Journal of Experimental Psychology:
Human Perception and Performance
2006, Vol. 32, No. 2, 199–209

0096-1523/06/$12.00 DOI: 10.1037/0096-1523.32.2.199

Once More: Is Beauty in the Eye of the Beholder? Relative Contributions of Private and Shared Taste to Judgments of Facial Attractiveness

Johannes Hönekopp
Technische Universität Chemnitz

Misconstruing the meaning of Cronbach's alpha, experts on facial attractiveness have conveyed the impression that facial-attractiveness judgment standards are largely shared. This claim is unsubstantiated, because information necessary for deciding whether judgments of facial attractiveness are more influenced by commonly shared or by privately held evaluation standards is lacking. Three experiments, using diverse face and rater samples to investigate the relative contributions of private and shared taste to judgments of facial attractiveness, are reported. These experiments show that for a variety of ancillary conditions, and contrary to the prevalent notion in the literature, private taste is about as powerful as shared taste. Important implications for scientific research strategy and laypeople's self-esteem are discussed.

Keywords: attractiveness, faces, reliability, agreement

The proverb states that "beauty is in the eye of the beholder." When it comes to the attractiveness of faces, this proverb is seemingly wrong: Although some researchers in this area have pointed to the importance of interindividual differences between judges of attractiveness (e.g., Dion, 2002; Lucker, Beane, & Guire, 1981; Little, Penton-Voak, Burt, & Perrett, 2002), the majority of scholarly writers have stressed that consensus among raters is high. As a summary of their meta-analysis on the reliability of physical-attractiveness judgments, Langlois et al. (2000), for example, wrote that "raters agreed about the attractiveness of both adults and children" (p. 399). Similar statements are plentiful, even if one considers only articles in leading journals (Chen, German, & Zaidel, 1997; Cunningham, Roberts, Barbee, Druen, & Wu, 1995; Kowner, 1996; Langlois & Roggman, 1990; Mealey, Bridgstock, & Townsend, 1999; Perrett, May, & Yoshikawa, 1994; Rhodes, Zebrowitz, et al., 2001).

In this article, I argue that researchers who give the impression that taste[1] is largely shared misinterpret the findings on which they base this claim. I also argue that the data reported so far cannot answer the question of to what extent private and shared taste contribute to judgments of facial attractiveness. This gives rise to a curious situation. Whereas proverbial wisdom stresses the importance of private taste, and researchers emphasize the importance of shared taste, everyday experience suggests that both factors may be significant: Whereas the pay of top models bespeaks the importance of shared taste, the recollection of any discussion between yourself and a friend about the attractiveness of passersby probably advocates for the importance of private taste. The aim of this article is to inform about the relative contributions of private and shared taste to judgments of facial attractiveness. I define s*hared taste* as subsuming all attractiveness standards that let, on average, two judges agree to some extent about the attractiveness of faces; I define *private taste* as subsuming all attractiveness standards of a single judge that give rise to replicable disagreement with shared taste.

I first describe the standard paradigm in the field. I then explain why it gives rise to misinterpretations and why it is insufficient to disentangle the relative impact of shared and private taste. Then, I demonstrate why it is important to provide a correct answer to the question of how strongly private taste and shared taste affect attractiveness ratings. Afterward, I suggest a more suitable approach to this issue before discussing the likely impact of sample composition. Finally, I present three experiments examining the relative contributions of private and shared taste to judgments of facial attractiveness.

Why the Standard Paradigm Cannot Inform About the Relative Contributions of Private and Shared Taste

Most studies on facial-attractiveness judgments share the same simple paradigm: Each judge evaluates each target face on a rating scale. For the sake of efficient communication, I call each face's average rating its *face score*; I call the average of a single judge's ratings a *judge score*. The bulk of research has been stimulus centered. Face scores, which reflect the shared taste of the judge sample, are usually computed, and relationships between them and facial properties are examined. One good example of this paradigm is the finding that faces with large eyes tend to receive more favorable face scores (e.g., Cunningham, Barbee, & Pike, 1990; Geldart, Maurer, & Carney, 1999; McArthur & Apatow, 1983–1984). Being able to partly explain face scores is remarkable, and

I express my gratitude for the help of Steffi Anhut, Bernd Marcus, Sandra Hopps, Andreas Keinath, David Kenny, Ilona Rappholt, Frank Renkewitz, Peter Sedlmeier, and Udo Rudolph.

Correspondence concerning this article should be addressed to Johannes Hönekopp, Institut für Psychologie, Technische Universität Chemnitz, Wilhelm-Raabe Strasse 43, D-09120, Chemnitz, Germany. E-mail: johannes.hoenekopp@phil.tu-chemnitz.de

[1] The scope of this article is restricted to facial attractiveness. Therefore, *taste* is always meant to refer to facial attractiveness.

findings like this certainly challenge the proverbial notion of beauty being in the eye of the beholder. It is this fact that probably motivates many scholars to stress that taste is largely shared. On what evidence does this claim rest? As already mentioned, most studies on facial attractiveness seek to explain face scores. Virtually all studies report very high Cronbach's alpha reliability coefficients for these face scores. It is such findings that induce the prevailing statements in the scientific literature that raters agree about who is and who is not attractive.

Agreement Between Samples of Judges

The basic idea behind the use of Cronbach's alpha in facial-attractiveness research is to treat each judge as an item of a test that sets out to measure the average attractiveness judgment that each face would receive from the total population of judges. The face score of any face serves as an estimate of this population average. Cronbach's alpha indicates the reliability of face scores. It indicates what correlation should be expected between the obtained face scores and a second set of face scores that stems from another, equally large sample of judges. Thus, a high Cronbach's alpha signifies that the obtained face scores approximate the average evaluations of the population of judges. This is a happy and important finding—but is it good enough evidence to claim that perceptions of facial attractiveness are predominantly shared? I argue that the answer is no.

