MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master‘s Thesis

Variance of Effect Sizes of Facial Attractiveness

verfasst von / submitted by

Patrick Smela, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie UG2002
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Psych. Dr. Helmut Leder

Acknowledgements

I would like to thank Lei Zhang, whose seminar on Bayesian modelling has inspired the design of this study. I would further like to thank Koyo Nakamura and Cliodhna Quigley for the discussions and insights they provided. Additionally, I would like to thank Christina Krumpholz and Helmut Leder for giving me the opportunity to join the project this thesis is based on and for providing me the freedom to be creative with the research design. Lastly, I would like to thank Eda Kir and Dominique Vogler for their support during the study and writing process.

Table of Contents

Abstract

There is a growing recognition for the need to further differentiate effects in psychological facial attractiveness research, such as distinguishing between top-down and bottom-up components (Stolier et al., 2018). Disentangling the internal mechanisms of the effects investigated in facial attractiveness studies is beneficial for the entire field because it helps to broaden our understanding of how people perceive each other. This study will emphasize the importance of critically examining prevalent effects that are heavily discussed in the field. Three aspects will be of particular interest: (1) taking a holistic approach by examining the mutual influence of multiple variables simultaneously; (2) contextualizing main effects in terms of individual face’s effects; (3) expanding on the conventional hypothesis testing approach by instead using effect sizes of component traits to assess the practical meaningfulness of facial traits and attractiveness judgements. To achieve these goals, a hierarchical multilinear regression using Bayesian inferential statistics was employed. This model allows a quantification of effect sizes for individual faces as well as of the overall effect, typically discussed as the general effect. Therefore, the interstimulus variance of an association between two variables - a good indicator of the universality of an effect - can be investigated. The use of hierarchical multilinear models to investigate individual face’s effects and understand practical meaningfulness for the included variables will thus be exemplified.

Keywords: Facial Attractiveness; Hierarchical Model; Multilinear Model; Bayesian inference; Effect Size Estimation

Zusammenfassung

Es gibt eine wachsende Anerkennung für die Notwendigkeit, die Effekte psychologischer Gesichtsattraktivitätsforschung besser zu differenzieren, wie zum Beispiel zwischen top-down und bottom-up Komponenten zu unterscheiden (Stolier et al., 2018). Die internen Wirkungsmechanismen der Effekte die in Studien zur Gesichtsattraktivität untersucht werden zu entflechten ist für das gesamte Feld von Vorteil, da dies dazu beiträgt, unser Verständnis darüber zu erweitern, wie sich Menschen gegenseitig wahrnehmen. Diese Studie betont die Wichtigkeit sich kritisch mit den wichtigen Effekten auseinanderzusetzen, die in diesem Bereich besonders ausführlich diskutiert werden. Drei Aspekte sind dabei von besonderem Interesse: (1) eine holistischen Ansatz zu wählen, bei dem der gemeinsame Einfluss mehrerer Variablen gleichzeitig berücksichtigt wird; (2) Haupteffekte im Bezug auf die Effekte individueller Gesichter zu verstehen; (3) den herkömmlichen hypothesentestenden Ansatz zu erweitern indem stattdessen Effektstärken zur Feststellung von praktischer Relevanz von Gesichtsmerkmalen und Attraktivitätsbewertungen herangezogen werden. Um diese Ziele zu erreichen wurde eine hierarchische multilineare Regression auf Grundlage bayesianischer Inferenzstatistik verwendet. Dieses Modell erlaubt es sowohl Effektstärken für individuelle Gesichter als auch einen Gesamteffekt zu quantifizieren. Auf diese Weise kann die Interstimulusvarianz von Zusammenhängen zweier Variablen untersucht werden, die einen guten Indikator für die Universalität eines Effektes darstellt. Die Anwendung hierarchische multilinearer Modelle zur Untersuchung von Effekten einzelner Gesichter und dem Verständnis der praktischen Bedeutung dieser Effekte wird in dieser Studie veranschaulicht.

Schlagworte: Gesichtsattraktivität; Hierarchische Modelle; Multilineare Modelle; Bayesianische Inferenzstatistik; Effektgrößenschätzung

Variance of Effect Sizes of Facial Attractiveness

> In one way or another esthetic pleasure is felt by all members of mankind. No matter how diverse the ideals of beauty may be, the general character of the enjoyment of beauty is of the same order everywhere. (Boas, 1955, p. 9)

Facial attractiveness is an omnipresent topic in our lives and is the topic of intensive psychological research. Previous studies have been concerned with the causal explanations of facial attractiveness, which in most cases means either researching the facial traits that lead to increased perceived facial attractiveness or researching the consequences of facial attractiveness on human partner choice as well as the general implications for social interactions in public and private spaces. Facial attractiveness research has found a myriad of ways how attractiveness influences the social environments (Brand et al., 2012; Langlois et al., 2000). These effects can be subtle, such as attractive faces capturing attention (Chen et al., 2012; Leder et al., 2010; Leder, Mitrovic, et al., 2016; Valuch et al., 2015) or shaping the way they are perceived emotionally, since attractive faces are associated with positive emotion, while unattractive faces are associated with negative emotions (Gerger et al., 2011). Attractiveness can even shape the way humans think about each other and influence judgements such as competence or morality judgements (Landy & Sigall, 1974; Tsukiura & Cabeza, 2011).

Since facial attractiveness is such an important factor in social situations, the presented thesis aims to expand on the understanding of factors that lead to facial attractiveness. Previous efforts have been spent to confirm theoretical assumptions about the facial features that are perceived as attractive. This practice has led to a hypothesis testing approach in which the confirmation of theoretically assumed effects are of primary concern instead of a broader discussion which includes effect sizes and practical performance as well. Therefore, this study will aim to address these shortcomings by broadening the current methodological understanding that is generally used to assess effects. Instead of using an approach where data is used in a model that simply aggregates across a full dataset at once, a model is applied that makes use of variances between faces to get a more holistic understanding of how participants perceive individual faces and that takes into account that the reasons why faces are perceived as attractive might vary from face to face. Additionally, this study aims to demonstrating how modern statistical models help to understand the practical meaningfulness by applying the approach using the most important factors in the field of facial attractiveness research.

In order to realize the goals of this study, it is important to understand the methodological developments it draws from. In total, three different kinds of developments that have shaped the design of this study: 1) the trend from singular causal research designs towards more interactive research designs incorporating multiple variables, 2) the trend towards more ecologically valid stimuli and methodological concepts, and 3) the trend towards more data-driven techniques that attempt to optimally describe data, thereby focussing on applicability rather than testing specific, isolated theoretical assumptions. All of these trends should not be understood as mutually exclusive from one another and in fact build upon each other to understand facial attractiveness, especially 1) and 2). Before discussing the methodological developments that drive this study, the following paragraphs will first introduce the theoretical foundation of facial attractiveness research as well as the variables included in this study to contextualize the considerations these developments are based on and which gaps in the field they are trying to address.

Theoretical Background

Facial attractiveness research is closely linked to mating behaviour (Buss & Barnes, 1986; Grammer & Thornhill, 1994; Jones, 1995) and is also often associated with aesthetics in general. There is a clear evolutionary undertone in a lot of conceptualizations that are used in the field, and often the development of an aesthetic sense is linked to the development strategies such as the tail the male peacock waves around to attract females (Dissanayake, 2015). In this view, facial attractiveness perception is innately rooted in the brain, a claim supported by the finding that even infants show preferences for attractive faces (Langlois et al., 1991). Previous research has also shown that attractiveness is processed very rapidly in the brain (Schacht et al., 2008), as well as activating brain regions associated with reward (Aharon et al., 2001; Chatterjee et al., 2009). The effects of beauty in general seem to be so deeply ingrained in human biology that they even extends to the perception of usability of an object (Tractinsky et al., 2000) and they are often discussed in other forms of aesthetic experiences, most prevalent in art perception (Leder et al., 2004; Leder & Nadal, 2014; Mitrovic et al., 2020). Taken together, these findings illustrate a strong biological foundation for an aesthetic sense in humans.

The evolutionary framework underlying mate choice research also presents researchers with mechanisms to explain and predict the effects of facial attractiveness theoretically. These arguments are derived from sexual selection principles since behaviour in the evolutionary framework birthing and raising the following generation is seen as the primary goal of mate choice. Sexual selection criteria are therefore often at the core of many arguments and explanations why certain facial traits are perceived as more attractive than others. Consequently, sex differences are frequently emphasized and researched in the field of mate choice (Buss & Barnes, 1986). For example, women should seek men with large resources to help them raise offspring, while men on the other hand seek to find fit, attractive partners that have a higher chance to give birth to healthy offspring.

There is, however, still some discussion about the exact mechanism of sexual selection because scholars disagree on whether it is mainly driven by an aesthetic component or rather if certain traits might signal some underlying genetic fitness to the perceiver. Contemporary literature seems to make frequent use of the latter interpretation to formulate causal arguments (Prum, 2012). This may be due to the fact that it allows researchers to derive arguments and causal mechanisms by understanding facial traits an indicator of genetical fitness (Little et al., 2011). For example, the symmetry of an organism has been linked to developmental conditions and might therefore signal stable development which in turn is perceived as attractive (Gangestad et al., 1994).

One implicit assumption of the evolutionary framework is the universality of causal relationships between facial traits and facial attractiveness. It is assumed that more or less every human finds these traits attractive in a very similar manner or that all faces are attractive for the same reasons. Typically only sex (for example, Buss & Barnes, 1986) or cultural differences (for example, Che et al., 2018) are accounted for. These assumptions typically leave out important factors such as personal preferences or face to face variance. The presented work aims to fill this gap by exemplifying how these factors can be taken into account by applying modern statistical methods to seven facial traits that stem mostly from the biological framework. The included variables are: averageness, sexual dimorphism, health, likeability, trust, dominance and beauty. Since evidence for the causal relationship between each of these traits and facial attractiveness has been demonstrated in previous studies discussed in the following section, the main focus is to address not the causal relationships but rather to investigate the overall effect these traits have on facial attractiveness and how their effect sizes vary across individual faces.

Averageness

Averageness is one of the most prevalent traits in facial attractiveness research. It refers to the concept that a face becomes more attractive the closer it resembles the mathematical average across all faces for a population (Rhodes, 2006). The more average a face is the less distinctive it is. Averageness is perceived as attractive since it might signal developmental stability (Thornhill & Møller, 1997) and thereby health and better genes.

One popular method to experimentally investigate the effect of averageness is called morphing, the process in which facial images are computer generated by merging different sets of faces together (Langlois et al., 1994). One of the most influential papers that used the morphing technique dates back to the early 1990s where Langlois and Roggman (1990) demonstrated that the computer-generated faces were perceived as more attractive than their natural counterparts. Many studies since have followed up on these findings by using digitally manipulating faces and have found similar associations between averageness and facial attractiveness (Langlois et al., 1994; Little & Hancock, 2002; Vingilis-Jaremko et al., 2014).

The downside of the morphing techniques is that faces do not only become more average but consequently also more symmetrical by averaging across a large number of faces. This might mean that the effect averageness has on facial attractiveness might be in part an experimental artifact as a result of the morphing technique and explained by the symmetry of the face alone, a facial trait which has also been linked to facial attractiveness (Cárdenas & Harris, 2006; Gangestad et al., 1994), albeit in a disputed manner (Leder et al., 2019). Rhodes et al. (1999) found however, that the effect of averageness remains, even if the effects of symmetry are statistically controlled for.

A second technique is based on caricaturing the facial stimuli presented in an experiment in order to exaggerate the individual face from the average face. This technique employs the opposite approach to morphing, reducing the averageness of the face and increasing its distinctiveness. Rhodes and Tremewan (1996) found that attractiveness ratings were lower for caricatures, indicating that averageness is associated with higher attractiveness. Similarly, Trujillo et al. (2014) found that average faces were perceived as more attractive, while distinctiveness showed a negative relationship to facial attractiveness.

Not many studies have been conducted with natural, unmanipulated stimuli, however, focussing on male faces, Light et al. (1981) found that the more attractive faces were rated as less distinctive, while Catz and Lewis (2020) found that averageness increased facial attractiveness for male faces but not female faces. To properly assess the association between averageness and facial attractiveness Rhodes (2006) ran a meta-analysis that included 45 studies. A large overall effect was reported (0.52 ± 0.41, M ± SD), but studies using manipulated faces reported larger effects (0.67 ± 0.43, M ± SD) than studies using real faces (0.40 ± 0.33, M ± SD). In this case the effects might be more pronounced for the manipulated faces since the manipulation increases differences between average and non-average faces to a larger degree than natural fluctuation and therefore leads to more pronounced effects.

Nonetheless, these findings illustrate, that there might be effect size variations between the different faces.

Sexual Dimorphism

Sexual dimorphism is based on the sex differences in in secondary sexual characteristics between men and women and as such depend on sex typical facial features, and the effects are generally discussed in terms of femininity and masculinity separately. Typically, masculine features include such things as a large jaw, prominent cheekbones and a prominent brow ridge (Little & Hancock, 2002), while feminine features include such things as high cheekbones and baby-facedness like large eyes and lips and a small nose. Sexual dimorphic characteristics might be perceived as attractive since they are based on hormonal changes during puberty and signal good hormonal development and a strong immune system (Rhodes et al., 2000). Similar to the techniques used for averageness, some studies used digitally manipulated faces (Grammer & Thornhill, 1994; Perrett et al., 1998), while others used natural photographs (Cunningham, 1986; Jones & Hill, 1993).

Several studies have found a positive link between femininity and perceived facial attractiveness (Catz & Lewis, 2020; Koehler et al., 2004; Perrett et al., 1994; Rhodes et al., 2000, 2003) in female faces. Despite the concept being based on sex specific facial features, the effects of femininity are not just limited to female faces. It has also been found that feminized not just female but also male faces are also perceived as more attractive (Nakamura & Watanabe, 2020; Perrett et al., 1998; Tan & Zheng, 2020). It has also been found that effects of femininity depend on general social knowledge about a person (Wang et al., 2020).

The results for masculinity in male and female faces on the other hand are more ambiguous. The findings of femininity increasing attractiveness in male faces call into question whether male or female facial attractiveness would be positively associated with masculinity. This might be the reason why some studies reported findings for specific masculine features instead of masculinity itself: for example Cunningham et al. (1990) and Grammer and Thornhill (1994) who found that women preferred large jaws in men. Sometimes masculine features are also discussed in terms of their relation with dominance (Grammer & Thornhill, 1994; Perrett et al., 1998), one of the traits further discussed later on. Nevertheless, a lot of studies have demonstrated an effect for masculinity (Johnston et al., 2001; Koehler et al., 2004). The effects of masculinity are also more pronounced in attractive faces (Hu et al., 2018) and like femininity seem to depend on the social knowledge about a person (Wang et al., 2020).

Differing results have been reported when the kind of relationship a woman seeks is taken into account. Preferences for masculinity for long-term relationship compared to short-term relationships have been reported (Little et al., 2001) but the preference has been reported for short-term over long-term relationships has been reported too (Little et al., 2002). Additionally, preferences for less feminine faces during the high-conception-risk phase of the female menstrual cycle have been demonstrated (Penton-Voak et al., 1999). In a cross-cultural study, Marcinkowska et al. (2019) found that women living in countries with a higher Human Development Index also prefer masculine men to a higher degree. These findings are contrary to the idea that sexual dimorphism effects might be perceived as attractive because it is perceived as an indicator of a strong immune system since the developed countries tend have healthy populations in general, making health less of a concern for mate choice.

These findings make it difficult to understand how sexual dimorphism is associated with perceived facial attractiveness and illustrates the need to further differentiate findings between different contexts to which they apply. Nevertheless, a meta-analysis by Rhodes (2006) found that based on 18 included articles, a large connection existed between femininity and attractiveness (0.64 ± 0.39, M ± SD), while the effect for masculinity in comparison was smaller (0.36 ± 0.27, M ± SD). The amount of conflicting findings however, paints a picture of sexual dimorphism that emphasis how important it is to understand the variance in effect sizes that might be based in groups of faces or even individual faces.

