuli using their adjustment procedure (Experiment 3.1). We run this experiment as a check on the adjustment procedure, to make sure we are able to measure an effect when we know one exists.

Then, to test MacLin and Malpass' results, we start by creating a set of morphed Afro/Caucasian faces using face-morphing software. We create two version of each face, one with the hairstyle of the original Caucasian face, the other with the hairstyle of the original African-American face. In Experiment 3.2a, we determine which morphed faces look Caucasian when wearing the Caucasian hairstyle but African-American when wearing the African-American hairstyle. In Experiment 3.2b, we have participants match the skin-tone of these morphed faces using the same adjustment procedure as in Experiment 3.1. To obscure the purpose of the experiment, we also have participants match the gray-level of other objects that are not human faces and we inform them that the experiment is about "how people perceive the shading of objects," with no mention of "faces" or "race." If the lightness distortion is maintained, this provides strong evidence that race can affect lightness perception. However, if the lightness distortion is not maintained, then this brings previous results into question.

Experiment 3.1: Replicating Levin and Banaji (2006)

For this experiment, our goal was to replicate Levin and Banaji's results with their original face stimuli, using their adjustment procedure. Participants matched the shading of the faces by adjusting the gray-level of uniform, rectangular patches. Following Levin and Banaji's original protocol, there was no deception involved, as participants were told the experiment was about "how people perceive the shading of faces from different races." The purpose of this experiment was to test the adjustment procedure and make sure we are able to measure an effect when we know one exists.

Participants

We recruited 31 participants (11 male) through the University of California subjects pool. We ended up excluding one female participant because of the number of times she made no adjustment on two consecutive trials (which was 4.1 standard deviations above the mean for all participants). After excluding this participant, the racial make-up (gathered by self-report) was as follows: 11 Asian/Pacific Islander, 13 Hispanic/Latino and 6 White/Caucasian. The mean age was 20.3 years (range 18-24 years). All subjects were undergraduate students at the University of California, Irvine and all received extra credit for participation.

Apparatus

Stimuli were presented using Dell computers attached to 17-in. Dell LCD monitors set at a resolution of 1280x1024 and a refresh rate of 60Hz. Participants were run in groups of one to five in a room containing 8 computers. Participants responded using the computer's keyboard and presentation was controlled by a program written in Matlab, using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007).

Stimuli

The stimuli consisted of three morphologically distinct faces, representing a prototypical Caucasian face, a prototypical African-American face and a racially Ambiguous face. These faces were graciously provided by Levin and Banaji and are displayed in Figure 3.3. The Caucasian and African-American prototypes were created by blending a set of 16 faces from the race the prototype was to represent and the Ambiguous face was created by morphing the Caucasian face and the African-American face in a 50-50 blend (for more details on how the faces were created, see Levin, 1996; Levin & Banaji, 2006).

For each face, we created 13 gray-level variations. The initial images (level 0) all had a mean 8-bit gray-level of 141 out of 256 (all gray-levels are reported using these 8-bit units), which is equivalent to 37.2 cd/m^2, as measured using a Photo Research PR-670 spectroradiometer on one of the computers used in the experiment. The remaining 12 gray-level variations were created by changing the mean gray-level, in increments of 5, to six levels below and six levels above the 0 level. This lead to a set of faces whose (scaled) gray-levels varied from -30 to +30 in increments of 5 (for some sample stimuli, see Figure 2B in Levin & Banaji, 2006). The contrast range for all faces – as measured by the standard deviation of their gray-level histograms – was kept relatively constant. Each face measured 58 (horizontal) by 73 (vertical) pixels at 72 dots per inch. Each gray patch measured 80 (horizontal) by 100 (vertical) pixels.

Figure 3.3. The face stimuli used in Experiment 3.1 (at 0 gray-level). These represent a prototypical African-American face (left), a racially Ambiguous face (middle) and a prototypical Caucasian face (right). These faces were graciously provided by Levin and Banaji.

Procedure

Participants used the method of adjustment to give lightness judgments for the faces. On each trial, they viewed a reference face and an adjustable gray patch, and were informed to match the patch to the shade[14] of the skin on the reference face. They used the up and down arrow keys on the keyboard to adjust the gray-level of the patch (in increments of 5) and pressed the spacebar to move on to the next adjustment once they perceived a match. After each stimulus, there was a 500ms interstimulus interval with Gaussian noise that was just large enough to cover the area of the reference face and the gray patch. The reference face and the gray patch were vertically centered and horizontally 10cm apart on the screen (center-to-center) with a viewing distance of approximately 50cm. At this viewing distance, the faces subtended approximately 2.2 degrees of visual angle, and the gray patches subtended approximately 3.0 degrees of visual angle. All stimuli were set against a white background. On each trial, the reference face was set to one of five different gray-levels (-10, -5, 0, +5, +10), and the initial gray-level of the adjustable patch was offset from the reference face by +/-10 or +/-20. Thus, there were 20 trials for each

[14] We used the words "shade" and "shading" in the instructions to match the instructions given by Levi and Banaji (2006).

face, for a total of 60 trials per subject. The trials were randomly ordered and the screen-side of the reference face (left vs. right) was counterbalanced across subjects.

In accordance with Levin and Banaji's original protocol, there was no deception – subjects were told that this study is about "how people perceive the shading of faces from different races." To gauge each subject's knowledge of the experiment, we also gave them a post-experiment survey where we asked them to guess with free response what they thought the experiment was about.

Results

For all subjects, we calculated the average lightness distortion for each face prototype. The lightness distortion is defined as the chosen gray-level of the adjustment patch minus the actual mean gray-level of the reference face. Thus, a lightness distortion of 0 indicates perfect matching by the subject, while positive or negative lightness distortions indicate that the subject rated the face to be lighter or darker than it actually is respectively.

The average distortion for each face prototype is shown in Figure 3.4. A repeated-measures ANOVA revealed a significant effect of race on lightness distortion, $F(2,58) - 27.8$, $p < .001$. While there is a bias for all faces to be rated lighter than they actually are – Levin and Banaji (2006) reported a similar bias – we are interested in the differences between faces. Paired t-tests revealed that the African-American face ($M = 7.05$, $SD = 4.59$) was rated significantly darker than both the Caucasian face ($M = 15.0$, $SD = 4.95$), $t(29) = -5.77$, $p < .001$, $d = 0.98$,[15] and the Ambiguous face ($M = 10.6$,

[15] Cohen's *d* gives a standardized difference between two group means (Cohen, 1988). It is a measure of effect size, intended to signify the magnitude of the difference between the two groups (Kelley & Preacher, 2012). Cohen

$SD = 2.39$), $t(29) = -4.39, p < .001, d = 0.44$, and the Ambiguous face was rated significantly darker than the Caucasian face, $t(29) = -4.73, p < .001, d = 0.59$. The mean difference in judgments between the Caucasian face and the African-American face was 7.95 gray-level units, which corresponds to approximately1.60 cd/m^2. Although this difference is small in luminance, it was consistent across subjects: 26 out of 30 subjects chose darker samples for the African-American face than for the Caucasian face, $\chi^2(1, N = 30) = 16.1, p < .001$.

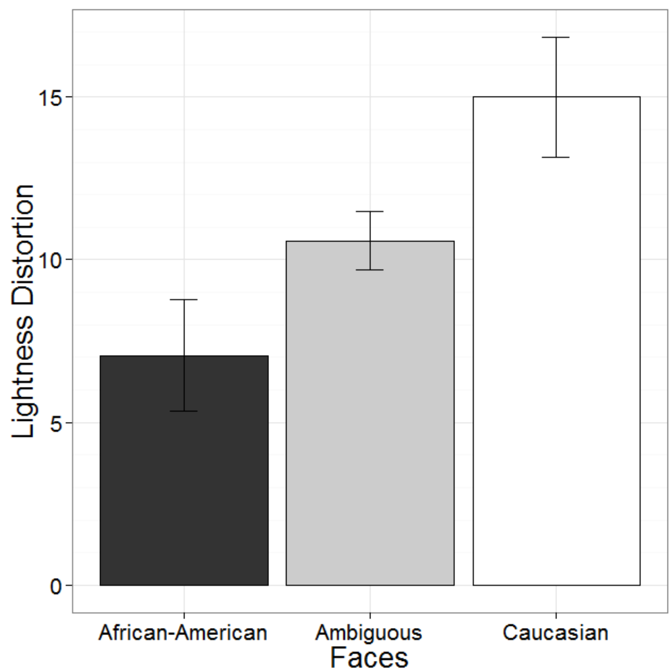


Figure 3.4. Lightness distortion for the faces in Experiment 3.1. Error bars represent within-subjects 95% confidence intervals (Cousineau, 2005; Morey, 2008). Overall, the African-American face was rated as darker than the Ambiguous face which was rated as darker than the Caucasian face.