The answer is no because the reliability of a test partly depends on the number of items (here, judges) used, and lengthening a test will yield a more reliable measure (e.g., Cortina, 1993). Thus, items with only weak intercorrelations may constitute a very reliable (although long) test. For example, using 80 judges whose judgments only reach a correlation of .10 with each other would yield a very respectable Cronbach's alpha of .90 for the resulting face scores.[2] Although Cronbach's alpha is high in this case, one would hardly assume that taste is largely shared. This shows why Cronbach's alpha is of little use for determining to what extent taste is private or shared.

Agreement Within Pairs of Judges

When Cronbach's alpha and n are known, an intraclass correlation can be computed that reflects how strongly, on average, the ratings of two judges are correlated. Thornhill and Gangestad (1999), in a literature review concerning facial attractiveness, reported that "two raters' judgments typically [correlate] in the range 0.3–0.5" (p. 452). A shared amount of variance between 9% and 25% is certainly not very impressive. Does it indicate that taste is predominantly private? Again, the answer is no, and a fictitious experiment shows why: Imagine that one tested each judge twice. Further, assume that one obtained a typical interjudge agreement of .40 and an average retest reliability of .40. In this case, the agreement between two raters would be as high as the agreement between the average rater and her- or himself. Because it is hard to see how interjudge agreement could exceed retest reliability, the interjudge agreement here—although low in absolute numbers—would strike one as very remarkable. Moreover, such a result would clearly indicate the superior strength of shared taste: All judgment variance that is stable over time could be explained by a shared attractiveness standard.

Why One Should Know About the Relative Importance of Shared and Private Taste

Accordingly, neither the reliability of face scores (which is usually very high) nor interjudge agreement (which is usually moderate to low) can inform as to whether the facial-attractiveness standards of different judges are largely similar or dissimilar. And, therefore, the widespread notion that taste is largely shared is—although not necessarily wrong—at least unwarranted.

It is important to scrutinize this claim for at least four reasons. First, the face strongly contributes to overall physical attractiveness (e.g., Alicke, Smith, & Klotz, 1986; Furnham, Lavancy, & McClelland, 2001), and physical attractiveness matters: Physically attractive children and adults of both sexes are more favorably judged and treated by others than are their less attractive peers (Eagly, Ashmore, Makhijani, & Longo, 1991; Feingold, 1992; Langlois et al., 2000). Economic numbers, which reflect the subjective importance of physical attractiveness in the modern, industrialized world, complement this scientific evidence: On the day I write these lines, the cosmetic company L'Oréal is almost 3 times as valuable as the world largest car manufacturer (General Motors). And the world literature suggests that the importance of physical attractiveness is neither a new phenomenon nor one that is restricted to the West. For example, 2,500 years ago, Rachel's beauty captured Jacob's heart, and to marry her he served 14 years (Gen. 29, Revised Standard Version); and around the year 1000, the Japanese court lady Murasaki Shikibu praised the beauty of Prince Genji, the hero of the world's first novel. Today, several non-Western cultures place substantially more weight on good looks when it comes to marriage than Western cultures do (Buss, 1989). Apparently, attractiveness moves people all over the world and has done so throughout history.

The second reason pertains to scientific research strategy: Unwarranted claims about taste being largely shared bring with them the danger of overlooking an important field of investigation. If private taste substantially contributes to judgments of facial attractiveness, researchers should explain it, not simply dismiss it as random noise. It is interesting to note that biologists who work on mate choice have begun to regard interindividual preference differences as an important field of interest (Forstmeier & Birkhead, 2004; Jennions & Petrie, 1997). Research on interindividual preference differences in humans is scarce but promising (Johnston, Hagel, Franklin, Fink, & Grammer, 2001; Little et al., 2002; Little & Perrett, 2002; Perrett et al., 2001).

The third reason concerns the way in which evolutionarily based hypotheses about attractiveness should be empirically corrobo-

---

[2] This can easily be computed using the following definition of Cronbach's alpha:

$$(j^2 M_{cov}) \sum_{var/cov}^{-1},$$

where j is the number of judges, M_{cov} is the mean interjudge covariance, and

$$\sum_{var/cov}$$

is the sum of all elements in the variance–covariance matrix (this computation being based on the assumption that all judges' variances are equal; Cortina, 1993).

rated: Such hypotheses have had a beneficial effect on facial-attractiveness research by providing theoretical guidelines (e.g., Symons, 1995; Thornhill & Gangestad, 1993) and by directing attention to hitherto unexplored phenomena like facial symmetry (e.g., Hume & Montgomerie, 2001; Rhodes, Proffitt, Grady, & Sumich, 1998; Scheib, Gangestad, & Thornhill, 1999) and facial averageness (e.g., Fink, Grammer, & Thornhill, 2001; Langlois & Roggman, 1990). Most evolutionary approaches in this field share the idea that a preference for certain facial properties is an evolved adaptation that guides mate choice toward individuals of high "quality." For example, Thornhill and Gangestad (1993) argued that facial symmetry signals parasite resistance, which may be the ultimate cause for humans preferring symmetrical faces. To be judged as an adaptation, any trait (e.g., preference for symmetrical faces) must be reliable (e.g., Buss, 2003, p. 40; Cosmides & Tooby, 1992, p. 61). This entails that (almost) all members of the species who would benefit from this trait have this trait. Therefore, the claim that some facial preference is an adaptation should be substantiated by showing that this preference holds for (almost) all relevant judges; for example, a positive correlation between facial symmetry and perceived attractiveness should result for almost all relevant judges. However, such an analysis is hardly ever reported. Instead, the relationship between the examined facial property and face scores is reported (e.g., Grammer & Thornhill, 1994). This is sufficient as long as it can be assumed that taste is largely shared and, thus, that most judges' evaluations closely resemble the face scores. In this case, any relationship between facial properties and face scores will also hold for most judges. But if taste has a substantial private component, the assumption would not be granted. As such, researchers should change their method of analysis and report for how many judges any proposed relationship between a facial property and attractiveness judgments was obtained.