Health

Perceived health is included in this study as a direct measure of perceived fitness. Based on the biological reasoning it should be one the best predictors for facial attractiveness since it should be a direct signal of mate quality. The biggest issue including health is there is no clearly defined metric to measure the perceived health of a person (Little et al., 2011).

Health has instead often been discussed in terms of skin health or healthy skin colour, since they function as direct visible cues for the health of a person, even cross-culturally (Stephen et al., 2012). Apparent health of skin has been shown to be linked with facial attractiveness in men (Jones et al., 2004) as well as women, especially when skin colour was homogenous (Fink et al., 2006). More recent findings, however, have shown that health might be better understood by the averageness of a face rather than skin colour (Jones, 2017). Symmetry has also been found to have a larger influence on facial attractiveness than skin tone (Vera Cruz, 2018).

Besides skin colour and skin health, face and body shape have also been used to investigate how people make health judgements. Tinlin et al. (2013) demonstrated that the waist-to-hip ratio and Body Mass Index convey information about women's health and Coetzee et al. (2011) showed that the most attractive faces are also considered the healthiest across different levels of adiposity, thereby underlining the association between health and facial attractiveness.

Despite the unclear nature of how people perceive other people's health, support for the positive association has been shown for a positive effect of health judgements on perceived facial attractiveness. Additionally, when asked to rate the perceived health of a face, participants showed high levels of agreement (Jones et al., 2001, 2005), even if it is unclear what these judgements are based on. It has also been found that participants are more willing to show trust in decision games towards partners that they rated as healthier (Krupp et al., 2011). Since there is no clear consensus how to quantify health, the trait was included to understand how much the effect size of subjectively rated health varies across faces.

Likeability

Besides the biologically driven traits, there are also some primarily social concepts and variables that are discussed in the literature, putting the emphasis on social structure rather than biological traits. Following the "what is beautiful is good" stereotype, likeability or other emotionally favourable facial traits might be highly related to facial attractiveness.

In line with this reasoning, smiling has been found to enhance male attractiveness in the context of long-term relationships cross-culturally (Okubo et al., 2015). But people do not only have to depend on their own judgments. Investigating the social influence of face perception, Jones et al. (2007) showed that participants heavily rely on social cues consisting of either smiling or neutral faces and responded more favourably to target faces when the cue faces were smiling. This suggests that the impression participants form about the faces they are confronted with depends at least partly on the social status of the face and that men and women might form these opinions in different ways. Underlining the importance of social perception, Kniffin and Wilson (2004) investigated how sportsmen and women in fraternities and sororities rated each other's talent and effort in their respective sports and their physical attractiveness. They found that non-physical traits showed strong effects on physical attractiveness that these judgments can last for decades.

Most prominently, personality has been reported to be one of the most important factors in partner choice for a long time (Botwin et al., 1997; Buss, 1989; Buss & Barnes, 1986) and attractive people are generally rated to possess greater socially desirable traits compared to unattractive people (Dion et al., 1972). Women also seem to be more unwilling to choose partners with negative personality traits compared to men (Wang et al., 2015).

Zhang et al. (2014) found that experimentally manipulating the personality information that is given to participants about a person influences their evaluation of the persons facial attractiveness. Similar findings have been reported by Lewandowski et al. (2007). Personality traits that are positively associated with facial attractiveness include kindness, honesty and friendliness (Tsukiura & Cabeza, 2011).

Findings are very complex and varied between the different concepts and in order to keep the design simple, personality judgements were condensed into a single rating of likeability. Attention has been shown to be guided by what participants like (Goller et al., 2019) and likeability has also been linked to attractiveness (Little et al., 2006; Rhodes et al., 2001). Not much is known about the overall effect size of likeability, yet the literature implies a positive connection to facial attractiveness and an emphasis on what perceivers personally like. Therefore, at least partial variation in effect sizes might exist between participants and faces.

Trustworthiness

Trustworthiness has been often used as another variable that results from social environments. Kaisler and Leder (2017) used a social cue design in which pairs of people were shown in either frontal or two-thirds view from the side. The people were sometimes set up to look at each other for the gaze cue condition or towards the participants in the neutral condition. They found that the gaze cue condition effected social evaluations and concluded that trustworthiness builds more on other social criteria than structural properties of a face. Trustworthy faces are also seen as more positively valent, while negative valence is due to untrustworthy faces being more threatening (Aguado et al., 2011).

Threat on the other hand has been shown to lead to lower attractiveness (Faust et al., 2018). In another experiment, participants had to look at pairs of people in natural settings, which had previously been framed as either threatening or a normal social situation (Leder et al., 2010). Fixation times of participants were measured using an eye-tracking device. Threat seemed to guide viewing behaviour since attractive faces received longer fixation times in the social situation, whereas in the threat condition the fixation times of attractive male faces dropped.

Similar to health, it is not quite clear which facial features lead to perceived trustworthiness. However, there is consensus that faces that appear smiling or happy are associated with higher trustworthiness (Krumhuber et al., 2007; LaFrance & Hecht, 1995; Oosterhof & Todorov, 2009; Todorov et al., 2013), while angry looking faces are associated with low trustworthiness (Oosterhof & Todorov, 2009; Said et al., 2009; Todorov et al., 2013). Furthermore, trustworthiness perception is based on face shape, for example faces that appear smiling are rated more trustworthy (Lee et al., 2017).

The biggest support of the effects of trustworthiness comes from a study using Principal Component Analysis (PCA). Oosterhof and Todorov (2008) found two dimensions, trustworthiness, and dominance, explain 80% of the variance and can therefore explain social and affective judgments of facial appearance. Trustworthiness was by far the largest component accounting for 63% of total variance. This means that despite the lack of focus on trustworthiness as a predictor of facial attractiveness in hypothesis-driven research, there is strong data-driven evidence that trustworthiness might be one of the most practically relevant traits in order to describe the attractiveness of a face.

Dominance

So far dominance has been discussed as a by-product of masculinity in sexual dimorphism (Grammer & Thornhill, 1994; Perrett et al., 1998). But much like trustworthiness, the biggest support for dominance does not come from theoretical assumptions but rather from studies employing PCA. While trustworthiness explains the biggest portion of variance, dominance commonly is the second most important factor. Depending on the study, the explained variance of dominance ranges from around 8% to 20% of total variance in facial attractiveness (Nakamura et al., 2020; Oosterhof & Todorov, 2008; Todorov et al., 2008). Data-driven models therefore suggest that just like trustworthiness, dominance should be one of the most relevant traits for describing face attractiveness.

Beauty

Neuroscientific findings have revealed that beauty is the potential to please the senses (Chatterjee & Vartanian, 2016). Beauty is the most ambiguous trait included in the study, since there is often very little differentiation between beauty and attractiveness in the literature. Sometimes beauty is even used interchangeably, such as in the case of Schacht et al. (2008) who despite titling their paper “The appraisal of facial beauty is rapid but not mandatory” never actually discuss beauty but rather facial attractiveness in their text. Despite the clear lack of differentiation, beauty was nevertheless included as an exploratory trait, in order to investigate whether participants would show a practical difference in their ratings of attractiveness and beauty.

Methodological Developments

To properly understand the previous findings discussed above, it is important to recognize the circumstances under which these results have been found. For example, most studies aim to specifically test causal hypotheses under laboratory conditions. When statistically significant effects are found, they are usually understood to be universal in the evolutionary framework. These conditions and theoretical assumptions limit the interpretability of previous findings in multiple ways. Laboratory research aims to control the conditions in which stimuli are presented in order to assess a clear effect and allow causal interpretations of the results. While this is the right approach to test causal relationships, and helps to develop psychological theory, it does not capture the relationship in a holistic manner and therefore limits the assessment of practical applicability of a variable. In other words, just because a causal relationship exists between two variables does not mean they influence each other in a large and impactful capacity. Viewing effects as universal construct on the other hand, limits the kinds of research questions asked. This understanding typically leads to approaches in which average ratings are preferred level of observation rather than investigating variance in the data, which leads to a huge amount of ignored or unutilised data and potential research questions.

All of these limitations are nothing new and should not be understood as criticism of individual studies. Rather the research process should be understood as series of articles progressively building on each other to translate initial laboratory findings in an increasingly complex fashion into a more applied context. In facial attractiveness research such a development can clearly be seen in three different methodological developments discussed in the following section. Each of these developments has influenced the design of this study in a unique manner detailed below.

The trend towards more integrative research designs

Early facial attractiveness research has often emphasized theoretically driven hypothesis testing to confirm causal relationships of singular facial traits and facial attractiveness (for example, Langlois & Roggman, 1990). This approach stems from rigorous laboratory research in which all conflicting variables are eliminated to assess the causal relationship between two variables. In a literature review, Rhodes (2006) gave an overview on the effects of averageness, the symmetry and sexual dimorphism of a face and discusses how they explain facial attractiveness ratings. As it included a meta-analysis, one concern of the paper was to assess the effect size for each of the traits individually. Notably, the only discussion of an interaction between variables is how these facial traits relate to health in order to underline the biological mechanisms, but not amongst interactions of the traits themselves. In some instances, the integration of multiple variables was rather driven by accounting for alternative explanations for causal relationships rather than exploring joint explanations (Rhodes et al., 1999).

While the approach of assessing causal relationships between two variables is a reasonable starting point towards understanding an issue, causal relationships might also depend on specific circumstances or different conditions. A good example can be found in the previously discussed findings for the effects of masculinity when the type of relationship a person seeks and whether the person seeks a relationship at all is taken into account (Little et al., 2001, 2002; Mitrovic et al., 2018; Penton-Voak et al., 1999, 2003). The evolutionary driven sexual selection framework also has a long tradition of cultural-comparisons to account for the social and environmental conditions that might shape these effects (Che et al., 2018), which has helped to uncover how societies might influence preference for certain facial traits. The comparison between western and eastern faces as well as perceivers is the most common cultural comparison in facial attractiveness research (Fang et al., 2019, 2020; Koda & Ruttkay, 2009; Nakamura & Watanabe, 2019; Perrett et al., 1998), but studies with a broader scope also exist that even include multiple cultures at the same time (Durkee & Ayers, 2020; Jones, 1995; Valentova et al., 2017). Comparing within- und between-culture agreement in their meta-analysis, Langlois et al. (2000) found high agreement rates for both the cross-cultural (r = .94) and the within-group reliability (r = .90).

Another common moderating factor that is part of the basic assumptions of biological theories is sex difference, either within or between sexes. The interaction between the sex of the perceiver and the perceived face is of particular interest. Leder et al. (2016) found that when viewing faces of the opposite sex, attractiveness captured attention stronger compared to viewing faces of the same sex. Furthermore, participants looked longer at faces that matched their sexual preference (Mitrovic et al., 2016). One intrasex factor that has been extensively researched has been the influence of the female menstrual cycle on facial attractiveness perception (Johnston et al., 2001; Penton-Voak et al., 1999). At peak fertility, women seem to prefer more masculine faces, compared to the rest of the menstrual cycle (Benedict C. Jones et al., 2008; Little & Jones, 2012), as well as more symmetrical ones (Little et al., 2007).

A gender-neutral factors of interest has been self- resemblance. DeBruine (2005) found that faces with a great self-resemblance to the perceiver were rated more trustworthy, albeit equally or less attractive than faces with lower self-resemblance and - not surprisingly - it was found that people also spent more time looking at pictures they personally found attractive, making personal taste a very important determinant of viewing behaviour (Mitrovic et al., 2020).

In summary, contemporary research commonly integrate or control for multiple variables to properly contextualize effects and understand how effects interact with each other to explain facial attractiveness. The presented study will take these developments into account by integrating all the facial traits discussed above in a single multilinear regression model. This design allows to assess the effect sizes of a trait under the consideration of all other traits, therefore creating more realistic estimations of the perception of facial attractiveness due to the more holistic examination.

The trend towards more ecologically valid stimuli and methodological concepts

The inclusion of more ecologically valid stimuli in experiments and the subsequent interpretation of the results using more sophisticated methodological concepts can be seen as a necessary co-development alongside the trend of incorporating multiple variables. Nevertheless, it is important to discuss these developments separately as they do not refer to the kinds of questions asked by researchers but rather how experiments are practically designed, and how findings are understood.

A lot of early research used computer-generated or natural faces that were reduced to the bare minimum face, excluding hairline, jewellery, and every other additional context factor until the bare minimum face was left (for example Little et al., 2002; Little & Jones, 2012; Penton-Voak et al., 2003). On top of that, faces were only presented in a frontal view. The stimuli were created this way because researchers aimed to investigate facial features without the interference of additional conflicting variables, which was seen as the optimal way to investigate causal relationships, whether singular approaches were used or interactive ones. The drawback to these stimuli is, that they appear highly unnatural, which is why more applied designs in modern research have now shifted to incorporate stimuli that are less standardised in terms of for example lighting angle (for example Goller et al., 2018; Jenkins et al., 2011), sometimes even explicitly researching the additional information different angles provide for the perceivers (Kaisler & Leder, 2017). Some studies have even gone so far as to display stimuli where the whole upper body is presented in a natural environment (Leder et al., 2010; Leder, Mitrovic, et al., 2016; Mitrovic et al., 2016, 2018).

Another ongoing trend to create even more realistic stimuli is to make use of audiovisual integration, whereby audio stimuli are presented alongside visual ones, sometimes in form of full videos (Rezlescu et al., 2015; Schall et al., 2009; Valentova et al., 2017; Wells et al., 2013). These types of stimuli extend the representation beyond facial features and towards a more holistic representation of the depicted person. For example, research questions such as the influence of voice pitch or volume can be investigated by digitally manipulating voices or changing different voice lines. This expands theoretical assumptions as well, since there is no reason why mate choice conceptions should be limited to facial features alone and might be understood in terms of full body perception instead.

On the conceptual side, the role of the perceiver has become so widely recognized that methodological conceptualizations have been developed to investigate the role of personal taste in facial attractiveness perception. Hönekopp (2006) has suggested that facial perception should be understood in terms of shared and private taste. Shared taste refers to the agreement between two or more raters of facial attractiveness and can be understood as the difference between the mean rating of an idividual face and the overall mean rating of all faces. Private taste on the other hand refers to how much judges disagree with each other and can be understood in a similar way to shared taste only that the difference between the mean rating of an individual rater and the overall mean rating for all raters is compared. This conceptual framework gives a clear path towards understanding whether facial attractiveness might be driven by general or personal factors.

This has raised the question which categories of objects impacted the most by private taste (Leder et al., 2016). The results indicated that around 40% of the variance of facial attractiveness could be explained by private taste, while around 75% of the variance in abstract artworks could be explained by private taste. This suggests that faces are a rather stable aesthetic object class since the shared taste component is rather large, but still leave considerably large room for personal taste.

A concept very similar conception to Hönekopp's (2006) private and shared taste has emerged in the form of the integration of top-down and bottom-up effects (Stolier, Hehman, & Freeman, 2018; Stolier, Hehman, Keller, et al., 2018). In this concept, an effect is split into the parts that are generated by the face (i.e., the bottom-up part) and the perceiver (i.e., the top-down part). Using this approach one can decipher whether the effects are mainly driven by characteristics inherent to the faces themselves or to the preferences of perceivers, by calculating how large the proportion of either the bottom-up or top-down effects are in the overall effect. This approach allows for an effect to be understood in a holistic fashion, where both top-down and bottom-up parts come together to create the overall impression.

In short, the use of stimuli and the underlying theoretical conceptualisations have gradually moved towards more ecologically valid research designs and conceptualisations. While more integrative research designs discussed earlier call for more complex stimuli, it is still important to recognize that the trend towards ecological validity does not necessarily rely on such designs, but rather attempts to adequately incorporate paradigms into laboratory research that are modelled to consider a more holistic perspective. This study follows these developments by displaying a combination of picture and video stimuli that make use of audiovisual integration.