We also analyzed the subjects' post-experiment guesses about the nature of the research. We had 5 independent raters judge each subject's response on a 5-point Likert

(1988) recommends distinguishing between 'small' (d = 0.2), 'medium' (d = 0.5) and 'large' effects (d = 0.8) as a helpful guide.

scale, where 1 means "They definitely did not know the research hypothesis" and 5 means "They definitely did know the research hypothesis." The raters considered the research hypothesis to be the following: "Identifying the race of a face will affect the perceived lightness/shading of the face. Specifically, African-American/Black faces will be perceived as darker than Caucasian/White faces." The average rating across subjects was 3.17 with standard deviation 1.16.

Discussion

In Experiment 3.1, we successfully replicated the results of Levin and Banaji (2006). Using Levin and Banaji's original stimuli, we showed that participants rate the African-American face as having darker skin on average than the luminance-matched Caucasian face using an adjustment task. Of course, these stimuli have already been shown to be confounded (Firestone & Scholl, 2015), so we did not run this experiment to determine whether race affects lightness perception. Rather, we wanted to show that we could measure a known effect with established stimuli using an adjustment procedure. We succeeded in that task as we measured the predicted effect with an effect size similar to that of Levin and Banaji (our effect size was 0.98. Theirs ranged from 0.75 to 1.65).

With the procedure established, we used it in the next experiment to test whether morphed Afro/Caucasian faces look darker when they are perceived to be African-American than when they are perceived to be Caucasian.

Experiment 3.2a: Creating morphed faces

For this experiment, our goal was to create faces that would be perceived as African-American when wearing an African-American hairstyle but Caucasian when wearing a Caucasian hairstyle. Thus, following MacLin and MacLin (2011), we first created several morph continuums with pairs of Caucasian and African-American faces and then tested which ones would be perceived as Caucasian with Caucasian hair but African-American with Afro hair.

Participants

We recruited 42 participants (14 male) through the University of California subjects pool. The racial make-up (gathered by self-report) was as follows: 22 Asian/Pacific Islander, 10 Hispanic/Latino, 9 White/Caucasian, and 1 Middle Eastern. The mean age was 21.2 years (range 18-37 years). All subjects were undergraduate students at the University of California, Irvine and all received extra credit for participation.

Apparatus

Same as in Experiment 3.1.

Stimuli

To create the stimuli, we first obtained photographs of Caucasian and African-American faces (150 African-American, 184 Caucasian) from the database of Meissner, Brigham, and Butz (2005), and used Adobe Photoshop to remove their hair. Then, following MacLin and MacLin (2011), we used image-morphing software (FataMorph 3) to create morph continuums between different Caucasian and African faces (see Figure 3.5). The starting Caucasian face represents 0% on the continuum, the starting African-American face 100%, and percentages in between represent percentages of African-American face in the blend. We used a pilot study to narrow the set of morph continuums down to six using the following criteria: (1) The pure African-American face should look sufficiently African-American, (2) the pure Caucasian face should look sufficiently Caucasian, and (3) blends in the middle should look relatively ambiguous and not strongly African, Caucasian, or some other race (as can happen; MacLin & MacLin, 2011). We used faces from these six continuums as stimuli in the experiment (see Figure 3.6 for a sample of some of the faces used).

Figure 3.5. An example morphological continuum between a Caucasian face (0%, left) and an African-American face (100%, right). Each step along the continuum represents a 16.67% increase in the proportion of the African-American face to the blend. Any stimuli used in an experiment would first be cleaned up using Adobe Photoshop (i.e., any "ghosting" would be removed) and grayscaled.

MacLin and MacLin (2011) found that the maximally ambiguous morph percentage (i.e., the morph percentage at which observers have close to a 50-50 chance of categorizing the face as either Caucasian or African-American) did not occur at 50% morph, but rather was biased toward the Caucasian face. Thus, for our stimuli, we chose the 42%, 46% and 50% morphs from each of the six morph continuums.

Furthermore, we thought the categorization boundary was stronger when the faces were lightened (perhaps because it is more common to see a "light-skinned" Black person than a "dark-skinned" White person. See, e.g., Peery & Bodenhause, 2008). Thus, we also included a lightened version of each face in addition to the original. To create the lighted version, we simply increased the mean gray-level of each image by 20 – 25 units, while keeping the contrast relatively constant.

Finally, we created two versions of each face: one wearing the hairstyle of the initial Caucasian face, and one wearing the hairstyle of the initial African-American face. We used

Adobe Photoshop to select and copy the hair from the starting faces onto the morphed faces.

Overall, we created 72 face stimuli: 6 morph continuums × 3 morph percentages from each continuum × 2 mean gray-levels × 2 hairstyles. All images were roughly 290 (vertical) by 160 (horizontal) pixels.

Procedure

On each trial, subjects were shown one of the faces and asked to categorize it as Caucasian (by pressing "c" on the keyboard) or African-American (by pressing "a" on the keyboard). Subjects were also informed that some of the faces might look ambiguous or multi-racial, but they should choose the race that the face looks most like. After choosing a racial category, subjects pressed the return key to submit their response. Each submission was followed by a 500ms interstimulus interval containing a 375 (vertical) by 300 (horizontal) pixel rectangle of Gaussian noise. All faces were shown against a white background at central fixation. Subjects sat roughly 50cm away from the screen. At this viewing distance, the faces subtended approximately 8.8 degrees of visual angle.

We were concerned that our results might be compromised if subjects saw a face morph with both hairstyles and recognized it to be the same face. Thus we had two groups of participants: The first group saw half of the faces with Caucasian hair and the other half with African-American hair. The second group saw the same faces as the first group, but the hairstyles were switched. Thus, the faces with the Caucasian hair in the first group had African-American hair in the second group and vice-versa. In this way, each group saw faces with both African-American and Caucasian hairstyles, but no subject saw the same face with two different hairstyles. Each subject saw a total of 36 different stimuli (18 unique faces at two different mean gray levels).

The stimuli were presented to subjects in pseudo-random order: Each morph type (of the 6) was shown once before another of that type was shown, but the morph-percentage, gray-level and order of the 6 morph-types was randomized.

Each participant ran through the entire experiment twice, both times in the same random order. Thus, they categorized each face twice.

Results

For a given face, we counted a subject's categorization of that face as either Caucasian, African-American or Ambiguous. Their categorization was counted as Caucasian if they rated the face Caucasian both times they saw it, as African-American if they rated the face African-American both times they saw it, and as Ambiguous if they rated the face Caucasian on one viewing but African-American on another.

Of the 6 unique face-morphs, there were 4 that participants consistently judged to be Caucasian on both viewings when they saw it wearing the Caucasian hair and African-American on both viewings when they saw it wearing the African-American hair. Figure 3.6 shows each of these four face-morphs at the gray-level and morph percentage that led to the highest percentage of desired racial categorization.