The fourth reason is an ethical one: Strong relationships between body self-esteem and global self-esteem (e.g., Mendelson, Mendelson, & Andrews, 2000; Secord & Jourard, 1953) imply that many people severely suffer from the thought that they are physically unattractive. Naturally, the notion that taste is largely shared must aggravate such worries, whereas the idea of beauty being predominantly in the beholder's eye must alleviate them. Because research on facial attractiveness is not confined to an ivory tower but, rather, receives considerable media coverage (at least in Germany and in Great Britain), potentially false claims from the scientific community about the universality of taste may actually do harm.

How to Measure Shared and Private Taste

Generalizability theory is a suitable statistical framework to measure the relative contributions of private and shared taste to attractiveness judgments (for an introduction, see Brennan, 2001; Shavelson & Webb, 1991). A basic idea behind this approach is that a particular psychological measurement is often not of interest in itself; rather, alternative measures would have served the same aim (i.e., if different but equally suitable test items, raters, occasions for observation, or the like had been chosen). Generalizability theory examines how readily one can generalize from the particular measures used to the universe of all measures deemed equally suitable. It does so by estimating variance components whose relative sizes reflect how much of the observed variance is attributable to the measurement objects, on the one hand, and to various facets of the measurement (i.e., items, occasions, etc.) on the other hand.

Given that each judge evaluates all faces more than once, it is possible to estimate the relative impact of private and shared taste from the variance components. The strength of shared taste is reflected in the variance component for faces relative to overall variance. The more judges agree, the larger the fraction of variance attributable to faces. To see why, it is helpful to consider two extremes: First, imagine that all judges in a sample give ratings at random. Consequently, there is, on average, no agreement between two judges (i.e., their ratings are not correlated), which, by definition, indicates the absence of shared taste. In this case, all faces receive similar face scores (note that for all faces, the expected mean across judges is the same). Thus, variance in face scores is low, and the estimated variance component for faces approaches zero. Now imagine the other extreme, that shared taste is maximized: All judges agree completely (i.e., give identical ratings for all faces). In this case, the variance of face scores is much larger and equal to overall variance. To measure shared taste—that is, the variance component for faces—it is sufficient that each judge evaluates each face only once.

The impact of private taste on judgments is reflected in the variance component for the interaction between faces and judges. To see why, imagine that the faces of Peter, Paul, and Mary receive face scores of 3, 4, and 5, respectively. Assume that Anne repeatedly rates these faces with 3, 4, and 7, respectively. One concludes that Anne's preferences are somewhat different from the average in that she regards Mary as more attractive than the average judge does. Thus, an interaction effect between judge and face indicates private taste (Kenny, 1994). To determine this interaction, it is necessary that judges rate the faces repeatedly; otherwise, the interaction cannot be separated from error.

Whereas the variance component for Judge $\times$ Face clearly reflects private taste, it is less clear if the latter is additionally reflected in the variance component for judges. To see why, imagine that Jessica rated the same three faces repeatedly with 1, 2, and 3, respectively. How should one interpret this? There are two possibilities. First, one could assume that the 2-point difference between Jessica's judge score and the average judge score (which is identical to the average face score and to the grand mean) reflects a meaningless difference in scale use. Accordingly, one would argue that Jessica's ratings completely agree with the face scores (i.e., with shared taste). More generally speaking, differences between individual judge scores and the mean judge score should be neglected because these differences have no bearing on the relative impact of private and shared taste. Second, one could assume that Jessica's low judge score reflects the fact that she likes the three faces less than Anne and the average judge do. At first sight, this difference seems unimportant; after all, Jessica, Anne, and the average judge all share the same preference order. Consequently, Jessica and Anne might both prefer Paul over Peter for going out on a date. Nonetheless, the overall rating difference may reflect important differences in behavior. Whereas Anne might actually meet with Paul, Jessica might prefer to read a book instead. In this case, we should assume that Jessica's low judge score reflects some genuine disagreement with the average taste.

Thus, Jessica's judge score being 2 points below average should be seen as indicative of private taste.

Which of the two interpretations is right? Do judge-score differences reflect meaningless differences in scale use, or do they reflect genuine differences in perception? Both views have been argued for (e.g., Cronbach, 1955, for the former; Kenny, 1994, for the latter), and there is no definite answer to this question. Likely, the truth lies somewhere in the middle. Therefore, I present all results from both perspectives.

What are the relative contributions of private and shared taste to judgments of facial attractiveness? There is now a method at hand to address this question properly: Provided that all judges rate all faces repeatedly, it is possible to estimate variance components for faces (representing shared taste), for the Judge × Face interaction (representing private taste), and for judges (representing, depending on perspective, either meaningless differences in scale use or private taste).