The methodological concept of this study has been inspired by the conceptual developments that strive for a deeper understanding of the mechanisms of the effects of facial traits that draws from more individualistic explanations rather than general or universal ones. While the exact concepts discussed above will not be used, the methodological development is recognized and each face will be understood as unique and might therefore not only vary in its perceived attractiveness from other faces, but also in the underlying reasons for why it is perceived as attractive might vary as well. For example, one face might be attractive because it might be highly average, while another face might be unaffected by averageness. This would be reflected in varying effect sizes between different faces.

To investigate these potential variances, the multilinear regression will be modelled in a hierarchical fashion. This kind of model allows for the estimation of an overall fixed effect of a facial trait with the addition of a random effect that captures the variance in effect sizes between the individual faces. The advantages of such a model are that the overall fixed effect can be interpreted in the same way as a flat model, as both effects are seen as the generalized output that are a good estimation to predict values of previously unknown faces, while the hierarchical model also offers additional context how this effect is shaped by its individual components. However, the fixed effect in a hierarchical model differs from the one in a flat model insofar as it can roughly be understood as the mean effect size of the effect sizes from the individual faces. This means that it is possible for both models might reach vastly different conclusions, as is the case when there is an overall positive association across the whole data set, while each individual face shows a negative association between two variables. In this case the flat model would conclude a positive effect size, while the hierarchical model would conclude an overall negative effect because it summarizes the associations of each individual face. Additionally, the random effect of the hierarchical model serves as a good indicator for the universality of an effect, since it summarized the variance across the per face effect sizes. Facial traits that show a small random effect across faces can be seen as showing stable and universal effects compared to those with larger random effects because these facial traits are mainly driven by the overall effect and all faces will show similar effects. If the random effects are large however, it is not possible to accurately predict how large the effect for any individual face is with estimating it empirically. The hierarchical model therefore helps to understand the in-depth mechanisms of an individual face’s attractiveness in level of detail that could not be reached in the conventional use of effect size estimation.

The trend towards more data-driven and Bayesian inferential techniques

Besides the conventional theoretically driven hypothesis testing, new developments are emerging that make use of data data-driven techniques. None of these techniques are particularly new, PCA and Bayesian inference have existed for decades, but the recent developments in computer technology have made these techniques more accessible and widely used. What sets them apart from the conventional approach is that they are not necessarily concerned with hypothesis testing or causal interpretations, but rather attempt to describe their depended variable in a mathematical fashion or employ a more inductive reasoning.

Early examples of the use of PCA include the findings for trustworthiness and dominance which have been discussed earlier (Nakamura et al., 2020; Oosterhof & Todorov, 2008; Todorov et al., 2008). These studies generally employ PCA to evaluate the performance of a variable in terms of explained variance in order to separate the well performing ones from the rest. Here, the causal relationship is of secondary nature, while practical performance is of primary concern. Often times these approaches also validate their findings by using the rating data for the faces to train an algorithm to mimic the facial structures that are correlated with the facial traits. The algorithm then creates more digital faces that resemble different expressions of these traits (most often low, middle and high) which are then rated by a different set of participants to confirm that the traits correlate with facial attractiveness (Nakamura et al., 2020; Nakamura & Watanabe, 2019, 2020). This technique turns the conventional scientific process somewhat upside-down. First, the practical performance of a trait is established, then stimuli are created based on the trait, and then the stimuli are used in an experiment to show that the trait is associated with higher facial attractiveness, which proves a causal relationship. There is immense opportunity in these models, mainly because they allow for custom made stimuli, whereby the same facial stimulus is manipulated alongside one trait dimension while the others are held constant. It has even been shown that data-driven models can rival or even outperform conventional theory-driven models (Holzleitner et al., 2019).

Bayesian inference on the other hand is an inferential paradigm that can be used for hypothesis testing. The conventional Frequentist approach depends primarily on significance testing, whereby the likelihood to find any given set of data based on the prior assumption (usually the null hypothesis) is calculated. This leads to a binary interpretation of the results in which the main concern is to establish whether an effect exists or not. In contrast, the Bayesian approach estimates the likelihood that any given effect size is the true effect size in the general population based on the data. The output is summarized in a probability density function (PDF) that describes the estimation. The Bayesian equivalent to significance testing would be to assess whether the estimated PDF of the effect size overlaps with a null-effect or the region of the smallest effect size of theoretical or practical interest. Even though this principle makes it possible to test hypotheses in a binary fashion, the emphasis of effect size estimation forces researchers to look beyond whether effects exist and take effect size and in turn theoretical or practical meaningfulness of an effect into account when interpreting and discussion results.

Data-driven models and Bayesian inference are a shift towards estimating practical performance rather than confirming theoretical explanations. Since the main goal of this study lies in the assessment of methodological concepts and practical meaningfulness of effects in facial attractiveness research instead of testing theoretical concepts, Bayesian inference will be used as it is seen as the superior inferential technique to assess effect sizes.

Research Question and goals of this study

Overall, this study aims to broaden the current methodological understanding of effects in facial attractiveness research. Using a Bayesian hierarchical multilinear model to analyse the data does this by shifting the understanding of facial attractiveness away from general explanations that view facial traits as universal causes of facial attractiveness and towards individual reasons why a face might appear attractive. This is due to the interaction between these two features of the model, whereby the hierarchical structure makes it possible to estimates whether the effect size of a facial trait varies across faces, while the multilinear regression includes multiple effects at the same time. The outcome of the model might for example show that one face has large effect for averageness but no effect for sexual dimorphism while another face shows the opposite pattern, thereby creating an individual explanation for each face’s attractiveness. Since the focus of this model lies primarily in the estimation of effect sizes and their variances, the model also helps to assess of the practical meaningfulness of the included facial traits.

The study design for data-collection was kept simple to keep the focus on demonstrating the features of the model. Facial stimuli were presented on a computer screen and all included variables will be measured using self-reports after each presentation to allow for straightforward interpretation. Instead, more effort was spent on collecting data from as many participants as possible to make model estimation more precise and reliable. To properly exemplify the strength of the hierarchical multilinear model, the data was analysed using multiple different models to compare their outputs.

The comparison between the different models will answer two key questions: How do hierarchical multilinear models help researchers conceptualize the effects of facial traits in an in-depth individualistic manner and how can practical meaningfulness of these effects be understood based on these types of models? To answer the first question the hierarchical multilinear model will be compared to a conventional multilinear model to assess differences in their overall conclusions. The main comparison will be between the overall fixed effect size estimation of the hierarchical model and the estimates of the conventional multilinear model. To answer the second question the practical meaningfulness of each facial trait will be judged based on three criteria: The overall effect size and the effect size variation from the hierarchical model as well as the amount of effect size compensation due to the other facial traits included in the model. These criteria were chosen for different reasons to judge practical meaningfulness based on multiple perspectives.

Most importantly, the larger an effect is, the more influential a trait is for the attractiveness of a face. The overall fixed effect sizes from the hierarchical model will therefore be taken into account as the most direct assessment of meaningfulness. It can be difficult to interpret effect sizes directly as they depend on multiple factors, such as standardisation or, in case of the multilinear model, the number of included predictors. In case of this study the relative effect size of one facial trait compared to the others will be of main interest.

The per face effect size variances that make up the random effects of the hierarchical model will also be taken into account. As previously discussed, the random effects component of the hierarchical model can be seen as an indication for universality of a trait. Universality is very desirable for effects, since these effects generalise reliably across faces. Therefore, smaller random effect sizes should be seen as an indication for a more practically meaningful effect.

The amount effect size compensation was modelled as the difference in fixed effect size between the multilinear model compared to the effects when traits are investigated separately. The larger this difference is, the more the effect of a facial trait is influenced by the consideration of other factors in the model. This also means that the trait is less relevant because it is more easily compensated by others.

In summary, to demonstrate how hierarchical models make it possible to understand the mechanisms of facial attractiveness in an individualistic fashion and how this understanding can guide the assessment of practical meaningfulness of effects a series of three models (hierarchical-multilinear; flat-multilinear; hierarchical-linear) will be employed to a data-set based on simple self-reports.

Materials and Methods

Materials

A stimulus set of 40 unique faces, 20 of which were female, was created to be used in the presented study as well as to be published as the Vienna Talking Faces database. The stimulus set consists of pictures as well as various videos of all the unique faces. The video stimuli consist of various different German sentences spoken by the actors, ranging from neutral greetings ("Hello it's me"; "Hallo ich bin's") over engaging calls to action ("Would you like to come drink a cup of coffee with me?"; "Willst du mit mir einen Kaffee trinken gehen?") to nonsense words ("Bido"; "Gali") and vowels ("a"; "e"; "i"; "o"; "u"). All stimuli in the database are filmed using three different viewing angles (frontal, three-quarters-view and profile view) and actors performed the different sentences in five different emotional conditions (neutral, happy, flirty, angry and sad). For the purposes of this study only the frontal angle in the neutral emotion condition was used to present picture and video stimuli. In line with the neutral approach, the neutral greeting "Hello it's me" ("Hallo ich bin's") was chosen for the video condition. All stimuli were presented in a resolution of 1000x1000 pixels and examples are provided in Figure 1.

Figure 1.

Example stimuli from the practice trials in the frontal angle with a neutral facial expression

Procedure

Since the restrictions during the Covid-19 pandemic forbid all in-person data collection, the study was carried out online via the Labvanced platform (Finger et al., 2017). Since video stimuli were presented alongside static pictures Labvanced was chosen because it offers solutions to the problem that unstable connections could potentially disrupt the data transfer. The platform does this by collecting data not via a live internet connection, but instead storing the data of the study on the local computer of the participant within the browser and retrieves the collected data as a whole once the participant has finished the study. Since technical problems with the video stimuli were found in the Firefox browser, where videos would not properly start playing, the study was limited to the run in Google Chrome. If participants started the study using the wrong browser, they received a notification telling them to reopen the study in Google Chrome instead. This procedure allowed for quick and efficient data collection, since participants could participate in the study at any time and any place but the drawback is that participants viewed the study on screens potentially varied in size or brightness and other non-standardised environmental conditions. Data was collected from January to July 2021.

When starting the study, participants were randomly assigned to the picture or video condition. To begin the study, participants had to provide informed consent by checking a box on the first page of the study. The data collection started out with a demographic questionnaire that included (german) language skills, age, country of living, gender and sexual orientation. Participants then completed a series of three practice trials to get used to the way stimuli were presented and rating scales were displayed, before starting with the rating task. An experimental trial consisted of a fixation cross presented for 5 second, followed by the presentation of the stimulus and a slide with the rating scales in the end. Stimuli in the picture condition were presented for a duration of 5 seconds, while the duration of the video stimuli varied slightly around three seconds, dependent on how fast the actors spoke the sentence. Stimuli were presented in randomized order. Participants were instructed to rate facial traits on a seven-point Lickert-scale with the following scales: Attractiveness: “How attractive do you find this person?” (“Wie attraktiv finden Sie die Person?”); Averageness: “How memorable do you find this person compared to other people?” (“Wie einprägsam finden Sie die Person im Vergleich zu anderen?”); Sexual Dimorphism: “How typically female or male do you find this person?” (“Wie typisch weiblich bzw. männlich ist die Person?”); Health: “How healthy does this person appear to you?” (“Wie gesund erscheint Ihnen diese Person?”); Likeability: “How likeable do you find this person?” (“Wie sympathisch finden Sie die

Person?”); Trustworthiness: “How trustworthy do you find this person?” (“Wie vertrauenswürdig finden Sie die Person?”); Dominance: “How dominant do you find this person?” (“Wie dominant erscheint Ihnen die Person?”); Beauty: “How beautiful do you find this person?” (“Wie schön finden Sie die Person?”). The extreme ends were labelled with (1) “*not at all*” (“*gar nicht*”) and (7) “*extremely*” (“*äußerst*”) as anchors in all cases. All scales were presented in randomized order to keep participants concentrated. It is noteworthy that due to the phrasing of the averageness measurement high values correspond to high memorability, which in the German language is close to distinctiveness of the face while the low values correspond to high averageness. The values for averageness were therefore recoded before the data-analysis so all values were poled in the same way to make the results more easily interpretable.

Participants

Unlike the Frequentist approach, in which a researcher can perform a power calculation, Bayesian methodology does not offer a clear path to perform a sample size estimation. In order to attempt to estimate at least a minimum effect size, the model was run with simulated data to understand how accurate the model would predict the correct values based on various sample sizes. Six different simulations with 50, 100, 200, 500, 1000 and 2000 simulated participants were run. Data was simulated for each of the included facial traits as well as an additional variable to add noise to the data. The facial attractiveness ratings were then computed based on the simulated variables using randomly generated effect sizes. The simulated effect sizes were stored and later compared with the estimations of the model. Since Bayesian inference estimates an effect as PDF, accuracy was not simply judged by the mean of the distribution but by the standard deviation as well. This can be understood like an effect and its confidence interval in the frequentist approach. The smaller the standard deviation in the Bayesian estimation the more precise the estimation is. According to the simulation attempts the mean of the effect size probability distribution matched the simulated effect sizes with the same accuracy for all sample sizes but the 50 participants simulation. The precision of the estimation did reach diminishing returns at 500 participants and the standard deviations did not get much smaller beyond that. Originally, it was planned to attempt to collect that many participants, however running the model after the first 50 participants were collected revealed that the model already showed the same standard deviation using the real data compared to 500 simulated participants. Since the model performance already reached a satisfying degree, the decision was made to keep collecting as many participants as possible, instead of sticking to a set minimum number.

206 participants were collected in total (128 female; mean age *M* = 24.50 years, SD = 6.28) were recruited through recruitment channels of the University of Vienna. 175 participants were recruited via the Laboratory Administration for Behavioral Science system, in which students receive course credit as compensation for participating in studies and 31 participants were recruited via the Vienna Cognitive Science Hub Study Participant platform in which volunteers could enrol to receive monetary reward of 5€ as compensation for their participation. The assignments for the two conditions were randomized and automatic, so there was no way to ensure that both conditions would end up with equal amounts of participants due to participants not finishing the study and quitting while the data collection was still active. Ultimately, the two conditions differed slightly in numbers of participants with 108 participants collected (65 female; mean age *M* = 23.96 years, SD = 4.84) for the picture and 98 participants collected (63 female; mean age *M* = 25.10 years, SD = 7.54) for the video condition.

Statistical Analysis

As previously discussed, multiple models were run to analyse the data and compare the conclusions of the models. The decision to model the random effects component of the hierarchical models as a function of faces rather than a function of perceivers was made for two reasons. First, conceptually facial attractiveness effects are usually discussed based on facial traits in the literature and are therefore implicitly located in the face itself rather than the perceivers and secondly because of the uneven amount of data between faces and participants. The amount of data that could be collected for each participant was limited to the 40 faces included in the stimulus set, while the amount of data that was collected per face included 206 participants. As discussed in the sample size simulation discussion, this imbalance results in drastically more accurate estimations and a better performing model if the random effect is set up in a per face manner.

Data was analysed in R Version 4.1.0 (R Core Team, 2021) using Markov-Chain-Monte-Carlo (MCMC) simulations for the Bayesian inference via the rstan package Version 2.21.2 (Stan Development Team, 2020). The priors of all models were given normal distribution with $\mu = 0$ and $\sigma = 1$, which is a very weak and uninformed prior in order to influence the results as little as possible. This was done despite the effect size estimations from a previous meta-analysis for averageness and sexual dimorphism (Rhodes, 2006), since the findings might be biased because they are calculated based on laboratory findings on the one hand and without taking the presence of other variables into account on the other hand. Giving informed priors to the two variables for which previous findings exist would also skew the model in their favour if every other prior is set to be as uninformed as possible. A total of 4 MCMC simulations were run for each model to confirm that chains converged and the results would be reliable. Each chain had a length of 4000 iterations with warm-up being the first 2000 iterations.

All analyses were based on the full data-set of the combined data from both the picture and the video conditions. This was done to maximise the amount of data which increases the model accuracy. To make sure that the effect size estimation is not overly influenced by differences by the two conditions, the hierarchical multilinear model was run for both conditions separately. If both subsets show similar effect sizes and patterns, the two subsets can be used in combination without overly compromising the results.