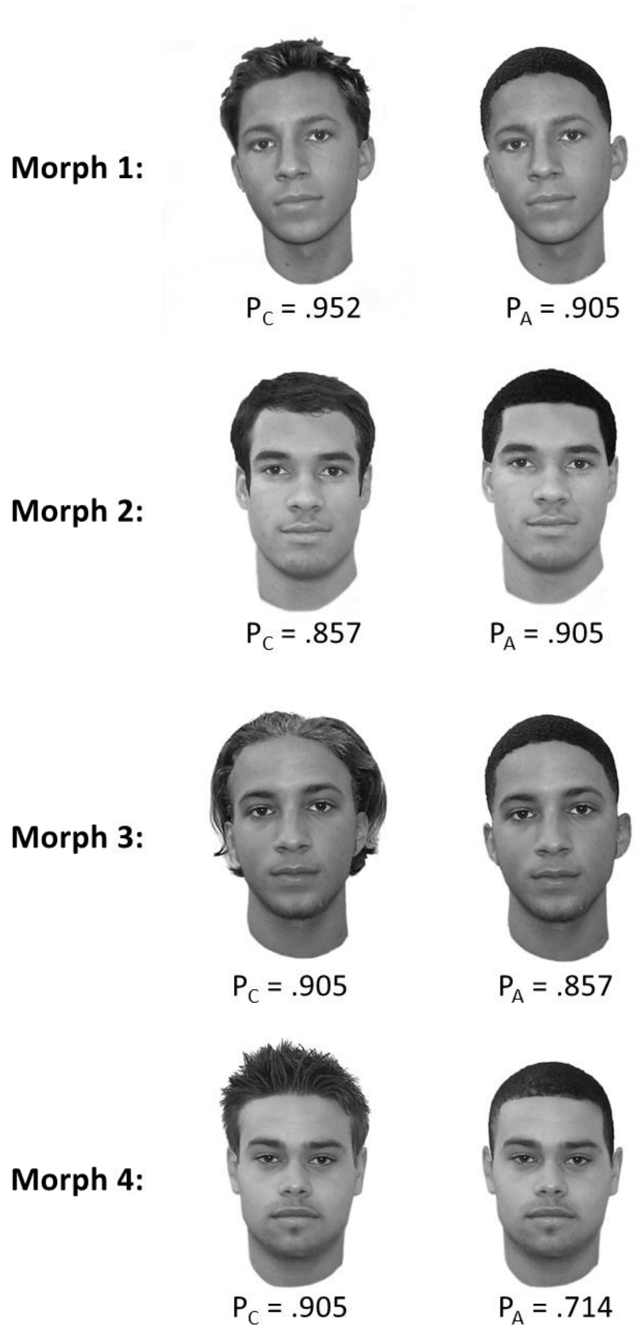


Figure 3.6. The four best face morphs from Experiment 3.2a, used in Experiment 3.2b. Along each row, the left face is identical to the right face, but the faces have different hairstyles. Each left face has the Caucasian hairstyle; each right face has the African-American hairstyle (the hairstyles come from the original, pre-morphed faces). The numbers represent the proportion of participants who judged the face to be either Caucasian (left faces) or African-American (right faces) both times they saw it.

The best face was categorized as Caucasian 95.2% of the time when wearing the Caucasian hair and as African-American 90.5% of the time when wearing the African-American hair. The worst face was categorized as Caucasian 90.5% of the time when wearing the Caucasian hair and as African-American 71.4% of the time when wearing the African-American hair (see Figure 3.6). We used these eight faces (4 morphs × 2 hairstyles) in the lightness-matching experiment.

Experiment 3.2b: Lightness-matching morphed faces

For the lightness-matching experiment, we wanted to see if the four face-morphs displayed in Figure 3.6 would be seen as lighter when wearing Caucasian hair than when wearing African-American hair. To test this, we used the same adjustment procedure as in Experiment 3.1 – we had subjects adjust the gray-level of a uniform rectangular patch to match the skin-tone of each face. To obscure the purpose of the experiment, we also included 9 other gray-scale images of things that were not human faces (three animals, three fruits, and three tools), and informed participants that the experiment was about “how people perceive the shading of objects” with no mention of “faces” or “race.”

Participants

We recruited 30 participants (12 male) through the University of California subjects pool. Assuming an effect size similar to that of MacLin and Malpass (2001), this should result in a power[16] of 0.90. We ended up excluding one male participant because of the number of times he made no adjustment on two consecutive trials (which was 4.8 standard deviations above the mean for all participants). After excluding this participant, the racial make-up (gathered by self-report) was as follows: 11 Asian/Pacific Islander, 7

[16] All power calculations computed using G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007).

Hispanic/Latino, 9 White/Caucasian, and 2 Black/African-American. The mean age was 22.4 years (range 18-58 years). All subjects were undergraduate students at the University of California, Irvine and all received extra credit for participation.

Apparatus

Same as in Experiment 3.1.

Stimuli

The stimuli consisted of 17 unique gray-scale images. This included the 8 faces from Experiment 3.2a (four face-morphs with both Caucasian and African-American hair, Figure 3.6), as well as 9 images of objects that were not human faces (three animals, three fruits, and three tools). The object images consisted of a dog, a cat, and a monkey, a saw, a clamp and a mallet, and a slice of watermelon, some blueberries, and a banana. They were collected from internet image searches. All faces were roughly 290×160 pixels, and all other object images were this size or slightly smaller.

To create the gray patches, we first measured the average gray-level of each object/face on an 8-bit channel (i.e., 0-255). We created the darkest gray-patch to be approximately 20 gray-levels darker than the darkest stimulus and the lightest gray-patch to be approximately 20 gray-levels lighter than the lightest stimulus. This corresponded to a darkest gray-level of 44/255 (luminance = 2.61 cd/m^2) and a lightest gray-level of 210/255 (luminance = 104.8 cd/m2) respectively (luminance measured using a Photo Research PR-670 spectroradiometer on one of the computers used in the experiment). The remaining gray-patches were constructed in increments of 2 gray-levels from darkest to lightest. Each gray patch was 375 (vertical) by 300 (horizontal) pixels.

Procedure

Participants used the same adjustment procedure as in Experiment 3.1 to give lightness judgments for the faces/objects. On each trial they viewed a reference object and an adjustable gray patch and were informed to match the shade of some particular feature of each object. For example, for the dog image, subjects were told to “match the shade of the dog’s fur” and for the banana image, subjects were told to “match the shade of the banana’s peel.” For the faces, subjects were told to “match the shade of the person’s skin.” They used the up and down arrow keys on the keyboard to adjust the gray patch (in increments of 2 gray-levels) and pressed the spacebar to move on to the next adjustment once they perceived a match. The object images and the patches were vertically centered on the screen and horizontally separated by 10cm, center-to-center. After each stimulus, there was a 500ms interstimulus interval with Gaussian noise that was just large enough to cover the area of the object image and the gray-patch. Subjects sat approximately 50cm from the screen. At this viewing distance, the object images subtended approximately 8.8 degrees of visual angle or slightly less, and the gray-patches subtended approximately 11.3 degrees of visual angle.

Each participant saw all of the faces, but the experiment was blocked so that they saw half of the faces with Caucasian hair and half with African-American hair in the first block. Then, in the second block, the faces with Caucasian hair in the first block were given African-American hair and those with Afrrican-American hair were given Caucasian hair. Participants took a mandatory break in between blocks and were informed that they would be completing the exact same experiment (with the same stimuli) after the break. They were not informed that the hair on the faces had been switched, and only one participant guessed as much during post-experiment questioning. Which faces had Afro/Caucasian hair first was counterbalanced across participants. All stimuli (faces and other objects) were shown in random order.

After completing all trials, participants guessed what they thought the experiment was about with free response.

Results

The average lightness distortion for the morphed faces is shown in Figure 3.7. A two-way repeated-measures ANOVA did not reveal a significant effect of race on lightness distortion, $F(1,28) = 0.183$, $p = 0.672$. While there was a significant effect of face morph on lightness distortion, $F(1,28) = 14.2$, $p < .001$, and each face was judged to be lighter than it actually is (Figure 3.7), we are only interested in the lightness difference between faces with different hairstyles.

Since the the ANOVA did not reveal a significant effect of race on lightness distortion, we decided to conduct a post-hoc analysis using Bayes factors since they take into account both evidence for and against the null hypothesis (see, e.g., Wagenmakers, 2007). A Bayes factor compares the data under the null hypothesis (that the faces look equivalent in lightness when wearing different hairstyles) to the data under the alternative hypothesis (that the faces look different in lightness when wearing different hairstyles)[17] to determine which hypothesis is more likely.

[17] Although the initial hypothesis was one-tailed (i.e., that the morphed faces should look darker when wearing the African-American hairstyle), we decided to run the post-hoc analysis with the two-tailed hypothesis since the data is not clearly trending in one direction or the other (see Figure 3.7).

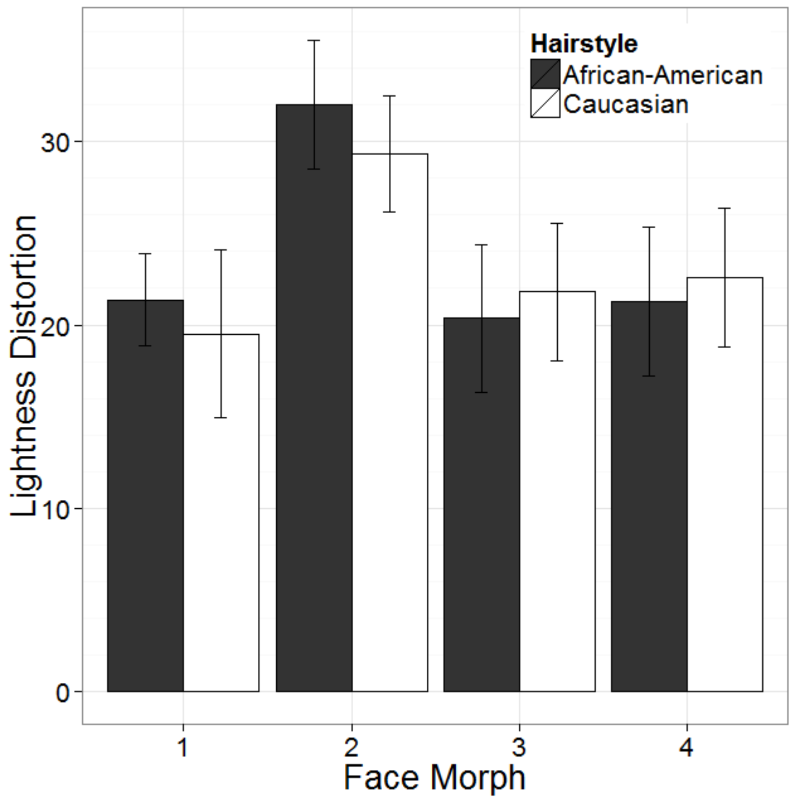


Figure 3.7. Lightness distortion for the face morphs in Experiment 3.2b. Error bars represent within-subjects 95% confidence intervals (Cousineau, 2005; Morey, 2008). Which hair the face morphs wear (African-American vs. Caucasian) does not appear to affect how light they look.