To conveniently describe the relative impact of private and shared taste, let me define *bi* (beholder index) as

$$\frac{variance_{\mathrm{pt}}}{(variance_{\mathrm{pt}}+variance_{\mathrm{st}})}$$

where $variance_{\mathrm{pt}}$ denotes the variance that is attributable to private taste, and $variance_{\mathrm{st}}$ denotes the variance attributable to shared taste. Resting on the assumption that judge-score differences are meaningless, bi_1 computes as

$$\frac{varcomp_{\mathrm{J\times F}}}{(varcomp_{\mathrm{J\times F}}+varcomp_{\mathrm{F}})}$$

where $varcomp_{\mathrm{J\times F}}$ denotes the estimated variance component for the Judge × Face interaction, and $varcomp_{\mathrm{F}}$ denotes the estimated variance component for faces. Resting on the assumption that judge-score differences are meaningful, bi_2 computes as

$$\frac{(varcomp_{\mathrm{J\times F}}+varcomp_{\mathrm{J}})}{(varcomp_{\mathrm{J\times F}}+varcomp_{\mathrm{J}}+varcomp_{\mathrm{F}})},$$

where $varcomp_{\mathrm{J}}$ denotes the estimated variance component for judges. Thus, a *bi* of .20 would indicate that 20% of the meaningful variance stable across time arises from private taste, and 80% arises from shared taste.

Estimated variance components, bi_1, and bi_2 are given for all experiments reported here. Although ways for computing standard errors on several ratios of variance components have been worked out (Burdick & Graybill, 1992), this is not true for the ratios that define bi_1 and bi_2; therefore, standard errors are only reported for the estimated variance components.

Overview of Experiments

Three experiments on the relative impact of private and shared taste are reported. Likely, results will somewhat depend on the face sample and the judge sample used. Consequently, the three experiments reported used diverse samples of judges and faces to inform about the effects of sample composition.

Four aspects of sample composition are likely to affect results. First, faces can be more or less homogeneous with respect to features that do not systematically relate to face scores. It seems likely that a more heterogeneous face sample will enhance *bi* simply because there are more differences in the faces on which private taste can bear. To address the influence of facial heterogeneity that is unrelated to face scores, Experiments 1 and 2 used face samples that differ very much in this respect. Whereas the first experiment drew solely on Caucasian faces, the second experiment maximized heterogeneity by using Asian, Black, and Caucasian faces (assuming that no race produces more attractive faces per se than do others).

Second, judges may have more or less similar tastes. The less similar they are, the higher *bi* will be. To address this point, Experiment 2 drew not only on an ethnically diverse face sample but also on a multiethnic judge sample comprising Asian, Black, and Caucasian judges. It can be expected that judges will rate pictures of their own race best; and this should increase the role of private taste, because this type of preference has no bearing on face scores. There are at least two reasons to expect an own-race bias. The first is in-group favoritism (e.g., Mullen, Brown, & Smith, 1992). The second is that a mere-exposure effect holds for facial attractiveness (e.g., Mita, Dermer, & Knight, 1977) and probably extends to new faces that are similar to ones previously seen (Rhodes, Halberstadt, & Brajkovich, 2001). The participants in Experiment 2 very likely had highest exposure frequencies to people of their own race, and this would be expected to contribute to a general own-race bias.

Third, *bi* might depend on the interaction of sex of judge and sex of face. This is suggested by results from Marcus and Miller (2003), who found that agreement about facial attractiveness is especially high when men rate women. Experiment 3 addressed potential sex effects.

Fourth, whether the faces in the sample are of similar attractiveness (i.e., produce low variance in face scores) will presumably have a strong impact on *bi*. Imagine a face sample consisting of Leonardo DiCaprio and George Clooney. Both faces will yield high face scores that will hardly differ; nonetheless, many judges will have a stable preference for one face over the other. Consequently, the variance component for faces will be negligible, but the variance component for the Face × Judge interaction will be moderate to high; consequently, *bi* will be very high. For the opposite reasons, *bi* will become very low if the sample of faces used is extremely heterogeneous in attractiveness. For example, a face sample consisting of movie stars and facially deformed people (Tobiasen, 1987) is likely to "prove" that private taste hardly exists. I investigate the impact of face homogeneity related to face scores in a reanalysis of the three experiments.

Experiment 1

Method

Stimuli. Pictures of 77 Caucasian models (56 women, 21 men) between the ages of 16 and 37 years ($M = 25.5$, $SD = 4.2$) were taken. Stimulus persons were approached at different continuing education sites in Münster, Germany, and told that the pictures would be used in a study on face perception. Overall, it was easy to find willing models. All photographs were taken in separate rooms before a bright, neutral background, using a 35-mm camera, a 70-mm lens, and black-and-white film. Shooting distance was 1 m. Both face halves were equally lit. All models were photographed as they were, except that eyeglasses were removed. Models posed with a neutral facial expression and faced the camera frontally. The men were beardless but were not required to be clean shaven.

The photographs were scanned into a PC at a resolution of 1,200 dots per inch, rotated to show the face in an upright position, and cropped to show the head from the top down to the upper cervical fold. These photographs were then scaled down to a height of 537 pixels.

Judges of facial attractiveness. Thirty-one Caucasian participants (21 women, 10 men), none of whom knew the photographed persons, rated all stimuli for attractiveness. The judges ranged in age from 19 to 46 years ($M = 28.5$, $SD = 8.2$). Fourteen participants were undergraduate psychology students from Westfälische Wilhelms-Universität (Münster, Germany); by participating, they fulfilled research requirements. All other participants were approached in the vicinity of the university and participated without payment.

Procedure. The participants first read short instructions informing them of the experiment's aim and procedure. The aim given was "to better understand the perception of facial attractiveness." The instructions also stressed that there were no right or wrong answers but that the participant's personal evaluations were of interest. Participants were then shown how the computer program presenting the stimuli and recording the ratings functioned. Next, in a first presentation series, each face was shown for 2 s to allow the judges to establish a stable internal standard for scale use. In the second presentation series, which immediately followed the first, each picture was shown for 5 s; after this, a noise mask appeared. Only during this second presentation were faces rated on a 7-point scale ranging from 1 (*not attractive*) to 7 (*very attractive*). Participants were neither provided with nor asked for a definition of "attractive." In both series, pictures were presented in the same order, which was randomly determined for each judge. Participants worked alone or in small groups seated separately in a large computer room.