Results

In order to get a better understanding of the data, first the descriptive indicators were taken into consideration. The descriptive statistics of facial attractiveness can be found in Table 1, while the tables for the remaining eight variables can be found in the Appendix: Table Directory. Large variances were found in each face's attractiveness ratings, with most faces showing a standard deviation of around SD = 1.50. Considering that data was collected using a seven-point Lickert-scale, these standard deviations mean that participants used the full range of the scale when rating attractiveness, which is also reflected in each face's minimum and maximum ratings. The findings for the other included variables show similar patterns. These findings highlight the importance of individual approaches to better understand the mechanisms why the same face appears attractive to some participants but not to others.

Table 1.
Attractiveness ratings for every face in the stimulus set

	Descriptive Metric				
Face	Mean	SD	Min	Med	Max
1	3.92	1.59	1.00	4.00	7.00
2	4.04	1.47	1.00	5.00	7.00
3	4.52	1.56	1.00	5.00	7.00
4	4.11	1.72	1.00	4.00	7.00
5	3.92	1.63	1.00	4.00	7.00
6	3.15	1.39	1.00	3.00	7.00
7	3.72	1.55	1.00	4.00	7.00
8	3.03	1.37	1.00	3.00	7.00

9	3.19	1.30	1.00	3.00	7.00
10	3.94	1.54	1.00	4.00	7.00
11	4.02	1.50	1.00	4.00	7.00
12	3.44	1.45	1.00	3.00	7.00
13	3.01	1.40	1.00	3.00	6.00
14	3.24	1.51	1.00	3.00	6.00
15	3.61	1.51	1.00	4.00	6.00
16	4.41	1.58	1.00	5.00	7.00
17	2.72	1.43	1.00	3.00	7.00
18	4.35	1.61	1.00	5.00	7.00
19	2.53	1.42	1.00	2.00	7.00
20	4.12	1.52	1.00	4.00	7.00
21	3.77	1.55	1.00	4.00	7.00
22	3.94	1.49	1.00	4.00	7.00
23	2.27	1.18	1.00	2.00	5.00
24	2.54	1.34	1.00	2.00	6.00
25	4.65	1.38	1.00	5.00	7.00
26	4.70	1.43	1.00	5.00	7.00
27	2.34	1.29	1.00	2.00	6.00
28	2.39	1.20	1.00	2.00	7.00
29	4.94	1.51	1.00	5.00	7.00
30	2.68	1.32	1.00	3.00	6.00
31	3.27	1.59	1.00	3.00	7.00
32	1.60	0.88	1.00	1.00	6.00
33	3.37	1.68	1.00	3.00	7.00
34	2.75	1.55	1.00	2.00	7.00
35	3.88	1.61	1.00	4.00	7.00
36	3.79	1.55	1.00	4.00	7.00
37	3.14	1.50	1.00	3.00	7.00
38	4.58	1.52	1.00	5.00	7.00
39	3.01	1.44	1.00	3.00	7.00
40	2.01	1.11	1.00	2.00	7.00

Note. Ratings are a combination of both picture and video condition, N = 206.

Model 1

Model 1 was designed to estimate the effect sizes of the facial traits discussed above hierarchically with random effects as a function of face. At first the model was assessed using the full data-set. The results, summarized in Table 2, show almost no effects for any facial trait with the exception of beauty and likeability. Closer inspection revealed a large correlation between facial attractiveness and beauty (r = .81) which might have caused the poor performance in the model. The decision was made to exclude beauty from the model for this very reason.

Table 2.

Estimated Effect Sizes of Model 1 including beauty

Variable	Effect Size Metric			
	$\text{Mean}_{\text{Fixed Effect}}$	$\text{SD}_{\text{Fixed Effect}}$	$\text{Mean}_{\text{Random Effect}}$	$\text{SD}_{\text{Random Effect}}$
Intercept	-0.419	0.061	0.049	0.037
Averageness	-0.034	0.008	0.020	0.010
Sexual Dimorphism	0.038	0.009	0.009	0.007
Health	0.037	0.009	0.012	0.009
Likeability	0.115	0.013	0.036	0.012
Trustworthiness	0.024	0.011	0.017	0.011
Dominance	0.042	0.008	0.009	0.007
Beauty	0.674	0.013	0.059	0.011

Note. The fixed effects summarize the overall effect size of a variable, while the random effect component summarizes the effect size variance across faces. *N* = 206

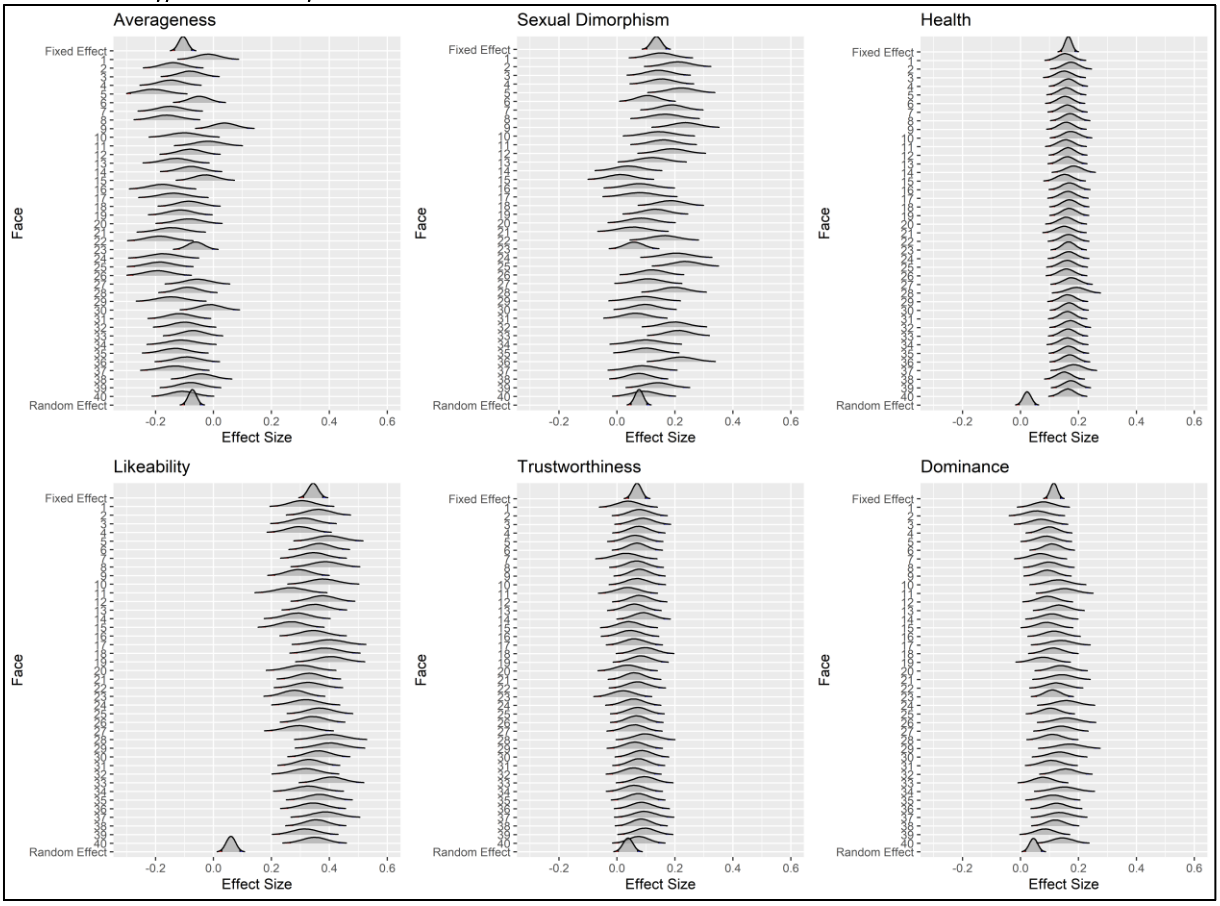
After excluding beauty, the model was run using the same parameters specified above. The results are summarized in Table 3 and the effects are graphically displayed in Figure 2 for a visual comparison that includes each individual faces effect size as well. The model revealed small to medium sized fixed effects for all included variables. Likeability showed by far the largest fixed effect (μ = 0.344; σ = 0.017), while health (μ = 0. 165; σ = 0.012), sexual dimorphism (μ = 0.136; σ = 0.017), dominance (μ = 0.115; σ = 0.012) and averageness (μ = -0.105; σ = 0.015) all showed very similar effect sizes. Notably, averageness was the only variable with a negative effect size. Only trustworthiness showed a small effect size (μ = 0.069; σ = 0.015). On the other hand, only small random effect sizes were found for all variables, the largest being the effect for sexual dimorphism (μ = 0.076; σ = 0.014).

Table 3.
Estimated Effect Sizes of Model 1 excluding beauty

	Effect Size Metric			
Variable	$Mean_{Fixed\ Effect}$	$SD_{Fixed\ Effect}$	$Mean_{Random\ Effect}$	$SD_{Random\ Effect}$
Intercept	-0.689	0.081	0.102	0.074
Averageness	-0.105	0.015	0.073	0.014
Sexual Dimorphism	0.136	0.017	0.076	0.014
Health	0.165	0.012	0.023	0.014
Likeability	0.344	0.017	0.060	0.016
Trustworthiness	0.069	0.015	0.038	0.017
Dominance	0.115	0.012	0.045	0.014

Note. The fixed effects summarize the overall effect size of a variable, while the random effect component summarizes the effect size variance across faces, $N = 206$

Figure 2.
Estimated Effect Sizes of Model 1 visualized

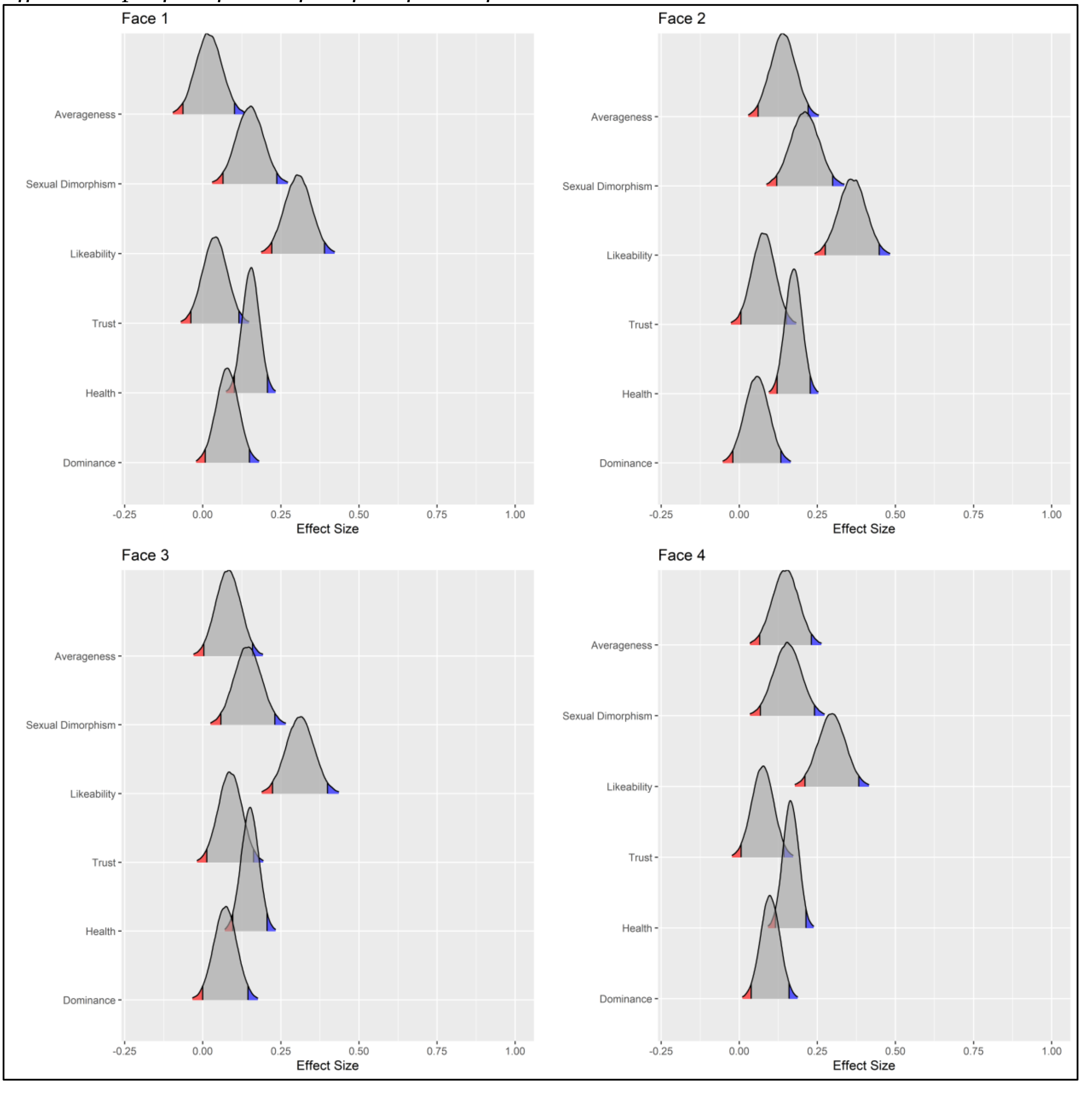


Note. The fixed effect is shown alongside every individual face's effect size estimation. The random effect is included as a summary of the variance between the faces at the bottom.

The results can alternatively be depicted in an individualistic manner in which the results are displayed by face. A graphical example is provided in Figure 3. This kind of depiction makes it possible to easily understand each faces effect size profile for the included variables of the model, and compare faces with each other. In case of this studies results, all faces have very similar profiles as not much effect size variances were found in the data.

Figure 3.

Effect size profiles for the first four faces of the data-set



Next, Model 1 was run separately for both stimulus conditions to investigate differences in effect sizes and patterns. The results are summarized in Table 4 and 5 respectively. Not many differences between the two conditions were found in either the fixed and the random effects as the estimated effect sizes were rather similar. The only exception was dominance as the effect was larger in the picture condition ($\Delta Mean_{\text{Fixed Effect}} = 0.091$).

Overall, the differences were seen as small enough to keep analysing the data using the combined data-set for the purposes of this study.

Table 4.
Estimated Effect Sizes of Model 1 for the picture condition

Variable	Effect Size Metric			
	$\text{Mean}_{\text{Fixed Effect}}$	$\text{SD}_{\text{Fixed Effect}}$	$\text{Mean}_{\text{Random Effect}}$	$\text{SD}_{\text{Random Effect}}$
Intercept	-0.709	0.110	0.091	0.069
Averageness	-0.077	0.016	0.055	0.019
Sexual Dimorphism	0.134	0.023	0.098	0.017
Health	0.158	0.016	0.025	0.017
Likeability	0.350	0.021	0.055	0.021
Trustworthiness	0.064	0.019	0.022	0.015
Dominance	0.157	0.017	0.062	0.021

Note. Based on the sample from the picture condition, $N = 108$

Table 5.
Estimated Effect Sizes of Model 1 for the video condition

Variable	Effect Size Metric			
	$\text{Mean}_{\text{Fixed Effect}}$	$\text{SD}_{\text{Fixed Effect}}$	$\text{Mean}_{\text{Random Effect}}$	$\text{SD}_{\text{Random Effect}}$
Intercept	-0.681	0.116	0.096	0.071
Averageness	-0.127	0.020	0.085	0.017
Sexual Dimorphism	0.130	0.018	0.032	0.018
Health	0.174	0.017	0.020	0.014
Likeability	0.346	0.023	0.068	0.025
Trustworthiness	0.080	0.023	0.064	0.024
Dominance	0.066	0.014	0.024	0.016

Note. Based on the sample from the video condition, $N = 98$

Model 2

Model 2 estimated the effects for all variables included in Model 1 in the conventional fixed effect fashion. This was done for methodological considerations to compare the overall fixed effect estimate approach to the hierarchical one to evaluate differences between the methodological conceptualizations of facial attractiveness effects. The results of the model are summarized in Table 6.