For each face morph, we estimated a Bayes factor (null/alternative) using the JZS prior with a scale factor of 1.0 (Jarosz & Wiley, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009). For Morph 1, this estimate suggested that the data were 5.48 : 1 in favor of the null; for Morph 2, the data were 3.80 : 1 in favor of the null; for Morph 3, the data were 6.12 : 1 in favor of the null; and for Morph 4, the data were 6.11 : 1 in favor of the null. In other words, the Bayes factor estimates suggested that, given the data, it is 3.8 to 6.1 times more likely that the faces look equivalent in lightness than different in lightness when wearing different hairstyles.

Additionally, participants were not consistent with their ratings for any of the face morphs. For faces wearing the African-American hairstyle, only 12 out of 29 subjects chose darker samples for Morph 1, $\chi^2(1, N = 29) = 0.862, p = .353$; 11 out of 29 chose darker samples for Morph 2, $\chi^2(1, N = 29) = 1.69, p = .194$; 13 out of 29 chose darker samples for Morph 3, $\chi^2(1, N = 29) = 0.310, p = .578$; and 15 out of 29 chose darker samples for Morph 4, $\chi^2(1, N = 29) = 0.035, p = .853$.

We also analyzed the participants" post-experiment guesses about the nature of the research. We had 5 independent raters judge each participant's response using the same 5-point Likert scale as in Experiment 3.1. The average rating across participants was 1.69 with standard deviation 0.85, which was significantly lower than the ratings for the participants in Experiment 3.1, $t(57) = -5.59, p < .001$. This indicates that we successfully concealed the research hypothesis.

Discussion

Using the morphed face stimuli from Experiment 3.2a, we failed to find support for the assertion that the perceived race of a face affects how light its skin-tone looks. That is, even though changing the hair on these faces changed how they were racially categorized (Experiment 3.2a), it did not change how light observers rated their skin-tone (Experiment 3.2b).

A critic might protest that we haven't shown there isn't an effect, just that our study was not powerful enough to find it. However, MacLin and Malpass (2001) reported an effect size of 0.55 in their study with blended Caucasian/Hispanic faces. Presumably we should find an even larger effect for blended Caucasian/African-American faces (Levin & Banaji, 2006), but if we conservatively assume the same effect size as MacLin and Malpass, then our 30 recruited participants would give us a power of 0.90, well above Cohen's (1988) recommended power level of 0.80.[18] Additionally, we established the efficacy of our adjustment procedure in Experiment 3.1. In other words, if an effect exists, then we had a well-established procedure with well-above the recommended power to find it.

An alternative criticism is that the hair confounds the stimuli. That is, since the African-American hair tends to be darker, it may cause the faces to look lighter by contrast which would offset any effect of race. However, such a criticism does not seem to be supported by the results. For example, Morphs 1 and 3 have the exact same African-American hair, but Morph 1 is rated as lighter when wearing the African-American hair, whereas Morph 3 is rated as darker. Furthermore, MacLin and Malpass (2003) found that observers gave equivalent lightness ratings for faces regardless of whether the background behind the face was white or black, suggesting that skin-tone judgments are not significantly affected by contrast. We also failed to find an effect when we ran this experiment with blurred versions of the faces where race can no longer be identified (not reported here), suggesting that a low-level confound is not influencing the results differentially from race.

Finally, we should comment on the results of the post-hoc analysis using Bayes factors. We didn't just fail to reject the null hypothesis of no effect, but our analysis actually supported the null. Given the data, the null hypothesis was 3.8 to 6.1 times more likely than the alternative hypothesis. Jeffreys (1961) considers Bayes factors in this range to provide "substantial" support for the null, and, in an analysis of 855 published *t* tests, Wetzels et al. (2011) found that Bayes factors in this range generally correspond to *p*-values of 0.01 or

[18] Since we ended up dropping a participant, we actually only ran 29 subjects, which still results in a power of 0.89.

less (supposing we are trying to prove the null hypothesis is true). In other words, our results provide fair support for the assertion that, even though the face morphs look different in race when wearing different hairstyles, they don't look different in lightness.

General Discussion

We set out to examine whether the race of a face can influence how light its skin looks regardless of how light its skin actually is. In review:

Levin and Banaji (2006) reported that a prototypical African-American face appears to have darker skin than a prototypical Caucasian face even when the faces are matched for mean luminance and contrast (Figure 3.1). However, Firestone and Scholl (2015) showed that this effect is driven by the local luminance variations in the faces (a low-level confound) and not their race. Levin and Banaji (2006) also reported a lightness distortion with line-drawn faces that do not vary in local luminance (and thus are not confounded, Figure 3.2), but we argued that this result is not convincing because it is not consistent across observers. Finally, MacLin and Malpass (2001, 2003) reported that blended African-American/Hispanic faces are labeled darker in complexion when they are perceived to be African-American than when they are perceived to be Hispanic. Reasoning that this last result is the most compelling current evidence that race influences lightness perception, we set out to test its veracity using morphed African-American/Caucasian faces. But, we failed to replicate the result.

Why the discrepancy? Why did MacLin and Malpass find that race affects lightness judgments, while we did not? There are two important distinctions between our experiment and MacLin and Malpass' (2001, 2003) experiments that may explain this discrepancy, both of which suggest that their results may be due to response bias rather than altered lightness perception.

First, while we had participants make direct lightness judgments using gray-level samples, MacLin and Malpass used text anchor points (*light* and *dark*) so their participants did not make direct lightness judgments. This might encourage participants to respond how they think they are supposed to respond rather than in a way that matches their perceptual experience. At the very least, an experimental design with text-based responses makes it difficult to determine whether response bias might be a plausible explanation for the results (Firestone & Scholl, 2014).

Second, while we took special care to obscure the purpose of the study from our participants, MacLin and Malpass did not – it was clear to their participants that the study was about how race affects how people view a face. In our own experiments, we found that participants in Experiment 3.1 – exposed to instructions similar to those of MacLin and Malpass – were much better at guessing the research hypothesis than participants in Experiment 3.2, where the research hypothesis was obscured. This is problematic because it places a demand on the participants to respond in certain ways regardless their visual perceptions (Orne, 1962). For example, participants who are aware of the research hypothesis may respond in a way they believe with confirm the hypothesis in order to be "good" participants and not "ruin" the research (Nichols & Maner, 2008; Orne, 1962; Rubin, Paolini, & Crisp, 2010). Furthermore, this kind of bias has already been explicitly demonstrated in recent perceptual psychophysics research. Bhalla and Proffitt (1999) claimed that students perceive hills to be steeper when wearing a heavy backpack, but Durgin et al. (2009, 2012) showed that this effect is driven entirely by subjects who correctly guess the research hypothesis (i.e., that wearing a heavy backpack makes hills look steeper) and that the effect disappears when participants are given a deceptive cover story that justifies the backpack's presence (for other recent examples of response bias in perceptual psychophysics, see Firestone & Scholl, 2014; Shaffer et al., 2013).

Thus, we propose that MacLin and Malpass' participants may have judged African-American-looking faces as darker than Hispanic-looking faces because they felt they were supposed to do so – they *know* that African-American faces are darker than Hispanic faces – not because they *perceived* African-American faces as darker.