All judges participated in a second rating session that was scheduled 1 week after the first session. The procedure of this second session was identical to that of the first one. This included the fact that for each participant, a new random presentation order for the faces was determined. Thus, each participant rated all faces twice, with a test–retest interval of 1 week and presentation orders differing between sessions.

Results and Discussion

In line with results cited in the introduction, the reliability of the face scores was very high (Cronbach's αs = .95 for both sessions), whereas the average interjudge agreement was moderate ($r = .40$, first session; $r = .39$, second session). Retest reliability was, on average, .74.

Variance components were estimated from analysis of variance (ANOVA) mean squares (Shavelson & Webb, 1991), treating judges, faces, and time of testing (first vs. second session) as random variables. The results are listed in Table 1. Results for *bi* were as follows: $bi_1 = .49$, and $bi_2 = .57$. Thus, whether one regards interindividual differences in judge scores as meaningful (bi_2) or not (bi_1), private and shared taste roughly equally accounted for the meaningful variance stable over time.

Experiment 2

As discussed above, any results obtained for *bi* presumably depend on the composition of both the stimulus and judge samples. To form an impression of how much *bi* depends on ancillary conditions, Experiment 2 used a heterogeneous sample of judges as well as a face sample that was heterogeneous with respect to facial properties unrelated to face scores. As discussed, both measures can be expected to maximize *bi*.

Method

Stimuli. The sample consisted of the portraits of 20 Asian, 20 Black, and 20 Caucasian models (9 or 10 women and 10 or 11 men in each group). Twenty-eight pictures were of women. The people photographed ranged in age from 18 to 37 years ($M = 23.9$, $SD = 5.0$). The pictures of the Caucasian models were randomly drawn from the picture pool of Experiment 1, with the restriction that half of them depict women. All other participants were approached at Ruhr-Universität Bochum (Bochum, Germany) or at Rice University. All photographs were treated as described for Experiment 1.

Judges of facial attractiveness. Thirty-one participants (12 women, 19 men) who were unacquainted with the photographed persons judged all stimuli for attractiveness. They were approached in Münster at Westfälische Wilhelms-Universität or at different meeting points that are popular among foreigners. Participants ranged in age from 19 to 35 years ($M = 25.4$, $SD = 4.1$). Ten participants were Asian, 11 were Black, and 10 were Caucasian. None of the Asian or Black judges had been born in Germany. On average, they had spent 2 years in Germany at the time of the study. All participants were paid €2.50 (~$2).

Procedure. The procedure was largely the same as described for Experiment 1—that is, all participants rated the faces twice, with a test–retest

Table 1
*Analysis of Variance Mean Squares (MSs), Estimated Variance Components, and Beholder Indices (*bis*) for Experiments 1 and 2*

	Experiment 1				Experiment 2			
Source of variation or *bi*	*df*	MS	Estimated variance component	% of total variance (*SE*)	*df*	MS	Estimated variance component	% of total variance (*SE*)
Face	76	45.0496	0.7207	34.0 (5.7)	59	34.8881	0.5328	26.3 (5.1)
Judge × Face	2204	1.7760	0.6944	32.7 (1.3)	1770	1.7943	0.6755	33.3 (1.5)
Judge	29	46.9003	0.2671	12.6 (3.7)	30	42.7927	0.3086	15.2 (4.6)
Time	1	1.1800	0.0000	0.0 (0.0)	1	0.0000	0.0000	0.0 (0.0)
Time × Judge	29	4.3721	0.0518	2.4 (0.6)	30	4.4077	0.0661	3.3 (0.9)
Time × Face	76	0.4192	0.0011	0.0 (0.1)	59	0.5064	0.0020	0.1 (0.4)
Error	2204	0.3871	0.3871	18.2 (0.6)	1770	0.4434	0.4434	21.9 (0.7)
bi_1			.49				.56	
bi_2			.57				.65	

Note. bi_1 = beholder index under the assumption that judge-score differences are meaningless; bi_2 = beholder index under the assumption that judge-score differences are meaningful.

interval of 1 week. The only procedural difference from Experiment 1 was that all judges completed their task at public meeting points or at home, but they did so under the supervision of an experimenter. Therefore, notebook computers were used.

Results

The choice of racially diverse face and judge samples was based on the assumptions that all three racial groups would be equally attractive but that judges would tend to rate faces of their own race above average. To verify these assumptions, I ran a 3 × 3 mixed-factor ANOVA with faces as the unit of analysis (Hönekopp, Becker, & Oswald, 2006). Thus, race of judge was a within-unit factor, and race of face was a between-units factor. For each face, three values were entered into the analysis—namely, the average ratings of all Asian, Black, and Caucasian judges, respectively. As assumed, no main race-of-face effect was obtained, $F(2, 57) = 0.7$, $p = .48$. Thus, the attempt to use a sample of faces characterized by great heterogeneity in appearance unrelated to face scores was successful. In line with the expectation of an own-race bias, a Race-of-Face × Race-of-Judge interaction occurred, $F(4, 114) = 6.3$, $p < .001$. To be sure that the obtained interaction effect represented in-group favoritism, I computed the average evaluation of all same-race judges and of all other-race judges for each face and entered the results into a paired-sample t test with faces as the unit of analysis. As expected, the evaluations of same-race judges ($M = 3.5$, $SD = 0.9$) proved to be significantly more favorable than the evaluations of other-race judges ($M = 3.3$, $SD = 0.8$), $t(59) = 2.5$, $p = .02$. How the two proposed mechanisms (in-group favoritism and exposure) contributed to this effect remains unclear.