Table 6.
Estimated Effect Sizes of Model 2

Variable	Effect Size Compensation Metric	
	$\text{Mean}_{\text{Fixed Effect}}$	$\text{SD}_{\text{Fixed Effect}}$
Intercept	-1.237	0.080
Averageness	-0.100	0.010
Sexual Dimorphism	0.169	0.012
Health	0.207	0.012
Likeability	0.369	0.015
Trustworthiness	0.072	0.015
Dominance	0.130	0.010

Note. Based on the combined data-set, N = 206

Overall, Model 1 and 2 showed very similar estimates for the fixed effects, with Model 1 showing slightly smaller effect sizes. Nonetheless, both models reach almost identical conclusions. The differences between the models are summarised in Table 7. Additionally, a visual comparison between the two models is provided in Figure 4.

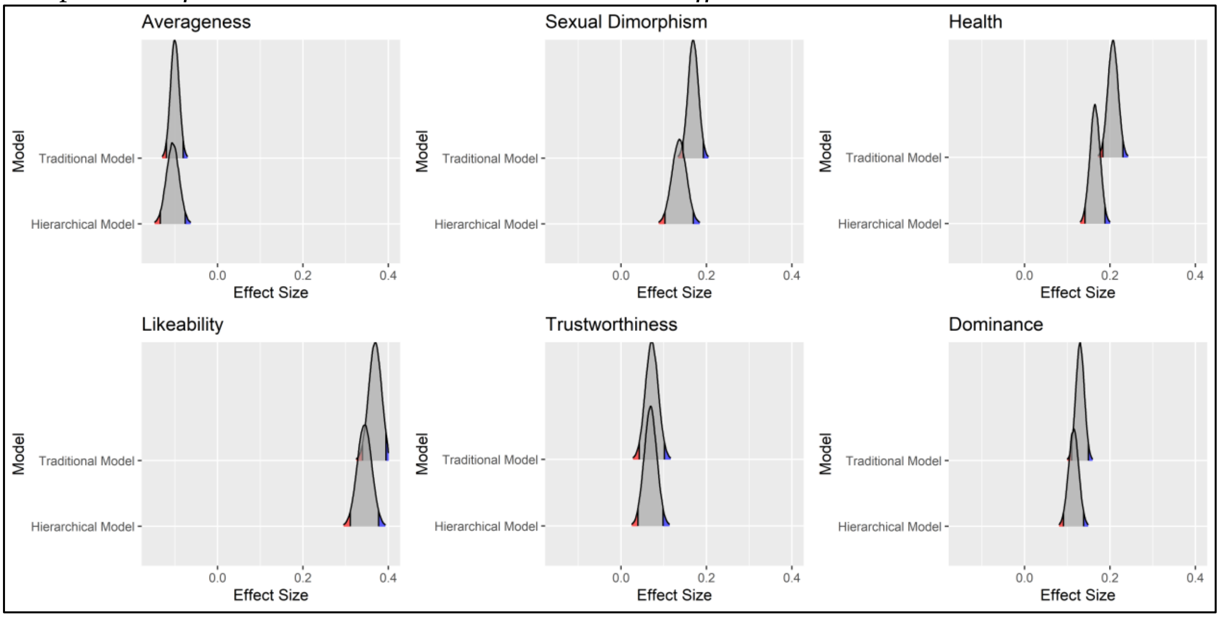
Table 7.
Comparison between the effect sizes estimated in Model 1 and 2

Variable	Effect Size Compensation Metric	
	$\Delta\text{Mean}_{\text{Fixed Effect}}$	$\Delta\text{SD}_{\text{Fixed Effect}}$
Averageness	-0.005	0.005
Sexual Dimorphism	-0.033	0.005
Health	-0.042	0.000
Likeability	-0.025	0.002
Trustworthiness	-0.003	0.000
Dominance	-0.015	0.002

Note. Numbers in the display the difference between the estimations of Model 1 compared to Model 2. Negative numbers indicate that the estimated size of the metric decreased in Model 1 compared to Model 2, N = 206

Figure 4.

Comparison of conventional and hierarchical model effect size estimates



Model 3

What is referred to as Model 3 is a series of six different models in which the variables included in Model 1 were estimated in a singular fashion, separately from each other. However, the hierarchical structure of the model was kept the same. This approach made it possible to investigate how the effect size of any given trait was compensated by the other variables in the multilinear model by comparing the results from Model(s) 3 and Model 1. The results of the model are summarized in Table 8.

Overall, fixed effect sizes as well as random effects were much larger compared to Model 1. This suggests that not just overall effect size is compensated by the multilinear model but also the effect size variance. Likeability still showed the largest effect size (μ = 0.506; σ = 0.020) with trustworthiness (μ = 0.412; σ = 0.019) and health (μ = 0.411; σ = 0.019) showing large effect sizes as well. Only dominance still showed a comparatively small effect size (μ = 0.146; σ = 0.016). Again, averageness shows a negative effect (μ = -0.238; σ = 0.024). The full comparison between Model(s) 1 and 3 estimations is summarized in Table 9.

Comparing effect size compensation, dominance showed almost no decrease between the models in either effect ($\Delta Mean_{Fixed\ Effect}$ = -0.031; $\Delta Mean_{Random\ Effect}$ = -0.032). Meanwhile, trustworthiness showed the largest compensation by other variables ($\Delta Mean_{Fixed\ Effect}$ = -0.343; $\Delta Mean_{Random\ Effect}$ = -0.054). The difference in size of compensation between the variables might partially be explained by the correlation between effect size and size of compensation (r=.465). averageness is the only trait for which effect size has seemingly increased

($\Delta \text{Mean}_{\text{Fixed Effect}}$ = 0.133; $\Delta \text{Mean}_{\text{Random Effect}}$ = 0.062), yet this increase is only due to the negative effect averageness showed in both models.

Table 8.
Estimated Effect Sizes of Model(s) 3

Variable	Effect Size Metric			
	$\text{Mean}_{\text{Fixed Effect}}$	$\text{SD}_{\text{Fixed Effect}}$	$\text{Mean}_{\text{Random Effect}}$	$\text{SD}_{\text{Random Effect}}$
Intercept	2.443	0.086	0.435	0.071
Averageness	-0.238	0.024	0.135	0.019
Intercept	1.690	0.088	0.348	0.083
Sexual Dimorphism	0.345	0.024	0.131	0.018
Intercept	1.489	0.092	0.450	0.073
Health	0.411	0.019	0.091	0.015
Intercept	1.223	0.078	0.382	0.064
Likeability	0.506	0.020	0.111	0.016
Intercept	1.609	0.102	0.561	0.081
Trustworthiness	0.412	0.019	0.092	0.015
Intercept	2.863	0.130	0.785	0.097
Dominance	0.146	0.016	0.077	0.016

Note. Each variable was calculated separately from each other, N = 206

Table 9.
Comparison between the effect sizes estimated in Model(s) 1 and 3

Variable	Effect Size Compensation Metric			
	$\Delta \text{Mean}_{\text{Fixed Effect}}$	$\Delta \text{SD}_{\text{Fixed Effect}}$	$\Delta \text{Mean}_{\text{Random Effect}}$	$\Delta \text{SD}_{\text{Random Effect}}$
Averageness	0.133	0.009	0.062	0.005
Sexual Dimorphism	-0.209	-0.007	-0.055	-0.004
Health	-0.246	-0.007	-0.068	-0.001
Likeability	-0.162	-0.003	-0.051	0.000
Trustworthiness	-0.343	-0.004	-0.054	0.002
Dominance	-0.031	-0.004	-0.032	-0.002

Note. Numbers in the display the difference between the estimations of Model 1 compared to Model 2. Negative numbers indicate that the estimated size of the metric decreased in Model 1 compared to Model(s) 3, N = 206

Discussion

This thesis aimed to demonstrate how hierarchical multilinear models make it possible to conceptualize the effects of facial traits in an individualistic manner and how these models help guide the understanding of the practical meaningfulness of these effects. These goals were set to be achieved by applying multiple different models to a data-set based on a simple self-report study. The comparison between the models will be the main concern of the following discussion.

A surprisingly high amount of variance was found in the ratings, exemplified by the facial attractiveness ratings (Table 1) but mirrored by all included variables (included in the Appendix). Participants made regular use of the full range of the Lickert-scale with most standard deviation being around SD = 1.50. These findings are somewhat contrary to Langlois et al. (2000) who found high cross-cultural and within-group agreement rates and exemplify the importance of understanding how personal taste and different facial traits and their interactions shape attractiveness perception. Perhaps personal taste or top-down effects in the context of the framework of Stolier et al. (2018) play a more important role than previously recognized. It further highlights the importance of good pre-selection of stimuli for experiments, since simply selecting stimuli based on their mean values of attractiveness, likeability, etc. might not lead to reliable results if the variance amongst the variables is high.

On a methodological level these results further illustrate the importance of specifically investigating variances in the data, rather than averages. Since participants seem to disagree with each other about the attractiveness of a face it is important to understand the mechanisms that make individual faces appear more or less attractive to different perceivers. In this case the approach was taken to investigate how the perceived attractiveness of a face changes alongside the perception of the other included variables. The hierarchical multilinear model achieves this by not only estimating an overall effect but also an effect size estimate for each individual face. Modelling data in this way conceptually bases the understanding of the overall effects in each face's effects which allows researchers to understand each individual face's effect size profile as demonstrated in Figure 3.

Comparing the results of Model 1 and 2 shows that both models reach almost identical conclusions which means that the presented results are in line with the conventional scientific approach. These results are promising for re-evaluation and recontextualization of previous findings because it suggests that they hold up under different methodological understanding. Nonetheless, Model 2 lacks the critical context Model 1 provides. Much like a mean score, an overall effect can be a bad indicator for individual value if the variances are high. Had Model

1 shown large random effects then the overall effect would not be a good indication for each individual face's effects and call assumptions of universality of effects into question. Like previously discussed, universality refers to the assumption that more or less every human finds these traits attractive in a very similar manner or that all faces are attractive for the same reasons. However, these assumptions are often made implicitly as a result of either the evolutionary framework or the structure of the statistical model. The advantage of hierarchical models takes have over the conventional flat approach is to make it possible to investigate these assumptions because they take an individualistic approach. In case of this study only small random effects were found, meaning all faces showed roughly the same effect. This suggests that the effects of the included traits can be seen as universal.

The multilinear structure of the model influences the results by taking multiple explanations of facial attractiveness into account at the same time. This is of great importance when evaluating the fixed and random effects of any given facial trait, since both of these components are influences by the presence of additional variables. However, the real advantage stems from comparing estimates from multiple different models to get a deeper understanding of the influence different facial traits have on facial attractiveness. For example, the fixed effect of sexual dimorphism appears quite large in isolation and the random effect is large enough to question the universality of the trait, as it has almost the same size as the fixed effect of dominance (Table 8). When taking the other facial traits into account the effect sizes become smaller, implying that the variance in effect size stems rather from other unaccounted variables than from large variation within the effect sizes of sexual dimorphism itself. This short example demonstrates that the strength of the hierarchical multilinear model lies in the practical understanding of the included variables, rather than testing theoretical causal assumptions.

The following part of the discussion will exemplify how the practical meaningfulness of the included facial traits might be understood using the results from Model 1 and its comparison to Model 3. Like previously discussed, interpreting the estimated effect sizes directly can be difficult which is why the relative effect sizes will be of main interest.

Examining the findings for the overall fixed effect sizes in Model 1 shows large differences between the different facial traits, with large effects being the most practically meaningful. Likeability showed by far the largest overall effect ($\mu = 0.344$; $\sigma = 0.017$), followed by health ($\mu = 0.165$; $\sigma = 0.012$), sexual dimorphism ($\mu = 0.136$; $\sigma = 0.017$), dominance ($\mu = 0.115$; $\sigma = 0.012$), averageness ($\mu = -0.105$; $\sigma = 0.015$) and trustworthiness ($\mu = 0.069$; $\sigma = 0.015$). Conventional would order the practical meaningfulness of the facial traits in this order, before taking the other two comparison criteria into consideration. These findings are contradictory to those from PCA-approaches (Nakamura & Watanabe, 2019; Oosterhof & Todorov, 2009), in which trustworthiness and dominance are by far the two predictors that account for most of the variance in the data, while they only rank 4th and 6th using the presented approach. This might be due to the different natures of the models, since a regression model minimizes prediction error from one variable to another, while PCA minimizes the overall variance of the data. Another surprising finding is the negative association between averageness and facial attractiveness. The findings might be partially influenced by the phrasing of the measurement, since memorability ("Einprägsamkeit") might differ in the public understanding from such a complex, academic construct like averageness that is hardly captured by any adjective without creating any unwanted connotations.

Examining the findings for the random effect sizes in Model 1 shows large similarities for the per face effect size variance between the different facial traits, with small effects being the most practically meaningful. Health showed the smallest random effect ($\mu = 0.023$; $\sigma = 0.014$), followed by trustworthiness ($\mu = 0.038$; $\sigma = 0.017$), dominance ($\mu = 0.045$; $\sigma = 0.014$), likeability ($\mu = 0.060$; $\sigma = 0.016$), averageness ($\mu = 0.073$; $\sigma = 0.014$) and sexual dimorphism ($\mu = 0.076$; $\sigma = 0.014$). One reason as to why sexual dimorphism shows the largest effect size variance might be due to the way it was measured. Usually the concept is investigated separately for femininity and masculinity (see Rhodes, 2006), yet the phrasing used asked participants to rate the faces based on their gender ("How typically female or male do you find this person?"). Therefore, females are rated in terms of their femininity, while males are rated in terms of their masculinity which might lead to differences in effect sizes. As all the random effect sizes are small and do not differ very much, the random effects are not as important as the other criteria included in the assessment of practical meaningfulness.

The comparison of the results between Model 1 and Model 3 reveals that the different traits show different sizes of compensation of the overall fixed effects in the multilinear model. Initially, the effect size estimation for dominance seems almost unaffected. This however might be explained by the correlation between the effect sizes in the singular models and the size of compensation in the multilinear model. Since dominance shows the smallest effect size in the singular model there is almost no potential to compensate the effect size further. The effect size of trustworthiness shows the opposite pattern, as it is estimated to have the second largest singular effect, but it is also the trait that is the most compensated by the multilinear model. The compensation is so large that as previously discussed, trustworthiness ends up with the smallest effect size in the multilinear model. This shows that the researchers need to spend a lot of effort validating the effects found in a controlled experimental setting and test them in an ecologically valid, holistic environment to look beyond simple tests for causality and discuss the practical meaningfulness of the variables within their models. Likeability is a good example of such a variable as it does not only show the largest effect sizes in both the singular and the multilinear model, it is also the third least compensated facial trait included in the model. Overall, the finding that traits were more compensated than others and the compensation does not just bring the effect sizes for traits closer together but even changes the order from largest to smallest effect between the singular models and the hierarchical one. This highlights the importance of holistic models as solely relying on findings from singular research design might not just overestimate effects in general but also gives a false impression of the meaningfulness of a trait.

All three criteria taken together, the presented findings would suggest likeability be to be the most meaningful facial trait to explain facial attractiveness, based on the overall effect size, the small amount of per face effect size variance, and the rather small level of compensation by other variables in the model. Health was the second most relevant predictor. It had the second largest effect size, is the most stable predictor due to having almost no per face variance. The only downside is that health showed the second largest compensation which might become more relevant if more variables were included in the model. Dominance was the third most meaningful trait, despite the effect being marginally smaller than the one of sexual dimorphism, but the effect was very stable across faces and is barely compensated by the presence of other variables. Sexual dimorphism beat averageness narrowly for the third least practically meaningful trait because the fixed effect size is slightly larger with very similar random effect and compensation sizes. Overall, the practical meaningfulness of averageness was surprisingly small for how prevalent this trait has been researched. This leaves trustworthiness as the least important trait, mostly because of the very small overall effect size and the high level of compensation in the multilinear model.