Furthermore, there is evidence of this kind of bias in the results of Levin and Banaji (2006). In their first experiment, they made no attempt to obscure the purpose of the experiment, telling participants that the experiment was about "how people perceive the shading of faces from different races" (p. 504). But, in their second experiment, they explicitly told participants the race of the faces – one group saw the Caucasian face labeled Caucasian and an Ambiguous face labeled African-American, and the other group saw the African-American face labeled African-American and the Ambiguous face labeled Caucasian – which makes it especially obvious to participants how they should respond. Indeed, Levin and Banaji reported that the effect size more than doubled (from 0.75 to 1.65) from their first experiment to their second experiment. They attributed this to a change in methodology (obtaining lightness judgments using gray-patches instead of other faces), but using the patch methodology, we obtained an effect size of only 0.96 for the African/Caucasian photographic faces (Experiment 3.1). Furthermore, their reported effect size of 1.65 is gigantic, more than twice what Cohen (1988) would consider a large effect (0.8). According to Cohen (1988, p. 27), an effect size of 0.8 should be blatantly obvious, or “grossly perceptible”, and he uses as his example the height difference between 13 and 18 year-old girls. In more technical terms, to obtain a power of 0.99 with an effect size of 1.65, one should only need 8 subjects. Now consider that this huge effect size was for the difference in lightness between the Ambiguous face and the African-American/Caucasian faces depicted in Figure 3.3. Is this difference in lightness more blatantly obvious than the difference in heights between 13 and 18 year-old girls? Certainly it’s not more *perceptually* obvious. We argue, then, that this difference must, at least in part, be due to response bias and, since MacLin and Malpass utilized similar instructions, that their results are also, at least in part, due to response bias.

Overall, we find no support for the assertion that the race of a face affects how light its skin looks. In this particular case then, it appears that there is no cognitive penetration of perception.

***

CONCLUSION

In this dissertation, I examined the surprising consistency of people's preferences for visual textures and for faces. I found that preferences for visual texture can be explained well using Palmer and Schloss' (2010) ecological valence theory. For faces, I found that none of the currently proposed mediators can explain why averaged configurations of faces are attractive and argued that averageness is, in itself, attractive. Finally, I examined whether low-level perception (specifically, of lightness) can be influenced by high-level cognitive factors. I found that the best current evidence for this does not hold up to scrutiny. Specifically, I found – contrary to Levin and Banaji (2006) – that the perceived race of a face does not influence how light its skin-tone looks.

Overall I found that, on the one hand, people are surprisingly consistent in their preferences, both for visual textures and for faces. On the other hand, I found no evidence that idiosyncratic cognitive factors can influence low-level perception.

References

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 229-251). Cambridge, MA: MIT Press.

Alley, T. R. (1988). The effects of growth and aging on facial aesthetics. In T. R. Alley (Ed.), *Social and applied aspects of perceiving faces* (pp. 51-62). Hillsdale, NJ: Erlbaum.

Alley, T. R., & Cunningham, M. R. (1991). Averaged faces are attractive, but very attractive faces are not average. *Psychological Science, 2*(2), 123-125.

Amadasun, M., & King, R. (1989). Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man and Cybernetics, 19*(5), 1264-1274.

Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision, 9*(11), 10.1-17.

Andersson, M. (1994). *Sexual Selection.* Princeton, NJ: Princeton University Press.

Apicella, C. L., Little, A. C., & Marlowe, F. W. (2007). Facial averageness and attractiveness in an isolated population of hunter-gatherers. *Perception, 36*, 1813-1820.

Ben-Shahar, O. (2006). Visual saliency and texture segregation without feature gradient. *Proceedings of the National Academy of Sciences of the USA, 103*(42), 15704-15709.

Bergen, J. R. (1991). Thoeries of visual texture perception. In D. Regan (Ed.), *Vision and visual dysfunction, vol. 10B* (pp. 114-134). New York: Macmillan.

Bhalla, M., & Proffitt, D. R. (1999). Visual-motor recalibration in geographical slant perception. *Journal of Experimental Psychology: Human Perception & Performance, 25*, 1076-1096.

Bomba, P. C., & Siqueland, E. R. (1983). The nature and structure of infant form categories. *Journal of Experimental Child Psychology, 35*(2), 294-328.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Brodatz, P. (1966). *Textures: a photographic album for artists and designers.* New York: Dover.

Bronstad, P. M., Langlois, J. H., & Russell, R. (2008). Computational models of facial attractiveness judgments. *Perception, 37*, 126-142.

Bruner, J. S., & Goodman, C. C. (1947). Value and need as organizing factors in perception. *Journal of Abnormal & Social Psychology, 42*, 33-44.

Carter, L. F., & Schlooler, K. (1949). Value, need, and other factors in perception. *Psychological Review, 56*, 200-207.

Chen, Y., Sobue, S., & Huang, X. (2008). Mapping functions of color image features and human KANSEI. *IIHMSP '08 International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 725-728). Harbin: IEEE.

Chronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 197-334.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Colins, J., & Olson, I. (2014). Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic Bulletin & Review, 21*, 843-860.

Coltman, D. W., & Slate, J. (2003). Microsatellite measures of inbreeding: a meta-analysis. *Evolution, 57*(5), 971-983.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Massons' method. *Tutorials in Quantitive Methods for Psychology, 1*, 42-45.

Daly, M., & Wilson, M. I. (1999). Human evolutionary psychology and animal behaviour. *Animal Behaviour, 57*, 509-519.

DeBruine, L. M., Jones, B. C., Unger, L., Little, A. C., & Feinberg, D. R. (2007). Dissociating averageness and attractiveness: Attractive faces are not always average. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1420-1430.

Deutsch, F. M., Zalenski, C. M., & Clark, M. E. (1986). Is there a double standard of aging? *Journal of Applied Social Psychology, 16*, 771-785.

Dong, Y., Tao, D., Li, X., Ma, J., & Pu, J. (2015). Texture classification and retrieval using shearlets and linear regression. *IEEE Transactions on Cybernetics, 45*(3), 358-369.

Dunning, D., & Balcetis, E. (2013). Wishful seeing: How preferences shape visual perception. *Current Directions in Psychological Science, 22*, 33-37.

Durgin, F. H., Baird, J. A., Greenburg, M., Russell, R., Shaugnessy, K., & Waymouth, S. (2009). Who is being deceived? The experimental demands of wearing a backpack. *Psychonomic Bulletin & Review, 16*, 964-969.

Durgin, F. H., Klein, B., Spiegel, A., Stawser, C. J., & Williams, M. (2012). The social psychology of perception experiments: Hills, backpacks, glucose, and the problem of generalizibility. *Journal of Experimental Psychology: Human Perception & Performance, 38*, 1582-1595.

Enquist, M., Ghirlanda, S., Lundqvist, D., & Wachtmeister, C.-A. (2002). An ethological theory of attractiveness. In R. Gillian, & L. A. Zebrowitz (Eds.), *Facial attractiveness: Evolutionary, cognitive, and social perspectives* (pp. 127-151). Westport, CT: Ablex Publishing.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods, 39*, 175-191.

Firestone, C. (2013). How "paternalistic" is spatial perception? Why wearing a heavy backpack doesn't -- and couldn't -- make hills look steeper. *Perspectives on Psychological Science, 8*, 455-473.

Firestone, C., & Scholl, B. J. (2014). "Top-down" effects where none should be found: The El Greco fallacy in perception research. *Psychological Science, 25*(1), 38-46.

Firestone, C., & Scholl, B. J. (2015). Can you experience "top-down" effects on perception?: The case of race categories and perceived lightness. *Psychonomic Bulletin & Review, 22*(3), 694-700.

Fodor, J. A. (1988). *The modularity of mind: An essay on faculty psychology.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Francis, G. (2012). The same old new look: Publication bias in a study of wishful seeing. *i-Perception, 3*, 176-178.

Galton, F. (1878). Composite portraits. *Nature, 18*, 97-100.

Galton, F. (1907). *Inquiries into human faculty and its development* (2nd ed.). New York: Elsevier North Holland.

Gatto, J. A., Porter, W. A., & Selleck, J. (2010). *Exploring visual design: The elements and principles* (4 ed.). Davis Publications.

Ghirlanda, S., Jansson, L., & Enquist, M. (2002). Chickens prefer beautiful humans. *Human Nature, 13*, 383-389.

Giese, M. A., & Leopold, D. A. (2005). Pysiologically inspired neural model for the encoding of faces spaces. *Neurocomputing, 65*, 93-101.

Grammer, K., & Thornhill, R. (1994). Human (Homo sampeins) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology, 108*, 233-242.

Guo, Z., Zhang, L., & Zhang, D. (2010). Rotation invariant texture classification using LBP variance (LBPV) with global matching. *Pattern Recognition, 43*(3), 706-719.

Haindl, M., & Filip, J. (2013). *Visual texture: accurate material appearance measurement, representation and modeling.* London: Springer-Verlag.

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., Pecher, D., Zeelenberg, R., Wai, L. I., & Winkielman, P. (2013). Two faces of attractiveness: Making beauty in averageness appear and reverse. *Psychological Science, 24*(11), 2343-2346.

Halit, H., de Haan, M., & Johnson, M. H. (2000). Modulation of event-related potentials by prototypical and atypical faces. *NeuroReport, 11*, 1871-1875.