As in Experiment 1, the reliability of the face scores was very high (Cronbach's $\alpha = .93$, first session; Cronbach's $\alpha = .94$, second session). The average interjudge agreement was somewhat lower than before ($r = .31$, first session; $r = .34$, second session), which is consistent with the expected smaller impact of shared taste in Experiment 2. Retest reliability was, on average, .72.

Variance components were estimated as for Experiment 1. A detailed account of the results is given in the right half of Table 1. Results for *bi* were as follows: $bi_1 = .56$, and $bi_2 = .65$.

Discussion

As expected, Experiment 2, using a set of faces and a sample of judges that were both heterogeneous with respect to race, yielded slightly higher values for *bi* than were found in Experiment 1. Both bi_1 and bi_2 rose moderately—each by 7%. Thus, somewhat more than half of the meaningful variance stable over time was attributable to private taste. For the purposes of Experiment 2, it was crucial to find a factor that would enhance the heterogeneity of the appearance of the judged faces while being unrelated to face scores. I can think of no more powerful factor than race here. Therefore, one might safely generalize that it is not of much consequence for the question of to what extent taste is private or shared whether the stimuli are more or less heterogeneous with respect to properties unrelated to face scores.

Experiment 3

As mentioned above, previous findings (Marcus & Miller, 2003) suggest that sex of judge and sex of face might affect the relative impact of private and shared taste. Experiments 1 and 2, with unbalanced numbers of female and male participants, could not properly address this issue. Experiment 3 did so. Moreover, participants' judgment latencies were recorded, because these might have some bearing on the question of whether interindividual differences in judge scores are meaningful. It seems plausible that judges look longer at attractive faces than at unattractive faces. If this is the case, a positive relationship between participants' judge scores and their mean looking times would indicate that judge-score differences are meaningful. In this case, judges giving, on average, low ratings would look for shorter periods of time at the faces than would judges giving, on average, high ratings, which would suggest that the former indeed liked the faces better than the latter. If, on the contrary, judge scores were not correlated with judgment latencies, this would suggest that judge-score differences are meaningless.

Method

Stimuli. The sample consisted of the portraits of 54 Caucasian models (27 women, 27 men). The mean age of the models was 22.1 years ($SD = 2.9$). Prospective models were approached at the Westfälische Wilhelms-Universität Law School. Pictures were taken with a digital camera at a resolution of 1,200,000 pixels. Because some of the portraits showed unnatural colors, all pictures were converted into monochromes. These were then standardized as described for Experiment 1.

Judges of facial attractiveness. One-hundred students from Technische Universität Chemnitz (Chemnitz, Germany; 50 women, 50 men) judged all stimuli for attractiveness. The judges ranged in age from 19 to 42 years ($M = 23.4$, $SD = 3.6$). The larger portion of the sample consisted of undergraduate psychology students; by participating, they fulfilled research requirements. All other participants were paid €3 (~$2.50).

Procedure. The procedure was identical to that of Experiment 1. Thus, all participants rated all faces twice, with a test–retest interval of 1 week. Judgment latencies were taken by the computer program that presented the stimuli and recorded the attractiveness judgments.

Results

Again, the reliability of the face scores was high (Cronbach's αs = .99 for both sessions), whereas the average interjudge agreement was moderate ($r = .43$, first session; $r = .42$, second session). Average retest reliability was .73. Variance components were estimated as for Experiments 1 and 2; Table 2 shows the results for the entire data set and for each of the four subgroups (women rating men, women rating women, etc.). For the entire data set, results for *bi* were as follows: $bi_1 = .44$, and $bi_2 = .57$ (similar to the findings of Experiment 1). When the data set was split into four groups by sex of face and sex of rater, moderate differences in *bi* were found. Overall, private taste was somewhat stronger in men ($bi_1 = .47$, $bi_2 = .60$) than in women ($bi_1 = .39$, $bi_2 = .54$). Given the size of the confidence intervals and the lack of a plausible explanation for this pattern, these differences may be best attributed to random error. The same holds for differences in interjudge agreement, which is reflected in the variance accounted for by faces. In contrast to the suggestion of Marcus and Miller (2003), agreement was not especially high when men rated women. In this condition, faces accounted for 30% of the variance; in the other conditions, this value varied between 28% and 37%.

Overall, judges looked longer at attractive than at unattractive faces; face scores and mean judgment latencies (averaged across

Table 2
*Analysis of Variance Mean Squares (MSs), Estimated Variance Components, and Beholder Indices (*bi*s) for Experiment 3*

Source of variation or *bi*	All data				Data split by sex of face and sex of rater			
	df	MS	Estimated variance component	% of total variance (*SE*)	W rate M	W rate W	M rate W	M rate M
Face	53	157.400	0.7774	32.9 (6.5)	37.2 (10.5)	35.4 (10.0)	29.5 (8.4)	28.1 (8.0)
Judge × Face	5247	1.676	0.6107	25.8 (0.7)	22.3 (1.3)	23.5 (1.3)	28.1 (1.6)	22.6 (1.4)
Judge	99	53.022	0.4389	18.5 (3.0)	19.0 (4.4)	19.9 (4.4)	13.6 (3.6)	22.0 (5.1)
Time	1	56.560	0.0096	0.4 (0.6)	0.0 (0.1)	0.2 (0.4)	0.5 (1.0)	1.3 (2.0)
Time × Judge	99	4.403	0.0731	3.1 (0.5)	3.1 (0.8)	2.2 (0.6)	5.6 (1.3)	4.2 (1.0)
Time × Face	53	0.693	0.0024	0.1 (0.0)	0.1 (0.1)	0.1 (0.1)	0.2 (0.2)	0.1 (0.2)
Error	5247	0.454	0.4544	19.2 (0.3)	18.3 (0.7)	18.6 (0.8)	22.6 (0.9)	21.8 (0.9)
bi_1			.44		.37	.40	.49	.45
bi_2			.57		.53	.55	.59	.61