Beauty was not considered in the discussion of practical relevance, since it was excluded from the model due to its high correlation with facial attractiveness which suggests that the two concepts might be similar in practice in a way that participants can not properly tell them apart. The inclusion of beauty in the model had therefore led to poor estimations for the other variables. This also suggests that beauty might not be a good variable to include in future models, not for the lack of association with facial attractiveness, but rather because of the inability of participants to properly differentiate between the two concepts.

Despite not being the main target of comparison, the use of two different kinds of stimuli conditions will briefly be discussed. Comparing the picture and video condition, not much difference between the conditions has been found. These findings suggest that ecologically valid video stimuli that make use of the same underlying mechanisms and coherences between facial traits as picture stimuli do. This justified not only the effect size estimation based on both kinds of stimuli but also generally the growing use of ecologically valid stimuli for studies that are not relying on rigorous laboratory designs to strictly confirm causal explanations. Dominance was the only exception out of all tested traits, with the effect size being more than twice as large in the picture condition. This might be because videos give participants additional cues like movements and voice to judge dominance. Therefore, the judgment of dominance might be based on social cognition rather than facial structure, which might change the dynamics of the connection to facial attractiveness. The random effects component also stayed largely the same between the conditions.

Limitations

One limitation of the presented study is directly tied to one of its biggest strengths, namely the use of self-reports and ecologically valid stimuli. This means that all presented findings are not only based on the personal evaluation of the independent variables, but also that these variables are based on the natural variances across faces, instead of experimental manipulations that specifically pre-select stimuli or manipulates them to induce variance. This prevents the results from being interpreted in a causal manner.

Similarly, the sample population as well as most depicted stimuli are very homogenous groups focussing mostly on central European perceivers and faces. This limits the generalizability of the findings towards other cultures. These concerns as well as the previous limitation highlight that the results of this study should be seen as a demonstration how hierarchical multilinear models can broaden methodological perspectives in the field of facial attractiveness research.

Another limitation lies in the structure of the model and the data. The random effect was modelled as a function of face to reflect that effects are commonly understood as a result of facial traits. Like previously discussed, this design was also chosen because the amount of participant data per face was much larger than the amount of face data was per person. The accuracy of modelling per participant variation would therefore be very compromised. Therefore, the structure of the model does not capture the personal preferences of perceivers that have been shown to have a large effect on facial attractiveness perception (Leder, Goller, et al., 2016). This might explain why not a lot of variance is found within the effect sizes of facial traits and it further highlights the importance of understanding both sides of the perceptual process, much like Stolier et al. (2018) conceptualized.

The way data was acquired also compromises the comparability between participants somewhat. Moving the data collection to an online platform limited the amount of control about the environments in which stimuli were displayed. On the technical side, this might include different screen sizes, resolutions, brightness or colour settings. This might influence how stimuli are displayed and in turn perceived. On the environmental side, this might include the lighting of the room, noise distractions, time of the day or other factors that compromise attention. Overall, the data was not collected in a standardized environment.

Further, despite taking the seven most important variables into account before excluding beauty for methodological concerns, this should by no means be understood as a comprehensive investigation of all variables in the field.

Further Studies

Further studies might pay closer attention to the previously discussed personal taste or top-down perceiver effects. This might mean expanding on the number of facial stimuli in an experiment and attempting to maximise variance across all facial dimensions. Such an attempt might make use of data-driven models to create such a stimulus set and make sure that each included facial trait is controlled for when setting up the stimuli. Using a sufficiently large stimulus set, the same model presented in the current study could be used to model random effects as a function of participant. Making use of both a large stimulus set as well as a large sample size both approaches could be combined to fully realize the model of Stolier et al. (2018).

Another consideration might be the use of culture as an additional layer in the hierarchical model. This might help to understand cultural differences better if, since effect sizes and therefore the underlying structure of facial attractiveness perception might vary across cultures more prominently than they vary along the population of a culture. Further studies might therefore collect data and facial stimuli that vary between cultures. The same might be true for the sex of a face or the sex of the perceiver. It seems less likely however, because the presented findings do not show a lot of variance across faces as they would if there were sex differences.

The findings presented here only took the effect size of traits into account while the other traits were kept stable. One final consideration might be the additional inclusion of interaction effects. This approach has to be taken carefully since including as many traits and interaction terms might lead to overfitting the model and compensate effect sizes so much each individual terms effect size ends up close to zero. Such a model might be useful to describe facial attractiveness, yet it would be hard to interpret the findings other than in terms of mathematical accuracy and include them into psychological theory.

Conclusion

In conclusion, the presented findings exemplify how hierarchical multilinear models can broaden methodological perspectives by understanding overall effects as a summary of individual face's effects. Despite reaching the same conclusions, the hierarchical model showed clear conceptual advantages over the conventional flat model by resembling modern conceptualizations of effects in facial attractiveness research more closely. A parallel can be drawn between the structure of the presented hierarchical model and Hönekopps (2006) concept of shared and private taste or the perspective of bottom-up and top-down effects (Stolier, Hehman, & Freeman, 2018; Stolier, Hehman, Keller, et al., 2018). While the current model understands effects in a bottom-up fashion, building models that take a top-down perspective or combine bottom-up and top-down components might be the next step to merge these emerging methodological concepts.

Furthermore, the assessment of practical meaningfulness these kinds of models provide help guide future research in two major ways. First by shedding light on the kinds of variables that might be important to include in future studies such as likeability and health. Since these variables are seen to be most important for practical settings, understanding what kinds of facial shapes or behaviours influence the judgements of these variables seems to be a promising approach in understanding facial attractiveness. Secondly, the effect size estimations can be used as a basis for future sample size estimations. Especially the effect size comparisons between Model 1 and 3 will help provide a perspective on how much researchers have to adjust their expectations when including multiple variables in their design.

Funding

This study has been funded by the Vienna Science and Technology Fund (Wiener Wissenschafts-, Forschungs- und Technologiefonds).

Declaration of Conflicting Interests

I hereby declare no potential conflict of interests regarding the research or funding of this work.

References

Aguado, L., Román, F. J., Fernández-Cahill, M., Diéguez-Risco, T., & Romero-Ferreiro, V. (2011). Learning about Faces: Effects of Trustworthiness on Affective Evaluation. *The Spanish Journal of Psychology*, *14*(2), 523–534. https://doi.org/10.5209/rev_sjop.2011.v14.n2.1

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

Boas, F. (1955). *Primitive art*. New York, NY: Dover.

Botwin, M. D., Buss, D. M., & Shackelford, T. K. (1997). Personality and mate preferences: Five factors in mate selection and marital satisfaction. *Journal of Personality*, *65*, 107–136. https://doi.org/10.1111/j.1467-6494.1997.tb00531.x

Brand, R. J., Bonatsos, A., D'Orazio, R., & Deshong, H. (2012). What is beautiful is good, even online: Correlations between photo attractiveness and text attractiveness in men's online dating profiles. *Computers in Human Behavior*, *28*, 166–170. https://doi.org/10.1016/j.chb.2011.08.023

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences*, *12*(1), 1–14.

Buss, D. M., & Barnes, M. (1986). Preferences in human mate selection. *Journal of Personality and Social Psychology*, *50*(3), 559–570. https://doi.org/10.1037/0022-3514.50.3.559

Cárdenas, R. A., & Harris, L. J. (2006). Symmetrical decorations enhance the attractiveness of faces and abstract designs. *Evolution and Human Behavior*, *27*, 1–18. https://doi.org/10.1016/j.evolhumbehav.2005.05.002

Catz, O., & Lewis, M. B. (2020). Exploring distinctiveness, attractiveness and sexual dimorphism in actualized face-spaces. *Visual Cognition*, *28*(9), 453–469. https://doi.org/10.1080/13506285.2020.1797967

Chatterjee, A., Thomas, A., Smith, S. E., & Aguirre, G. K. (2009). The neural response to facial attractiveness. *Neuropsychology*, *23*(2), 135–143. https://doi.org/10.1037/a0014430

Chatterjee, A., & Vartanian, O. (2016). Neuroscience of aesthetics. *Annals of the New York Academy of Sciences*, *1369*(1), 172–194. https://doi.org/10.1111/nyas.13035

Che, J., Sun, X., Gallardo, V., & Nadal, M. (2018). Cross-cultural empirical aesthetics. *Progress in Brain Research*, *237*, 77–103. https://doi.org/10.1016/bs.pbr.2018.03.002

Chen, W., Liu, C. H., & Nakabayashi, K. (2012). Beauty hinders attention switch in change detection: The role of facial attractiveness and distinctiveness. *PLoS ONE*, *7*(2). https://doi.org/10.1371/journal.pone.0032897

Coetzee, V., Re, D., Perrett, D. I., Tiddeman, B. P., & Xiao, D. (2011). Judging the health and attractiveness of female faces: Is the most attractive level of facial adiposity also considered the healthiest? *Body Image*, *8*(2), 190–193. https://doi.org/10.1016/j.bodyim.2010.12.003

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: Quasi-experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology*, *50*(5), 925–935. https://doi.org/10.1037/0022-3514.50.5.925

Cunningham, M. R., Barbee, A. P., & Pike, C. L. (1990). What do women want? Facialmetric assessment of multiple motives in the perception of male facial physical attractiveness. *Journal of Personality and Social Psychology*, *59*(1), 61–72. https://doi.org/10.1037/0022-3514.59.1.61

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

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*(3), 285–290. https://doi.org/10.1037/h0033731

Dissanayake, E. (2015). “Aesthetic Primitives”: Fundamental biological elements of a naturalistic aesthetics. *Aisthesis*, *8*(1), 6–24. https://doi.org/10.13128/Aisthesis-16203

Durkee, P. K., & Ayers, J. D. (2020). Is facial width-to-height ratio reliably associated with social inferences? A large cross-national examination. *PsyArXiv*, 1–15. https://doi.org/10.31234/osf.io/tpngz

Fang, X., Sauter, D. A., & van Kleef, G. A. (2020). Unmasking smiles: the influence of culture and intensity on interpretations of smiling expressions. *Journal of Cultural Cognitive Science*, *4*, 293–308. https://doi.org/10.1007/s41809-019-00053-1

Fang, X., van Kleef, G. A., & Sauter, D. A. (2019). Revisiting cultural differences in emotion perception between easterners and westerners: Chinese perceivers are accurate, but see additional non-intended emotions in negative facial expressions. *Journal of Experimental Social Psychology*, *82*, 152–159. https://doi.org/10.1016/j.jesp.2019.02.003

Faust, N. T., Chatterjee, A., & Christopoulos, G. I. (2018). The effect of unrelated social exchanges on facial attractiveness judgments. *Journal of Experimental Social Psychology*, *79*, 290–300. https://doi.org/10.1016/j.jesp.2018.08.010

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). *LabVanced: a unified JavaScript framework for online studies. In International Conference on Computational Social Science (Cologne)*.

Fink, B., Grammer, K., & Matts, P. J. (2006). Visible skin color distribution plays a role in the perception of age, attractiveness, and health in female faces. *Evolution and Human Behavior*, *27*(6), 433–442. https://doi.org/10.1016/j.evolhumbehav.2006.08.007

Gangestad, S. W., Thornhill, R., & Yeo, R. A. (1994). Facial attractiveness, developmental stability, and fluctuating asymmetry. *Ethology and Sociobiology*, *15*(2), 73–85. https://doi.org/10.1016/0162-3095(94)90018-3

Gerger, G., Leder, H., Tinio, P. P. L., & Schacht, A. (2011). Faces versus patterns: Exploring aesthetic reactions using facial EMG. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(3), 241–250. https://doi.org/10.1037/a0024154

Goller, J., Leder, H., Cursiter, H., & Jenkins, R. (2018). Anchoring effects in facial attractiveness. *Perception*, *47*(10–11), 1043–1053. https://doi.org/10.1177/0301006618802696

Goller, J., Mitrovic, A., & Leder, H. (2019). Effects of liking on visual attention in faces and paintings. *Acta Psychologica*, *197*, 115–123. https://doi.org/10.1016/j.actpsy.2019.05.008

Grammer, K., & Thornhill, R. (1994). Human (Homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology*, *108*(3), 233–242.

Holzleitner, I. J., Lee, A. J., Hahn, A. C., Kandrik, M., Bovet, J., Renoult, J. P., Simmons, D., Garrod, O., DeBruine, L. M., & Jones, B. C. (2019). Comparing theory-driven and data-driven attractiveness models using images of real women's faces. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1589–1595. https://doi.org/10.1037/xhp0000685

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 199–209. https://doi.org/10.1037/0096-1523.32.2.199

Hu, Y., Abbasi, N. ul H., Zhang, Y., & Chen, H. (2018). The effect of target sex, sexual dimorphism, and facial attractiveness on perceptions of target attractiveness and trustworthiness. *Frontiers in Psychology*, *9*, 1–8. https://doi.org/10.3389/fpsyg.2018.00942

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Johnston, V. S., Hagel, R., Franklin, M., Fink, B., & Grammer, K. (2001). Male facial attractiveness: Evidence for hormone-mediated adaptive design. *Evolution and Human Behavior*, *22*(4), 251–267. https://doi.org/10.1016/S1090-5138(01)00066-6

Jones, A. L. (2017). The influence of shape and colour cue classes on facial health perception. *Evolution and Human Behavior*, *39*(1), 1–11. https://doi.org/10.1016/j.evolhumbehav.2017.09.005

Jones, B. C., DeBruine, L. M., Little, A. C., Burriss, R. P., & Feinberg, D. R. (2007). Social transmission of face preferences among humans. *Proceedings of the Royal Society B: Biological Sciences*, *274,* 899–903. https://doi.org/10.1098/rspb.2006.0205

Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2004). When facial attractiveness is only skin deep. *Perception*, *33*(5), 569–576. https://doi.org/10.1068/p3463

Jones, B. C., Little, A. C., Penton-Voak, I. S., Tiddeman, B. P., Burt, D. M., & Perrett, D. I. (2001). Facial symmetry and judgements of apparent health: Support for a “good genes” explanation of the attractiveness-symmetry relationship. *Evolution and Human Behavior*, *22*(6), 417–429. https://doi.org/10.1016/S1090-5138(01)00083-6

Jones, B. C., Perrett, D. I., Little, A. C., Boothroyd, L., Cornwell, R. E., Feinberg, D. R., Tiddeman, B. P., Whiten, S., Pitman, R. M., Hillier, S. G., Burt, D. M., Stirrat, M. R., Smith, M. J. L., & Moore, F. R. (2005). Menstrual cycle, pregnancy and oral contraceptive use alter attraction to apparent health in faces. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1561), 347–354. https://doi.org/10.1098/rspb.2004.2962

Jones, Benedict C., DeBruine, L. M., Perrett, D. I., Little, A. C., Feinberg, D. R., & Law Smith, M. J. (2008). Effects of menstrual cycle phase on face preferences. *Archives of Sexual Behavior*, *37*, 78–84. https://doi.org/10.1007/s10508-007-9268-y

Jones, D. (1995). Sexual selection, physical attractiveness, and facial neoteny. *Current Anthropology*, *36*(5), 723–748. https://doi.org/10.1086/204427

Jones, D., & Hill, K. (1993). Criteria of facial attractiveness in five populations. *Human Nature*, *4*(3), 271–296. https://doi.org/10.1007/BF02692202

Kaisler, R. E., & Leder, H. (2017). Combined effects of gaze and orientation of faces on person judgments in social situations. *Frontiers in Psychology*, *8,* 1–10. https://doi.org/10.3389/fpsyg.2017.00259

Kniffin, K. M., & Wilson, D. S. (2004). The effect of nonphysical traits on the perception of physical attractiveness. Three naturalistic studies. *Evolution and Human Behavior, 25*(2), 88–101. https://doi.org/10.1016/S1090-5138(04)00006-6

Koda, T., & Ruttkay, Z. (2009). Cultural differences in using facial parts as cues to recognize emotions in avatars. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. Högni Vilhjálmsson (Eds.), *Intelligent Virtual Agents* (pp. 517–518). Springer.