Hampton, J. A. (1997). Psychological representations of concepts. In M. A. Conway (Ed.), *Cognitive models of memory* (pp. 81-110). Hove: Psychology Press.

Hampton, J. A. (2006). Concepts as prototypes. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 46, pp. 79-113). San Diego: Academic Press.

Haralik, R. M., Shanmugam, K., & Dinstein, I. (1973). Texture features for image classification. *IEEE Transactions on Systems, Man and Cybernetics, 3*(6), 610-621.

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

Hayes, A. F. (2013). *Indroduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York: The Guilford Press.

Henss, R. (1991). Perceiving age and attractiveness in facial photographs. *Journal of Applied Psychology, 21*, 933-946.

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2006). How direction of illumination affects visually perceived surface roughness. *Journal of Vision, 6*(5), 634-648.

Hoss, R. A., Ramsey, J. L., Griffin, A. M., & Langlois, J. H. (2005). The role of facial attractiveness and facial masculinity/femininity in sex classification of faces. *Perception, 34*, 1459-1474.

Iyengar, A., Kulkarni, R., & Vidya, T. N. (2015). Koinophilia and human facial attractiveness. *Resonance, 20*(4), 311-319.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving, 7*(1), 2-9.

Jeffreys, H. (1961). *Theory of probability.* Oxford, UK: Oxford University Press.

Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2004). When facial attractiveness is only skin deep. *Perception, 33*, 569-576.

Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q., Wang, J. Z., Li, J., & Luo, J. (2011). Aesthetics and emotions in images. *IEEE Signal Processing Magazine, 28*(5), 94-115.

Julesz, B. (1962). Visual pattern discrimination. *IEEE Transactions on Information Theory, 8*(2), 84-92.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature, 290*, 91-97.

Kawamoto, N., & Soen, T. (1993). Objective evaluation of color design II. *Color Research and Application, 18*(4), 260-266.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2), 137-152.

Kim, J., & Anderson, B. L. (2010). Image statistics and the perception of surface gloss and lightness. *Journal of Vision, 10*(9), 3.

Kim, S., Kim, E. Y., Jeong, K., & Kim, J. (2006). Emotion-based textile indexing using colors, texture and patterns. In *Advances in Visual Computing* (pp. 9-18). Berlin: Springer.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36*, 14. (ECVP Abstract Supplement).

Koeslag, J. H. (1990). Koinophilia groups sexual creatures into species, promotes stasis, and stabilizes social behaviour. *Journal of Theoretical Biology, 144*, 15-35.

Koeslag, J. H., & Koeslag, P. D. (1994). Koinophilia. *Journal of Theoretical Biology, 167*, 55-65.

Komori, M., Kawamura, S., & Ishihara, S. (2009a). Averageness or symmetry: Which is more important for facial attractiveness? *Acta Psychologica, 131*(2), 136-142.

Komori, M., Kawamura, S., & Ishihara, S. (2009b). Effect of averageness and sexual dimorphism on the judgment of facial attractiveness. *Vision Research, 49*(8), 862-869.

Landy, M. S., & Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision Research, 31*(4), 679-691.

Landy, M. S., & Graham, N. (2004). Visual perception of texture. In L. M. Chalupa, & J. S. Werner (Eds.), *The Visual Neurosciences* (pp. 1107-1118). Cambridge, MA: MIT Press.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1*(2), 115-121.

Langlois, J. H., Roggman, L. A., & Musselman, L. (1994). What is average and what is not average about attractive faces? *Psychological Science, 5*(4), 214-220.

Lee, J., & Park, E. (2011). Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia, 13*(5), 1031-1039.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *nature Neuroscience, 4*, 89-94.

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

Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces: the role of race categories. *Journal of Experimental Psychology: General, 135*, 501-512.

Lie, H. C., Rhodes, G., & Simmons, L. W. (2008). Genetic diversity revealed in human faces. *Evolution, 62*, 2473-2486.

Light, L. L., Hollander, S., & Kayra-Stuart, F. (1981). Why attractive people are harder to remember. *Personality and Social Psychology Bulletin, 7*, 269-276.

Little, A. C., & Hancock, P. J. (2002). The role of masculinity and distinctiveness in judgments of human male facial attractiveness. *British Journal of Psychology, 93*, 451-464.

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Philosophical Transactions of The Royal Society B, 366*, 1638-1659.

Liu, J., Lughofer, E., & Zeng, X. (2015). Could linear model bridge the gap between low-level statistical features and aesthetic emotions of visual textures? *Neurocomputing, 168*, 947-960.

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neurosciences, 8*, 1386-1390.

Lucassen, M. P., Gevers, T., & Gijsenij, A. (2011). Texture affects color emotion. *Color Research and Application, 36*, 426-436.

MacLin, O. H., & MacLin, M. K. (2011). The role of racial markers in race perception and racial categorization. In R. Adams, N. Ambady, K. Nakayama, & S. Shimojo (Eds.), *The science of social vision* (pp. 321-346). New York: Oxford University Press.

MacLin, O. H., & Malpass, R. (2003). The ambiguous-race face illusion. *Perception, 32*, 249-252.

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces. *Psychology, Public Policy, & Law, 7*, 98-118.

Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy & Phenomenological Research, 84*, 24-62.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A, 7*(5), 923-932.

Malpass, R. S., & Hughes, K. D. (1986). Formation of facial prototypes. In H. D. Ellis, M. A. Jeeves, F. Newcombe, & F. Young (Eds.), *Aspects of face processing* (pp. 154-162). Boston: Martinus Nijhoff Publishers.

Markow, T. A., & Gottesman, I. I. (1993). Behavioral phenodeviance: A lerneresque conjecture. *Genetica, 89*, 297-305.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: W.H. Freeman.

Mathes, E. W., Brennan, S. M., Haugen, P. M., & Rice, H. B. (1985). Ratings of physical attractiveness as a function of age. *Journal of Social Psychology, 125*, 163-166.

Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: A duel process approach. *Applied Cognitive Psychology, 19*, 545-567.

Mitton, J. B., & Grant, M. C. (1984). Associations among protein heterozygosity, growth rate and developmental homeostasis. *Annual Review of Ecology and Systematics, 15*, 479-499.

Montoya, A. K., & Hayes, A. F. (submitted). Two-condition within-participant statistical mediation analysis: A path-analytic framework.

Montoyishi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447*, 206-209.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau. *Tutorials in Quantitative Methods, 4*, 61-64.

Morris, P. E., & Wickham, L. H. (2001). Typicality and face recognition: A critical re-evaluation of the two-factor theory. *Quarterly Journal of Experimental Psychology, 54*(A), 863-877.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology, 135*(2), 61-64.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 7*(11), 776-783.

Palmer, S. E., & Schloss, K. B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences of the USA, 107*, 8877-8882.

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology, 64*, 77-107.

Peery, D., & Bodenhause, G. V. (2008). Black + white = black: Hypodescent in reflexive categorization of racially ambiguous faces. *Psychological Science, 10*, 973-977.

Perrett, D. I., Burt, D. M., Penton-Voak, I. S., Lee, K. J., Rowland, D. A., & Edwards, R. (1999). Symmetry and human facial attractiveness. *Evolution & Human Behavior, 20*, 295-307.

Perrett, D. I., Lee, K. J., Penton-Voak, D., Rowland, S., Yoshikawa, D. M., Burt, S. P., . . . Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature, 394*, 884-887.

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

Plylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22*, 341-365.

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complext wavelent coefficients. *International Journal of Computer Vision, 40*(1), 49-71.

Principe, C. P., & Langlois, J. H. (2011). Faces differing in attractiveness elicit corresponding affective responses. *Cognition & Emotion, 25*(1), 140-148.

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

Proffitt, D. R. (2006). Embodied perception and the economy of action. *Perspectives on Psychological Science, 1*, 110-122.

Randen, T., & Husy, J. H. (1999). Filtering for texture classification: a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(4), 291-310.

Rao, A. R., & Lohs, G. (1993). *Towards a texture naming system: Identifying relevant dimensions of texture perception.* IBM Research Technical Report, No. RC19140.

Rao, A. R., & Lohse, G. L. (1996). Towards a texture naming system: Identifying relevant dimensions of texture. *Vision Research, 36*(11), 1649-1669.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*, 199-226.

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological Science, 7*(2), 105-110.

Rhodes, G., Halberstadt, J., Jeffery, L., & Palermo, R. (2005). The attractiveness of average faces is not a generalized mere exposure effect. *Social Cognition, 23*(3), 205-217.