Note. Right half of the table shows percentages of total variance (with standard errors in parentheses). W = women; M = men; bi_1 = beholder index under the assumption that judge-score differences are meaningless; bi_2 = beholder index under the assumption that judge-score differences are meaningful.

judges and sessions) were correlated, with $r = .73$ ($n = 54$, $p < .001$). Then, do judge scores correlate with judges' mean latencies? To answer this question, I computed four scores for each judge: the judge scores for female and male faces (averaged across both sessions) and the mean latencies for female and male faces (again, averaged across both sessions). Split for sex of judge, the results were as follows: women judging women, $r = .29$ ($p = .022$); women judging men, $r = .29$ ($p = .022$); men judging women, $r = .23$ ($p = .052$); and men judging men, $r = .31$ ($p = .014$; $n = 50$, with one-sided testing in all cases). Thus, in three out of four groups, judges who gave, on average, more favorable ratings looked, on average, significantly longer at the faces. Moreover, the smallest (nonsignificant) correlation did not significantly differ from the highest ($Z = .77$, $p > .40$).

Discussion

Experiment 3 used a face sample and a judge sample similar to those of Experiment 1, and the obtained results were similar. Again, private and shared taste explained attractiveness judgments about equally well. Splitting data by sex of face and sex of judge into four subgroups gave rise to moderate variation in *bi*. Given the lack of a cogent explanation for the emergent pattern, these between-groups differences may be best interpreted as arising from random error.

Marcus and Miller (2003) argued, from an evolutionary standpoint, that agreement about attractiveness should be higher for opposite-sex than for same-sex judgments and that it should be highest for men rating women. In their study, unacquainted women and men met in small groups, took a good look at each of the other group members, and then rated their attractiveness. The authors' hypothesis was supported: Agreement between men rating women (41% variance attributable to stimulus persons) was higher than agreement between women rating women (29%) or men rating men (20%); agreement between women rating men was in-between (31%). The pattern of agreement obtained in Experiment 3 hardly resembles the one obtained by Marcus and Miller ($r = .11$, $n = 4$). Also, Marcus and Miller found much larger agreement differences between the four groups than were found in the present experiment. It seems unlikely that these differences between studies arise from the fact that Marcus and Miller used live ratings whereas the present experiment used photographs, because both methods yield very similar face scores (Howells & Shaw, 1985). The fact that judges saw the whole person in the Marcus and Miller study but saw only photographs of faces in the present experiment might be more important. Attractiveness of face and body are only moderately correlated for women (Thornhill & Grammer, 1999) and men (Hönekopp, Rudolph, Beier, Liebert, & Müller, 2006). Thus, agreement concerning the attractiveness of bodies might be different from agreement concerning faces; the sex-dependent pattern of agreement proposed by Marcus and Miller may hold only for bodies (see also Kerr & Kurtz, 1978, for another study on facial attractiveness with divergent results). Future research should address this issue.

Experiment 3 gathered judgment latencies because these may shed light on the question of whether interindividual differences in judge scores are meaningful. With faces as the unit of analysis, a strong correlation between attractiveness and mean judgment latencies was found, indicating that judges looked longer at attractive than they did at unattractive faces. This finding was paralleled by small but mostly significant correlations between judge scores and judges' mean latencies. Thus, judges who gave, on average, low ratings looked at the faces for a shorter time, indicating that they indeed deemed the faces less attractive than did judges who gave, on average, higher ratings. Therefore, interindividual differences in judge scores at least partly reflect substantial differences in perception and not differences in scale use exclusively. Consequently, bi_1, which completely ignores these differences, underestimates the relative contribution of private taste to attractiveness judgments.

The Influence of Attractiveness Heterogeneity: A Reanalysis of Experiments 1–3

The present experiments, using very different face and rater samples, found private and shared taste to be roughly equally powerful. This indicates that neither the composition of the rater sample nor the homogeneity of the faces with respect to properties that are unrelated to face scores exert a significant influence on *bi*; the same holds for the sex of judges and faces. However, as

discussed, a set of attractiveness-heterogeneous stimuli should decrease *bi*, whereas the opposite should be true for a set that is homogeneous with respect to the stimuli's average attractiveness.

A reanalysis of the three experiments illustrates this phenomenon: For each experiment, the stimuli were split into a homogeneous set—which comprised the faces that received intermediate face scores (2nd and 3rd quartiles)—and a heterogeneous set (faces with face scores in the 1st and 4th quartiles). For each subset, *bi* was computed as before (for brevity, only bi_1 is reported). The results are shown in Figure 1. As can be seen, *bi* strongly depended on the variance in face scores. Averaged across experiments, bi_1 rose to .84 with homogeneous face samples and dropped to .35 with heterogeneous face samples. It is interesting to note that whether the face sample was homogeneous or heterogeneous affected only the variance component for faces. In absolute numbers, this was 0.13 for homogeneous faces, 0.68 for all faces, and 1.26 for heterogeneous faces (all values unweighted averages across the three experiments). The variance component for Judge × Face interactions hardly changed across face sets. In absolute numbers, this was 0.64 for homogeneous faces, 0.66 for all faces, and 0.68 for heterogeneous faces (all values unweighted averages across the three experiments).