Koehler, N., Simmons, L. W., Rhodes, G., & Peters, M. (2004). The relationship between sexual dimorphism in human faces and fluctuating asymmetry. *Proceedings of the Royal Society B: Biological Sciences*, *271*, 233–236. https://doi.org/10.1098/rsbl.2003.0146

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, *7*(4), 730–735. https://doi.org/10.1037/1528-3542.7.4.730

Krupp, D. B., DeBruine, L. M., & Jones, B. C. (2011). Apparent health encourages reciprocity. *Evolution and Human Behavior*, *32*(3), 198–203. https://doi.org/10.1016/j.evolhumbehav.2010.10.001

LaFrance, M., & Hecht, M. A. (1995). Why smiles generate leniency. *Personality and Social Psychology Bulletin*, *21*, 207–214. https://doi.org/10.1177/0146167295213002

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer’s physical attractiveness. *Journal of Personality and Social Psychology*, *29*(3), 299–304. https://doi.org/10.1037/h0036018

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*(3), 390–414. https://doi.org/10.1037/0033-2909.126.3.390

Langlois, J. H., Ritter, J. M., Roggman, L. A., & Vaughn, L. S. (1991). Facial diversity and infant preferences for attractive faces. *Developmental Psychology*, *27*(1), 79–84. https://doi.org/10.1037/0012-1649.27.1.79

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*(2), 115–121. https://doi.org/10.1111/j.1467-9280.1990.tb00079.x

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. https://doi.org/10.1111/j.1467-9280.1994.tb00503.x

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetics appreciation and aesthetic judgment. *British Journal of Psychology*, *95*, 489–508. https://doi.org/10.1348/0007126042369811

Leder, H., Goller, J., Rigotti, T., & Forster, M. (2016). Private and shared taste in art and face appreciation. *Frontiers in Human Neuroscience, 10*, 1–7. https://doi.org/10.3389/fnhum.2016.00155

Leder, H., Mitrovic, A., & Goller, J. (2016). How beauty determines gaze! Facial attractiveness and gaze duration in images of real world scenes. *I-Perception*, 1–12. https://doi.org/10.1177/2041669516664355

Leder, H., & Nadal, M. (2014). Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode - Developments and challenges in empirical aesthetics. *British Journal of Psychology*, *105*(4), 443–446. https://doi.org/10.1111/bjop.12084

Leder, H., Tinio, P. P. L., Brieber, D., Kröner, T., Jacobsen, T., & Rosenberg, R. (2019). Symmetry is not a universal law of beauty. *Empirical Studies of the Arts*, *37*(1), 104–114. https://doi.org/10.1177/0276237418777941

Leder, H., Tinio, P. P. L., Fuchs, I. M., & Bohrn, I. (2010). When attractiveness demands longer looks: The effects of situation and gender. *Quarterly Journal of Experimental Psychology*, *63*(9), 1858–1871. https://doi.org/10.1080/17470211003605142

Lee, A. J., Wright, M. J., Martin, N. G., Keller, M. C., & Zietsch, B. P. (2017). Facial trustworthiness is associated with heritable aspects of face shape. *Adaptive Human Behavior and Physiology*, *3*(4), 351–364. https://doi.org/10.1007/s40750-017-0073-0

Lewandowski, G. W., Aron, A., & Gee, J. (2007). Personality goes a long way: The malleability of opposite-sex physical attractiveness. *Personal Relationships*, *14*(4), 571–585. https://doi.org/10.1111/j.1475-6811.2007.00172.x

Light, L. L., Hollander, S., & Kayra-Stuart, F. (1981). Why Attractive People are Harder to Remember. *Personality and Social Psychology Bulletin*, *7*(2), 269–276. https://doi.org/10.1177/014616728172014

Little, A. C., Burt, D. M., Penton-Voak, I. S., & Perrett, D. I. (2001). Self-perceived attractiveness influences human female preferences for sexual dimorphism and symmetry in male faces. *Proceedings of the Royal Society B: Biological Sciences*, *268*(1462), 39–44. https://doi.org/10.1098/rspb.2000.1327

Little, A. C., Burt, D. M., & Perrett, D. I. (2006). What is good is beautiful: Face preference reflects desired personality. *Personality and Individual Differences*, *41*, 1107–1118. https://doi.org/10.1016/j.paid.2006.04.015

Little, A. C., & Hancock, P. J. B. (2002). The role of masculinity and distinctiveness in judgments of human male facial attractiveness. *British Journal of Psychology*, *93*, 451–464. https://doi.org/10.1348/000712602761381349

Little, A. C., & Jones, B. C. (2012). Variation in facial masculinity and symmetry preferences across the menstrual cycle is moderated by relationship context. *Psychoneuroendocrinology*, *37*, 999–1008. https://doi.org/10.1016/j.psyneuen.2011.11.007

Little, A. C., Jones, B. C., Burt, D. M., & Perrett, D. I. (2007). Preferences for symmetry in faces change across the menstrual cycle. *Biological Psychology*, *76*(3), 209–216. https://doi.org/10.1016/j.biopsycho.2007.08.003

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: Evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1638–1659. https://doi.org/10.1098/rstb.2010.0404

Little, A. C., Jones, B. C., Penton-Voak, I. S., Burt, D. M., & Perrett, D. I. (2002). Partnership status and the temporal context of relationships influence human female preferences for sexual dimorphism in male face shape. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1496), 1095–1100. https://doi.org/10.1098/rspb.2002.1984

Marcinkowska, U. M., Rantala, M. J., Lee, A. J., Kozlov, M. V., Aavik, T., Cai, H., Contreras-Garduño, J., David, O. A., Kaminski, G., Li, N. P., Onyishi, I. E., Prasai, K., Pazhoohi, F., Prokop, P., Cardozo, S. L. R., Sydney, N., Taniguchi, H., Krams, I., & Dixson, B. J. W. (2019). Women's preferences for men's facial masculinity are strongest under favorable ecological conditions. *Scientific Reports*, *9*(3387), 1–10. https://doi.org/10.1038/s41598-019-39350-8

Mitrovic, A., Goller, J., Tinio, P. P. L., & Leder, H. (2018). How relationship status and sociosexual orientation influence the link between facial attractiveness and visual attention. *PLoS ONE*, *13*(11), 1–15. https://doi.org/10.1371/journal.pone.0207477

Mitrovic, A., Hegelmaier, L. M., Leder, H., & Pelowski, M. (2020). Does beauty capture the eye, even if it's not (overtly) adaptive? A comparative eye-tracking study of spontaneous attention and visual preference with VAST abstract art. *Acta Psychologica*, *209*, 1–10. https://doi.org/10.1016/j.actpsy.2020.103133

Mitrovic, A., Tinio, P. P. L., & Leder, H. (2016). Consequences of beauty: Effects of rater sex and sexual orientation on the visual exploration and evaluation of attractiveness in real world scenes. *Frontiers in Human Neuroscience*, *10*, 1–10. https://doi.org/10.3389/fnhum.2016.00122

Nakamura, K., Ohta, A., Uesaki, S., Maeda, M., & Kawabata, H. (2020). Geometric morphometric analysis of japanese female facial shape in relation to psychological impression space. *Heliyon*, *6*, 1–8. https://doi.org/10.1016/j.heliyon.2020.e05148

Nakamura, K., & Watanabe, K. (2019). Data-driven mathematical model of east-asian facial attractiveness: The relative contributions of shape and reflectance to attractiveness judgements. *Royal Society Open Science*, *6*, 1–11. https://doi.org/10.1098/rsos.182189

Nakamura, K., & Watanabe, K. (2020). A new data-driven mathematical model dissociates attractiveness from sexual dimorphism of human faces. *Scientific Reports*, *10*(16588), 1–12. https://doi.org/10.1038/s41598-020-73472-8

Okubo, M., Ishikawa, K., Kobayashi, A., Laeng, B., & Tommasi, L. (2015). Cool guys and warm husbands: The effect of smiling on male facial attractiveness for short- and long-term relationships. *Evolutionary Psychology*, *13*(3), 1–8. https://doi.org/10.1177/1474704915600567

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, *9*(1), 128–133. https://doi.org/10.1037/a0014520

Penton-Voak, I. S., Little, A. C., Jones, B. C., Burt, D. M., Tiddeman, B. P., & Perrett, D. I. (2003). Female condition influences preferences for sexual dimorphism in faces of male humans (homo sapiens). *Journal of Comparative Psychology*, *117*(3), 264–271. https://doi.org/10.1037/0735-7036.117.3.264

Penton-Voak, I. S., Perrett, D. I., Castles, D. L., Kobayashi, T., Burt, D. M., Murray, L. K., & Minamisawa, R. (1999). Menstrual cycle alters face preference. *Nature*, *399*(6738), 741–742. https://doi.org/10.1038/21557

Perrett, D. I., Lee, K. J., Penton-Voak, I. S., Rowland, D., Yoshikawa, S., Burt, D. M., Henzi, S. P., Castles, D. L., & Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, *246*(5429), 170. https://doi.org/10.1038/246170a0

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature*, *368*(6468), 239–242. https://doi.org/10.1038/368239a0

Prum, R. O. (2012). Aesthetic evolution by mate choice: Darwin’s really dangerous idea. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 2253–2265. https://doi.org/10.1098/rstb.2011.0285

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. https://www.r-project.org/

Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015). Dominant voices and attractive faces: The contribution of visual and auditory information to integrated person impressions. *Journal of Nonverbal Behavior*, *39*, 355–370. https://doi.org/10.1007/s10919-015-0214-8

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*, 199–226. https://doi.org/10.1146/annurev.psych.57.102904.190208

Rhodes, G., Chan, J., Zebrowitz, L. A., & Simmons, L. W. (2003). Does sexual dimorphism in human faces signal health? *Proceedings of the Royal Society B: Biological Sciences*, *270*, 93–95. https://doi.org/10.1098/rsbl.2003.0023

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition*, *19*(1), 57–70. https://doi.org/10.1521/soco.19.1.57.18961

Rhodes, G., Hickford, C., & Jeffery, L. (2000). Sex-typicality and attractiveness: Are supermale and superfemale faces super-attractive? *British Journal of Psychology*, *91*(1), 125–140. https://doi.org/10.1348/000712600161718

Rhodes, G., Sumich, A., & Byatt, G. (1999). Are average facial configurations attractive only because of their symmetry? *Psychological Science*, *10*(1), 52–58. https://doi.org/10.1111/1467-9280.00106

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological Science*, *7*(2), 105–110. https://doi.org/10.1111/j.1467-9280.1996.tb00338.x

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*(2), 260–264. https://doi.org/10.1037/a0014681

Schacht, A., Werheid, K., & Sommer, W. (2008). The appraisal of facial beauty is rapid but not mandatory. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(2), 132–142. https://doi.org/10.3758/CABN.8.2.132

Schall, S., Quigley, C., Onat, S., & König, P. (2009). Visual stimulus locking of EEG is modulated by temporal congruency of auditory stimuli. *Experimental Brain Research*, *198*, 137–151. https://doi.org/10.1007/s00221-009-1867-5

Stan Development Team. (2020). *RStan: The R interface to Stan. R package version 2.19.2. http://mc-stan.org/*. http://mc-stan.org/

Stephen, I. D., Scott, I. M. L., Coetzee, V., Pound, N., Perrett, D. I., & Penton-Voak, I. S. (2012). Cross-cultural effects of color, but not morphological masculinity, on perceived attractiveness of men's faces. *Evolution and Human Behavior*, *33*(4), 260–267. https://doi.org/10.1016/j.evolhumbehav.2011.10.003

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Sciences*, *22*(3), 197–200. https://doi.org/10.1016/j.tics.2017.12.003

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(37), 9210–9215. https://doi.org/10.1073/pnas.1807222115

Tan, R., & Zheng, L. (2020). Individual differences in preferences for sexual dimorphism in faces among Chinese adolescents in single-sex and mixed-sex environments. *Personality and Individual Differences*, *153*, 1–4. https://doi.org/10.1016/j.paid.2019.109648

Thornhill, R., & Møller, A. P. (1997). Developmental stability, disease and medicine. *Biological Reviews*, *72*(4), 497–548. https://doi.org/10.1111/j.1469-185X.1997.tb00022.x

Tinlin, R. M., Watkins, C. D., Welling, L. L. M., Debruine, L. M., Al-Dujaili, E. A. S., & Jones, B. C. (2013). Perceived facial adiposity conveys information about women's health. *British Journal of Psychology*, *104*(2), 235–248. https://doi.org/10.1111/j.2044-8295.2012.02117.x

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738. https://doi.org/10.1037/a0032335

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145. https://doi.org/10.1016/S0953-5438(00)00031-X

Trujillo, L. T., Jankowitsch, J. M., & Langlois, J. H. (2014). Beauty is in the ease of the beholding: A neurophysiological test of the averageness theory of facial attractiveness. *Cognitive, Affective and Behavioral Neuroscience*, *14*(3), 1061–1076. https://doi.org/10.3758/s13415-013-0230-2

Tsukiura, T., & Cabeza, R. (2011). Shared brain activity for aesthetic and moral judgments: Implications for the beauty-is-good stereotype. *Social Cognitive and Affective Neuroscience*, *6*(1), 138–148. https://doi.org/10.1093/scan/nsq025

Valentova, J. V., Varella, M. A. C., Havlíček, J., & Kleisner, K. (2017). Positive association between vocal and facial attractiveness in women but not in men: A cross-cultural study.

Behavioural Processes, *135*, 95–100. https://doi.org/10.1016/j.beproc.2016.12.005

Valuch, C., Pflüger, L. S., Wallner, B., Laeng, B., & Ansorge, U. (2015). Using eye tracking to test for individual differences in attention to attractive faces. *Frontiers in Psychology*, *6*, 1–13. https://doi.org/10.3389/fpsyg.2015.00042

Vera Cruz, G. (2018). The impact of face skin tone vs. face symmetry on perceived facial attractiveness. *Journal of General Psychology*, *145*(2), 183–198. https://doi.org/10.1080/00221309.2018.1459452

Vingilis-Jaremko, L., Maurer, D., & Gao, X. (2014). The influence of averageness on judgments of facial attractiveness: No own-age or own-sex advantage among children attending single-sex schools. *Journal of Experimental Child Psychology*, *120*(1), 1–16. https://doi.org/10.1016/j.jecp.2013.10.006

Wang, X., Liu, S., Han, S., Gan, Y., Li, W., Xu, Q., & Zhang, L. (2020). Roles of social knowledge and sexual dimorphism in the evaluation of facial attractiveness. *Journal of Experimental Social Psychology*, *88*, 1–8. https://doi.org/10.1016/j.jesp.2020.103963

Wang, Y., Yao, P., & Zhou, G. (2015). The influence of facial attractiveness and personality labels on men and women’s mate preference. *Acta Psychologica Sinica*, *47*(1), 108–118. https://doi.org/10.3724/sp.j.1041.2015.00108

Wells, T., Baguley, T., Sergeant, M., & Dunn, A. (2013). Perceptions of human attractiveness comprising face and voice cues. *Archives of Sexual Behavior*, *42*, 805–811. https://doi.org/10.1007/s10508-012-0054-0

Zhang, Y., Kong, F., Zhong, Y., & Kou, H. (2014). Personality manipulations: Do they modulate facial attractiveness ratings? *Personality and Individual Differences*, *70*, 80–84. https://doi.org/10.1016/j.paid.2014.06.033

Appendix: Table Directory

Table 1.
Averageness ratings for every face in the stimulus set

Face	Descriptive Metric				
	Mean	SD	Min	Med	Max
1	5.22	1.39	1.00	5.00	7.00
2	4.30	1.54	1.00	4.00	7.00
3	4.39	1.38	1.00	5.00	7.00
4	5.09	1.26	1.00	5.00	7.00
5	4.76	1.34	1.00	5.00	7.00
6	3.49	1.46	1.00	3.00	7.00
7	4.29	1.37	1.00	5.00	7.00
8	4.29	1.57	1.00	5.00	7.00
9	3.66	1.48	1.00	4.00	7.00
10	3.94	1.57	1.00	4.00	7.00
11	4.48	1.52	1.00	5.00	7.00
12	3.82	1.45	1.00	4.00	7.00
13	4.18	1.38	1.00	4.00	7.00
14	3.63	1.42	1.00	4.00	7.00
15	3.99	1.44	1.00	4.00	7.00
16	3.90	1.42	1.00	4.00	7.00
17	4.81	1.49	1.00	5.00	7.00
18	4.72	1.42	1.00	5.00	7.00
19	5.31	1.31	1.00	6.00	7.00
20	4.50	1.39	1.00	5.00	7.00
21	4.58	1.54	1.00	5.00	7.00
22	5.50	1.27	1.00	6.00	7.00
23	3.11	1.32	1.00	3.00	7.00
24	3.92	1.43	1.00	4.00	7.00
25	3.92	1.46	1.00	4.00	7.00
26	3.52	1.42	1.00	3.00	7.00
27	4.20	1.44	1.00	4.00	7.00
28	3.66	1.52	1.00	3.50	7.00
29	4.39	1.45	1.00	5.00	7.00

30	3.32	1.31	1.00	3.00	7.00
31	3.96	1.53	1.00	4.00	7.00
32	3.87	1.50	1.00	4.00	7.00
33	4.61	1.36	1.00	5.00	7.00
34	4.63	1.47	1.00	5.00	7.00
35	3.83	1.28	1.00	4.00	7.00
36	3.62	1.47	1.00	3.50	7.00
37	4.35	1.46	1.00	5.00	7.00
38	5.05	1.27	1.00	5.00	7.00
39	3.65	1.42	1.00	4.00	7.00
40	4.61	1.58	1.00	5.00	7.00

Note. Ratings are a combination both picture and video condition, *N* = 206.