Rhodes, G., Harwood, K., Yoshikawa, S., Nishitani, M., & McLean, I. G. (2002). The attractiveness of average facial configurations: Cross-cultural evidence and the biology of beauty. In G. Rhodes, & L. A. Zebrowitz (Eds.), *Facial attractiveness: Evolutionary, cognitive and social perceptives.* Westoprt, CT: Ablex.

Rhodes, G., Hickford, C., & Jeffery, L. (2000). Sex-typicality and attractiveness: Are supermale and superfemale faces super-attractive? *British Journal of Psychology, 91*, 125-140.

Rhodes, G., Jeffery, L., Watson, T. L., Jaquet, E., Winkler, C., & Clifford, C. W. (2004). Orientation-contingent face aftereffects and implications for face-coding mechanisms. *Current Biology, 14*, 2119-2123.

Rhodes, G., Profitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review, 5*, 659-669.

Rhodes, G., Sakiko, Y., Palermo, R., Simmons, L. W., Peters, M., Lee, K., . . . Crawford, J. R. (2007). Perceived health contributes to the attractiveness of facial symmetry, averageness, and sexual dimorphism. *Perception, 36*, 1244-1252.

Rhodes, G., Sumich, A., & Byatt, G. (1999). Are average facial configurations attractive only because of their symmetry? *Psychological Science, 10*(1), 52-58.

Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., & Akamatsu, S. (2001a). Attractiveness of facial averageness and symmetry in non-Western populations: In search of biologically based standards of beauty. *Perception, 30*, 611-625.

Rhodes, G., Zebrowitz, L. A., Clark, A., Kalick, M., Hightower, A., & McKay, R. (2001b). Do facial averageness and symmetry signal health? *Evolution and Human Behavior, 22*, 31-46.

Ritter, N. L. (2010). *Understanding a widely misunderstood statistic: Chronbach's alpha.* New Orleans, LA: Paper presented at Southwestern Educational Research Association Converence.

Roberts, S. C., & Little, A. C. (2008). Good genes, complementary genes and human mate preferences. *Genetica, 134*(1), 31-43.

Roberts, S. C., Little, A. C., Gosling, L. M., Perrett, D. I., Carter, V., Jones, B. C., . . . Petrie, M. (2005). MHC-heterozygosity and human facial attractiveness. *Evolution and Human Behavior, 26*, 213-216.

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-47). Hillsdale, NJ: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. B., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.

Rosenholtz, R. (2000). Significantly different textures: a computational model of pre-attentive texture segmentation. In V. David, *Computer Vision -- ECCV 2000* (pp. 197-2111). Berlin: Springer.

Rosenholtz, R. (2015). Texture perception. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization.* Oxford, UK: Oxford University Press.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Baysian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

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

Rubenstein, A. J., Langlois, J. H., & Roggman, L. A. (2002). What makes a face attractive and why: The role of averageness in defining facial beauty. In G. Rhodes, & L. A. Zebrowitz (Eds.), *Facial attractiveness: Evolutionary, cognitive, and social perspectives.* (pp. 1-33). Westport, CT: Ablex Publishing.

Rubin, M., Paolini, S., & Crisp, R. J. (2010). A processing fluency explanation of bias against migrants. *Journal of Experimental Social Psychology, 46*, 21-28.

Said, C. P., & Todorov, A. (2011). A statistical model of facial attractiveness. *Psychological Science, 22*(9), 1183-1190.

Schloss, K. B., & Palmer, S. E. (2014). The politics of color: preference for republican-red vs. democtratic-blue. *Psychonomic Bulletin and Reivew, 21*, 1481-1488.

Schloss, K. B., Goldberger, C. S., Palmer, S. E., & Levitan, C. A. (2015). What's that smell? An ecological approach to understanding preferences for familiar odors. *Perception, 44*, 23-38.

Schloss, K. B., Hawthorne-Madell, D., & Palmer, S. E. (in press). Ecological influences on individual differences in color preference. *Attention, Perception, and Psychophysics*.

Schloss, K. B., Poggesi, R. M., & Palmer, S. E. (2011). Effects of university affiliation and "school spirit" on color preferences: Berkeley versus Stanford. *Psychonomic Bulletin and Review, 18*, 498-504.

Sebe, N., & Lew, M. S. (2001). Texture features for content-based retrieval. In *Principles of visual information retrieval* (pp. 51-85). London: Springer-Verlag.

Shaffer, D. M., McManama, E., Swank, C., & Durgin, F. H. (2013). Sugar and space? Not the case: Effects of low blood-glucose on slant estimation are mediated by beliefs. *i-Perception, 4*, 147-155.

Soule, M., & Cuzin-Roudy, J. (1982). Allometric variation, 2: Developmental instability of extreme phenotypes. *American Naturalist, 120*, 765-786.

Strauss, E. D., Schloss, K. B., & Palmer, S. E. (2013). Color preferences change after experience with liked/disliked colored objects. *Psychonimic Bulletin and Review, 20*, 935-943.

Symons, D. (1979). *The evolution of human sexuality.* London: Oxford University Press.

Tamura, H., Mori, S., & Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics, 8*(6), 460-473.

Taylor, C., & Franklin, A. (2012). The relationship between color-object associations and color preference: further investigation of the ecological valence theory. *Psychonomic Bulletin & Review, 19*, 190-197.

Thornhill, R., & Gangestad, S. W. (1996). The evolution of human sexuality. *Trends in Ecology & Evolution, 11*, 98-102.

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences, 3*, 452-460.

Thumfart, S., Heidl, W., Scharinger, J., & Eitzinger, C. (2009). A quantitative evaluation of texture feature robustness and interpolation behaviour. In X. Jaing, & N. Petkov (Eds.), *Computer analysis of images and patters: Lecture notes in computer science* (Vol. 5702, pp. 1154-1161). Berlin: Springer.

Thumfart, S., Jacobs, R. H., Lughofer, E., Eitzinger, C., Cornelissen, F. W., Groissboeck, W., & Richter, R. (2011). Modeling human aesthetic perception of visual textures. *ACM Transactions on Applied Perception, 8*(4), Article 27.

Trujillo, L. T., Jankowitsch, J. M., & Langlois, J. H. (2014). Beaty is in the ease of the beholding: A neurophysiological test of the averageness theory of facial attractiveness. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 1061-1076.

Tuceryan, M., & Jain, A. (1998). Texture analysis. In C. H. Chen, L. F. pau, & P. S. Wang (Eds.), *The Handbook of Pattern Recognition and Computer Vision* (2nd ed., pp. 207-248). New Jersey: World Scientific Publishing Co.

Tyler, C. W. (2004). Theory of texture discrimination of based on higher-order perturbations in individual texture samples. *Vision Research, 44*(18), 2179-2186.

Um, J., Eum, K., & Lee, J. (2002). A study of the emotional evaluation models of color patterns based on the adaptive fuzzy system and the neural network. *Color Research and Application, 27*(3), 208-216.

Unnikrishnan, M. K. (2012). Koinophilia revisited: the evolutionary link between mate selection and face recognition. *Current Science, 102*(4), 563-570.

Valentine. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43*, 161-204.

Valentine, T., Darling, S., & Donnelly, M. (2004). Why are average faces attractive? The effect of view and averageness on the attractiveness of female faces. *Psychonomic Bullent & Review, 11*(3), 482-487.

Valenzano, D., Mennucci, A., Tartarelli, G., & Cellerino, A. (2006). Shape analysis of female facial attractiveness. *Vision Research, 46*, 1282-1291.

Varma, M., & Zisserman, A. (2005). A statistical approach to texture classification from single images. *International Journal of Computer Vision, 62*, 61-81.

Victor, J. D. (1988). Models for preattentive texture discrimination: Fourier analysis and local feature processing in a unified framework. *Spatial Vision, 3*(4), 263-280.

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

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291-298.

Wickham, L. H., & Morris, P. E. (2003). Distinctiveness and recognition of faces: Attractive faces can be typical or distinctive but are not better recognized. *The American Journal of Psychology, 116*(3), 455-468.

Wickham, L. H., Morris, P. E., & Fritz, C. O. (2000). Facial distinctiveness: Its measurements, distribution and influence on immediate and delayed recognition. *British Journal of Psychology, 91*, 99-124.

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

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17*, 799-808.

Witt, J. K. (2011). Actions effect on perception. *Current Directions in Psychological Science, 20*, 201-206.

Yokosawa, K., Yano, N., Schloss, K. B., Prado-Leon, L. R., & Palmer, S. E. (2010). Cross-culteral studies of color preferences: US, Japan, and Mexico. *Journal of Vision, 10*(7), 408.

Zadra, J. R., & Clore, G. L. (2011). Emotion and perception: The role of affective information. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*, 676-685.

Zebrowitz, L. A., Voinescu, L., & Collins, M. A. (1996). Wide-eyed and crooked-faced -- Determinants of perceived and real honesty across the lifespan. *Personality and Social Psychology Bulletin, 22*, 1258-1269.

Zeimbekis, J., & Raftopoulos, A. (Eds.). (2015). *The cognitive penetrability of perception: New philosophical perspectives.* New York: Oxford University Press.

Zhang, H., Augilius, E., Honkela, T., Laaksonen, J., Gamper, H., & Alene, H. (2011). Aanalyzing emotional semantics of abstract art using low-level image features. *Proceddings of the 10th International Conference on Advances in Intelligent Data Analysis* (pp. 413-423). Berlin: Springer-Verlag.

Appendix A

Table A1shows all of the information needed to run the calculations in Chapter 1, for the 47 visual textures that we predicted preferences for. Table A2 shows information for the remaining 15 textures whose preferences we did not predict. The title (first column for both tables) of each texture is either the description given by Brodatz (1966), with his numbering in parentheses, or the search term that produced the image on Shutterstock.

Table A1. For the 47 visual textures whose preference was predicted, shows the average preference rating, rating predicted by the ecological valence theory, number of descriptions, average premium rating, most frequently associated object description with its average valence rating, and namesake description with its average valence rating. All rating scales vary from -100 to +100. Note: there is no valence rating for pressed cork because it was not listed during the object description task, and the blurred chocolate and mud were not given namesake descriptions.

Visual Texture	Preference	Ecological Prediction	nDescriptions	Premium	Most Frequent Description (MFD)	Valence of MFD	Namesake Description (ND)	Valence of ND
wire mesh (D1)	2.46	2.05	17	14.77	fence / gate	8.67	wire mesh	-13.33
crocodile skin (D10)	-27.65	1.10	29	-7.98	reptile skin (snake, crocodile, lizard)	-25.66	reptile skin (snake, crocodile, lizard)	-25.66
Japanese rice paper (D107)	-11.96	3.99	24	-31.49	branches / roots (of a tree)	15.86	paper	25.66
grassy fiber (D110)	-9.16	6.44	16	-25.24	grass	31.57	grass	31.57
plastic bubbles (D112)	-16.52	5.31	24	-5.72	bee hive	-29.03	bubbles	35.6
straw (D15)	10.7	9.24	14	-34.35	grass	31.57	straw	10.94
raffia weave (D18)	21.64	11.44	21	-15.65	woven basket	20.31	woven basket	20.31
beach pebbles (D23)	31.31	20.13	12	7.75	rocks / pebbles / stones	22.35	rocks/pebbles/stones	22.35
brick wall (D26)	19.21	6.28	6	-6.41	bricks (brick wall / brick building / brick sidewalk)	4.42	bricks (brick wall / brick building / brick sidewalk)	4.42
reptile skin (D3)	-26.96	1.72	25	3.99	reptile skin (snake, crocodile, lizard)	-25.66	reptile skin (snake, crocodile, lizard)	-25.66
beach pebbles (D30)	32.32	12.90	13	21.21	rocks / pebbles / stones	22.35	rocks/pebbles/stones	22.35
pressed cork (D32)	2.76	4.43	22	-28.99	wall of a house / building	13.02	pressed cork	--
netting (D34)	-6	2.07	21	-6.64	net / netting	-6.08	net / netting	-6.08
water (D37)	26.7	24.46	9	14.85	ocean / sea	67.34	ocean / sea	67.34
lace (D41)	38.29	28.34	17	35.32	flowers	59.28	lace / embroidery	35.15
mica (D5)	6.13	13.64	14	-17.51	rocks / pebbles / stones	22.35	wall of a house / building	13.02
raffia woven with cotton (D50)	14.74	11.55	12	-10.07	tree bark	12.47	cloth / fabric	27.49
oriental straw cloth (D52)	-6.85	7.61	19	-17.04	cloth / fabric	27.49	cloth / fabric	27.49
handmade paper (D57)	-10.11	8.64	20	-29.71	wall of a house / building	13.02	paper	25.66
European marble (D58)	-12.87	-0.12	26	-28.28	ocean / sea	67.34	granite / marble	31.39
European marble (D61)	-15.84	4.76	24	-8.30	rocks / pebbles / stones	22.35	granite / marble	31.39
European marble (D62)	0.62	14.52	14	0.89	rocks / pebbles / stones	22.35	granite / marble	31.39
handwoven rattan (D64)	30.31	10.30	16	8.39	woven basket	20.31	woven basket	20.31
wood grain (D69)	19.19	13.40	11	28.18	tree bark	12.47	wood	28.52
wood grain (D71)	22.48	10.48	18	24.25	wood	28.52	wood	28.52
tree stump (D72)	19.39	13.01	6	-4.79	tree bark	12.47	tree bark	12.47
soap bubbles (D73)	-16.19	4.18	34	-32.20	bubbles	35.6	bubbles	35.6
coffee beans (D74)	13.69	13.04	12	17.61	coffee beans	40.37	coffee beans	40.37
Oriental straw cloth (D78)	10.15	14.26	18	-21.13	wood	28.52	cloth / fabric	27.49
ceiling tile (D86)	-17.99	5.65	22	-39.24	wall of a house / building	13.02	wall of a house / building	13.02
fossilized sea fan with coral (D87)	-2.99	6.94	23	-4.01	leaves	34.87	coral / coral reef	42.84
dried hop flowers (D88)	10.15	8.11	20	1.24	insects / bugs	-35.54	flowers	59.28
grass lawn (D9)	-21.42	2.72	21	-41.91	grass	31.57	grass	31.57
clouds (D90)	0.83	3.49	16	6.46	clouds	61.82	clouds	61.82
clouds (D91)	63.88	22.74	8	27.75	clouds	61.82	clouds	61.82
fur, hide of unborn calf (D93)	10.78	5.47	6	5.47	animal fur (e.g., dog, cat)	6.69	animal fur (e.g., dog, cat)	6.69
crushed rose quartz (D98)	18.11	12.65	16	8.60	rocks / pebbles / stones	22.35	crystals / diamonds / gems	52.46
chocolate spread	28.67	10.07	20	38.47	chocolate	61.7	chocolate	61.7
chocolate (Gaussian blurred)	-54.01	4.17	21	-43.37	X-ray / ultrasound	-1.04	--	--
mud	-51.91	-1.52	23	-49.88	mud	-21.09	mud	-21.09
mud (Gaussian blurred)	-66.27	1.64	21	-63.15	people	40.24	--	--
lettuce	49.81	22.30	15	34.32	flowers	59.28	lettuce	37.99
mold	-23.27	1.50	29	-41.54	dirt / gravel / soil	-9.74	mold	-66.16
rotted strawberry	-48.37	-0.27	27	-25.08	beans	8.09	rotten / spoiled food	-74.77
strawberry	-12.57	9.31	27	1.84	eggs	33.25	strawberry	49.51
infected skin	-26.48	0.78	28	-17.61	ocean / sea	67.34	infection	-71.47
tropical sea	36.73	31.83	9	30.20	ocean / sea	67.34	ocean / sea	67.34

Table A2. For the 15 visual textures whose preference was not predicted, shows the average preference rating, number of descriptions, average premium rating, and most frequently associated object description with its average valence rating**.** All rating scales vary from -100 to +100.

Visual Texture	Preference	nDescriptions	Premium	Most Frequent Description (MFD)	Valence of MFD
woven cane (D101)	1.88	19	15.22	cloth / fabric	27.49
woolen cloth (D11)	12.13	17	-6.75	clothing (shirt / pants / sweater / jacket)	57.07
brick wall (D25)	18.43	6	-11.12	bricks (brick wall / brick building / brick sidewalk)	4.42
beach sand (D29)	-5.24	17	-29.34	cement / concrete / pavement	7.43
beach pebbles (D31)	45.84	10	19.52	rocks / pebbles / stones	22.35
lace (D39)	11.18	28	7.95	flowers	59.28
lace (D40)	38.27	23	42.47	flowers	59.28
lace (D42)	35.24	23	44.54	curtain	20.1
straw matting (D55)	-2.11	20	-1.38	clothing (shirt / pants / sweater / jacket)	57.07
European marble (D63)	-21.67	31	-19.67	spider web	-40.72
Oriental straw cloth (D80)	-20.28	21	-28.33	carpet / rug	15.51
Oriental straw cloth (D82)	-1.41	24	-7.00	carpet / rug	15.51
woven matting (D83)	7.91	21	-0.58	cloth / fabric	27.49
dried hop flowers (D89)	-26.21	23	-8.20	insects / bugs	-35.54
brick wall (D94)	31.42	5	3.01	bricks (brick wall / brick building / brick sidewalk)	4.42