Table 2.
Sexual Dimorphism ratings for every face in the stimulus set

	Descriptive Metric				
Face	Mean	SD	Min	Med	Max
1	4.75	1.50	1.00	5.00	7.00
2	4.99	1.32	1.00	5.00	7.00
3	5.61	1.10	2.00	6.00	7.00
4	4.69	1.34	1.00	5.00	7.00
5	5.83	0.96	2.00	6.00	7.00
6	5.57	1.15	1.00	6.00	7.00
7	4.83	1.31	1.00	5.00	7.00
8	5.13	1.20	1.00	5.00	7.00
9	4.76	1.12	2.00	5.00	7.00
10	5.49	1.04	2.00	6.00	7.00
11	4.64	1.41	1.00	5.00	7.00
12	5.19	1.24	2.00	5.00	7.00
13	5.02	1.36	1.00	5.00	7.00
14	4.90	1.16	1.00	5.00	7.00
15	5.64	0.98	3.00	6.00	7.00
16	5.46	1.08	2.00	6.00	7.00
17	5.06	1.18	1.00	5.00	7.00

18	5.92	1.01	2.00	6.00	7.00
19	3.51	1.38	1.00	3.00	7.00
20	5.30	1.20	2.00	6.00	7.00
21	4.95	1.25	1.00	5.00	7.00
22	4.83	1.17	2.00	5.00	7.00
23	4.55	1.26	1.00	5.00	7.00
24	4.97	1.15	1.00	5.00	7.00
25	5.49	1.12	1.00	6.00	7.00
26	5.89	0.94	3.00	6.00	7.00
27	3.75	1.43	1.00	4.00	7.00
28	4.46	1.23	1.00	5.00	7.00
29	5.37	1.09	2.00	5.00	7.00
30	5.25	1.13	1.00	5.00	7.00
31	5.05	1.30	1.00	5.00	7.00
32	4.92	1.20	1.00	5.00	7.00
33	4.63	1.34	1.00	5.00	7.00
34	4.47	1.28	1.00	5.00	7.00
35	5.52	1.09	1.00	6.00	7.00
36	5.19	1.16	1.00	5.00	7.00
37	5.54	1.08	1.00	6.00	7.00
38	5.82	1.12	2.00	6.00	7.00
39	5.21	1.10	2.00	5.00	7.00
40	5.76	0.98	2.00	6.00	7.00

Note. Ratings are a combination both picture and video condition, $N = 206$.

Table 3.
Health ratings for every face in the stimulus set

	Descriptive Metric				
Face	Mean	SD	Min	Med	Max
1	4.99	1.33	2.00	5.00	7.00
2	5.22	1.16	2.00	5.00	7.00
3	4.79	1.55	1.00	5.00	7.00
4	5.13	1.31	1.00	5.00	7.00
5	5.59	1.17	1.00	6.00	7.00

6	5.14	1.29	1.00	5.00	7.00
7	5.02	1.36	1.00	5.00	7.00
8	4.75	1.36	1.00	5.00	7.00
9	4.17	1.50	1.00	4.00	7.00
10	3.19	1.71	1.00	3.00	7.00
11	5.29	1.16	2.00	5.00	7.00
12	4.54	1.43	1.00	5.00	7.00
13	5.17	1.17	1.00	5.00	7.00
14	5.32	1.11	2.00	5.00	7.00
15	5.28	1.18	1.00	5.00	7.00
16	5.35	1.12	2.00	5.50	7.00
17	3.19	1.57	1.00	3.00	7.00
18	5.74	1.05	1.00	6.00	7.00
19	4.53	1.29	1.00	5.00	7.00
20	4.99	1.30	2.00	5.00	7.00
21	4.64	1.37	1.00	5.00	7.00
22	4.78	1.35	1.00	5.00	7.00
23	4.00	1.49	1.00	4.00	7.00
24	3.93	1.54	1.00	4.00	7.00
25	5.77	0.99	3.00	6.00	7.00
26	5.58	1.01	3.00	6.00	7.00
27	4.69	1.32	1.00	5.00	7.00
28	3.84	1.42	1.00	4.00	7.00
29	5.15	1.22	2.00	5.00	7.00
30	4.17	1.51	1.00	4.00	7.00
31	4.20	1.43	1.00	4.00	7.00
32	2.59	1.32	1.00	2.00	7.00
33	4.74	1.43	1.00	5.00	7.00
34	4.31	1.56	1.00	5.00	7.00
35	5.44	1.14	1.00	6.00	7.00
36	4.71	1.34	1.00	5.00	7.00
37	4.89	1.34	2.00	5.00	7.00
38	5.97	0.97	2.00	6.00	7.00

39	4.03	1.60	1.00	4.00	7.00
40	4.59	1.27	1.00	5.00	7.00

Note. Ratings are a combination both picture and video condition, $N = 206$.

Table 4.
Likeability ratings for every face in the stimulus set

	Descriptive Metric				
Face	Mean	SD	Min	Med	Max
1	3.84	1.35	1.00	4.00	7.00
2	4.39	1.49	1.00	5.00	7.00
3	5.24	1.22	1.00	5.00	7.00
4	4.78	1.37	1.00	5.00	7.00
5	5.38	1.23	2.00	6.00	7.00
6	4.89	1.46	1.00	5.00	7.00
7	4.77	1.44	1.00	5.00	7.00
8	4.19	1.30	1.00	4.00	7.00
9	3.73	1.43	1.00	4.00	7.00
10	4.51	1.20	1.00	5.00	7.00
11	4.65	1.40	1.00	5.00	7.00
12	4.09	1.38	1.00	4.00	7.00
13	4.91	1.21	1.00	5.00	7.00
14	4.97	1.40	1.00	5.00	7.00
15	4.67	1.35	1.00	5.00	7.00
16	4.88	1.23	2.00	5.00	7.00
17	3.25	1.37	1.00	3.00	7.00
18	5.11	1.19	1.00	5.00	7.00
19	4.05	1.57	1.00	4.00	7.00
20	4.30	1.35	1.00	4.00	7.00
21	4.00	1.44	1.00	4.00	7.00
22	4.74	1.41	1.00	5.00	7.00
23	3.20	1.35	1.00	3.00	7.00
24	4.32	1.48	1.00	4.00	7.00
25	5.89	1.05	3.00	6.00	7.00
26	4.47	1.39	1.00	5.00	7.00

27	4.08	1.40	1.00	4.00	7.00
28	3.33	1.27	1.00	3.00	6.00
29	5.18	1.20	1.00	5.00	7.00
30	3.89	1.39	1.00	4.00	7.00
31	4.17	1.37	1.00	4.00	7.00
32	2.79	1.35	1.00	3.00	6.00
33	3.90	1.51	1.00	4.00	7.00
34	4.16	1.63	1.00	4.00	7.00
35	4.18	1.38	1.00	4.00	7.00
36	4.38	1.46	1.00	5.00	7.00
37	4.45	1.44	1.00	5.00	7.00
38	5.12	1.18	2.00	5.00	7.00
39	3.85	1.40	1.00	4.00	7.00
40	4.57	1.46	1.00	5.00	7.00

Note. Ratings are a combination both picture and video condition, N = 206.

Table 5.
Trustworthiness ratings for every face in the stimulus set

	Descriptive Metric				
Face	Mean	SD	Min	Med	Max
1	4.26	1.39	1.00	4.00	7.00
2	4.63	1.37	1.00	5.00	7.00
3	5.34	1.22	1.00	5.50	7.00
4	4.44	1.38	1.00	5.00	7.00
5	5.17	1.23	2.00	5.00	7.00
6	5.21	1.43	1.00	5.00	7.00
7	4.99	1.23	2.00	5.00	7.00
8	4.47	1.27	1.00	5.00	7.00
9	4.16	1.46	1.00	4.00	7.00
10	4.53	1.25	1.00	5.00	7.00
11	4.66	1.31	1.00	5.00	7.00
12	4.31	1.36	1.00	4.00	7.00
13	4.90	1.25	1.00	5.00	7.00
14	4.73	1.46	1.00	5.00	7.00

15	4.54	1.36	1.00	5.00	7.00
16	4.98	1.22	2.00	5.00	7.00
17	3.90	1.47	1.00	4.00	7.00
18	4.98	1.27	1.00	5.00	7.00
19	4.27	1.37	1.00	5.00	7.00
20	4.30	1.34	1.00	5.00	7.00
21	4.07	1.34	1.00	4.00	7.00
22	4.56	1.33	1.00	5.00	7.00
23	3.41	1.40	1.00	3.00	7.00
24	4.20	1.52	1.00	4.00	7.00
25	5.76	1.06	3.00	6.00	7.00
26	4.60	1.36	1.00	5.00	7.00
27	4.19	1.34	1.00	4.00	7.00
28	3.79	1.30	1.00	4.00	7.00
29	4.87	1.21	1.00	5.00	7.00
30	4.03	1.44	1.00	4.00	7.00
31	3.99	1.42	1.00	4.00	7.00
32	2.94	1.41	1.00	3.00	7.00
33	3.81	1.54	1.00	4.00	7.00
34	4.17	1.58	1.00	4.00	7.00
35	4.21	1.36	1.00	4.00	7.00
36	4.64	1.32	1.00	5.00	7.00
37	4.48	1.45	1.00	5.00	7.00
38	4.69	1.16	1.00	5.00	7.00
39	4.06	1.43	1.00	4.00	7.00
40	4.48	1.44	1.00	5.00	7.00

Note. Ratings are a combination both picture and video condition, *N* = 206.

Table 6.
Dominance ratings for every face in the stimulus set

	Descriptive Metric				
Face	Mean	SD	Min	Med	Max
1	4.77	1.48	1.00	5.00	7.00
2	3.83	1.63	1.00	4.00	7.00

3	3.93	1.43	1.00	4.00	7.00
4	3.60	1.51	1.00	3.00	7.00
5	3.92	1.32	1.00	4.00	7.00
6	3.96	1.48	1.00	4.00	7.00
7	3.97	1.51	1.00	4.00	7.00
8	3.35	1.45	1.00	3.00	7.00
9	4.23	1.75	1.00	4.00	7.00
10	2.74	1.36	1.00	3.00	7.00
11	4.47	1.52	1.00	5.00	7.00
12	4.14	1.65	1.00	4.00	7.00
13	3.59	1.46	1.00	4.00	7.00
14	3.29	1.33	1.00	3.00	6.00
15	4.33	1.37	1.00	4.50	7.00
16	3.61	1.45	1.00	3.00	7.00
17	4.21	1.76	1.00	5.00	7.00
18	4.51	1.37	1.00	5.00	7.00
19	4.32	1.58	1.00	5.00	7.00
20	4.13	1.45	1.00	4.00	7.00
21	4.05	1.54	1.00	4.00	7.00
22	4.26	1.40	1.00	5.00	7.00
23	2.95	1.62	1.00	3.00	7.00
24	3.56	1.34	1.00	4.00	7.00
25	3.24	1.28	1.00	3.00	6.00
26	4.20	1.50	1.00	4.50	7.00
27	3.36	1.50	1.00	3.00	7.00
28	3.60	1.56	1.00	3.50	7.00
29	3.69	1.50	1.00	4.00	7.00
30	3.53	1.43	1.00	3.00	7.00
31	3.94	1.50	1.00	4.00	7.00
32	3.65	1.49	1.00	4.00	7.00
33	4.15	1.50	1.00	4.00	7.00
34	3.12	1.44	1.00	3.00	7.00
35	4.44	1.37	1.00	5.00	7.00

36	3.61	1.31	1.00	4.00	7.00
37	3.66	1.42	1.00	4.00	7.00
38	4.34	1.44	1.00	5.00	7.00
39	3.70	1.53	1.00	4.00	7.00
40	3.73	1.42	1.00	4.00	7.00

Note. Ratings are a combination both picture and video condition, N = 206.

Table 7.
Beauty ratings for every face in the stimulus set

Face	Descriptive Metric				
	Mean	SD	Min	Med	Max
1	4.15	1.52	1.00	4.00	7.00
2	4.51	1.45	1.00	5.00	7.00
3	4.92	1.37	1.00	5.00	7.00
4	4.43	1.50	1.00	5.00	7.00
5	4.23	1.42	1.00	4.00	7.00
6	3.98	1.45	1.00	4.00	7.00
7	4.05	1.43	1.00	4.00	7.00
8	3.40	1.34	1.00	3.00	7.00
9	3.68	1.41	1.00	4.00	7.00
10	4.38	1.45	1.00	5.00	7.00
11	4.35	1.37	1.00	5.00	7.00
12	3.88	1.43	1.00	4.00	7.00
13	3.44	1.33	1.00	3.00	7.00
14	3.51	1.48	1.00	3.00	7.00
15	3.95	1.39	1.00	4.00	7.00
16	4.76	1.39	1.00	5.00	7.00
17	3.36	1.57	1.00	3.00	7.00
18	4.57	1.41	1.00	5.00	7.00
19	2.95	1.45	1.00	3.00	7.00
20	4.58	1.40	1.00	5.00	7.00
21	4.10	1.50	1.00	4.00	7.00
22	4.30	1.39	1.00	5.00	7.00
23	2.62	1.26	1.00	2.00	7.00

24	2.83	1.32	1.00	3.00	7.00
25	5.00	1.24	1.00	5.00	7.00
26	4.89	1.28	1.00	5.00	7.00
27	2.85	1.39	1.00	3.00	7.00
28	2.80	1.34	1.00	3.00	7.00
29	5.22	1.36	1.00	5.00	7.00
30	2.95	1.34	1.00	3.00	7.00
31	3.53	1.41	1.00	3.00	7.00
32	1.88	1.03	1.00	2.00	6.00
33	3.67	1.56	1.00	4.00	7.00
34	3.10	1.51	1.00	3.00	7.00
35	4.07	1.48	1.00	4.00	7.00
36	4.01	1.47	1.00	4.00	7.00
37	3.45	1.42	1.00	3.00	7.00
38	4.96	1.32	1.00	5.00	7.00
39	3.30	1.44	1.00	3.00	7.00
40	2.50	1.27	1.00	2.00	7.00

Note. Ratings are a combination both picture and video condition, $N = 206$.