The data in Figure 1 suggest that inflating *bi* is easier than deflating it. The reason becomes clear when one remembers that *bi* is defined as

$$\frac{\text{private taste}}{(\text{private taste} + \text{shared taste})}$$

The reanalysis suggests that private taste is not affected by faces' heterogeneity in attractiveness. Because private taste is approximately constant, *bi* hinges exclusively on shared taste. Because the variance in face scores can easily approximate 0 (i.e., all faces look on average equally attractive), *bi* can approximate 1; however, because the variance in face scores cannot approximate ∞, *bi* cannot approximate 0.

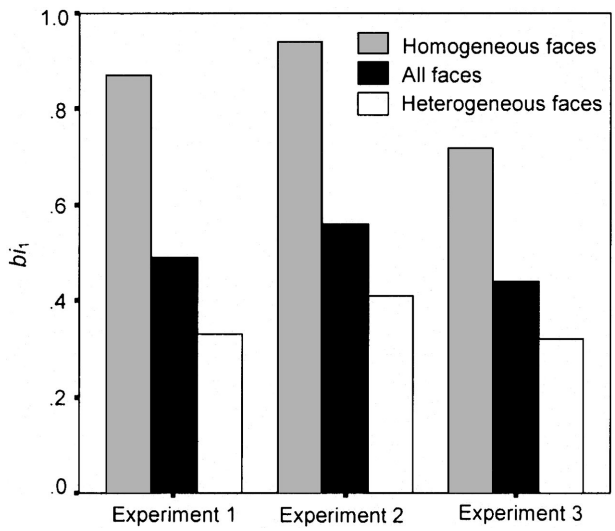

Figure 1. Reanalysis of Experiments 1–3. Beholder index under the assumption that judge-score differences are meaningless (bi_1) was determined for subsamples of faces that provided either homogeneous face scores (2nd and 3rd quartiles) or heterogeneous face scores (1st and 4th quartiles).

This reanalysis shows that *bi* is extremely sensitive to one aspect of sample composition: Faces very similar in attractiveness will yield high values for *bi*, whereas faces that vary widely in attractiveness will yield low values. Does this mean that the quest for a universal answer to the target question is futile? I do not think so. Although different face samples will yield different results, not all types of samples are equally suited to investigate the relative impact of shared and private taste. Studying judgments in a laboratory environment that does not match participants' natural environment(s) easily yields misleading results (e.g., Gigerenzer, Hoffrage, & Kleinbölting, 1991). But what constitutes an adequate natural face sample? First, it should adequately reflect the attractiveness spectrum within the examined age group, and it must neither be restricted to nor exclude extremes. I regard the sampling techniques used in the present study as appropriate in that respect. But how large should the age range from which faces are chosen be? An answer to the sampling problem should consider the function of attractiveness. As far as is known, evolutionary forces have shaped the human ability to regard others as physically (un)attractive because this is an adaptation that improves mate choice by directing one's desires toward those who would make an especially capable mate (e.g., Symons, 1995). Accordingly, attractiveness plays an important role in human dating behavior and mate choice (e.g., Buss & Barnes, 1986; Walster, Aronson, Abrahams, & Rottman, 1966). Consequently, judgments of attractiveness should be studied with a face sample that approximately reflects the natural environment in which mate search occurs. Most people will confine this search to people within a certain age range. Statistical data can provide information about this range: In Germany, in the year 2000, the standard deviation of the age difference in marrying couples was about 5.1 years (age of groom minus age of bride [estimated from aggregated data]; Statistisches Bundesamt, n.d.). To be ecologically valid, any face sample should come close to this age dispersion, as was the case with the samples in Experiments 1 (SD = 4.2 years) and 2 (SD = 5.0 years). Judged against this standard, the age dispersion in Experiment 3 was somewhat low (SD = 2.9 years). However, the fact that Experiment 3 gave rise to findings similar to those of the other studies suggests that this deviation did not matter much.

Although extremely homogeneous or heterogeneous face samples yield extreme values for *bi*, I regard the results of the present experiments—revealing that private and shared taste are about equally important—as a satisfying answer to the target question, because the samples used had acceptable ecological validity.

The Impact of Private Taste on Facial Attractiveness Judgments

It is easy to grasp the implications of shared taste: People tend to agree about the attractiveness of others. But what does it mean that private taste is about as influential as shared taste? A simple calculation may give an impression of the effect of private taste: For each judge, I averaged his or her first and second evaluation of each face, giving a fair account of the "real" impression each face evoked in each judge (the median of the reliability for this measure was .89 in Experiment 1, .85 in Experiment 2, and .86 in Experiment 3). Then, for each judge, I rank-ordered the faces. Thus, each

face received a rank for each judge reflecting the rank of that face in the judge's preference order. Figure 2 shows, for each face, the most and the least favorable evaluations from the 30 (Experiment 1), 31 (Experiment 2), and 50 judges (Experiment 3 [here, only opposite-sex judgments were considered]). As can be seen, a rater group of moderate size is already sufficient to produce extreme judgment differences for nearly all faces.

Conclusion

Although authors on facial attractiveness have often given the impression that taste is largely shared, no data existed to support this claim. Evaluations that are averaged across judges are highly reliable, whereas agreement between single judges is usually much lower. As outlined above, neither type of information is sufficient to answer the question of whether standards of facial attractiveness are primarily shared or primarily private. However, when judges evaluate the same faces twice, as they did in the present study, it is possible to estimate the fractions of overall variance that represent private taste and shared taste.

Three experiments addressing this issue were reported. All three mainly used faces of people between 20 and 30 years of age. Whereas the first and third experiments used Caucasian faces and raters only, the second used a racially diverse sample of faces and raters; the idea behind this procedure was that more heterogeneous samples would allow private taste to exert a greater influence. This was hardly the case, and all three experiments provided similar results—namely, that private and shared taste are about equally important.

The research presented here is not without shortcomings. First, the present experiments were restricted to facial photographs. Although faces are important contributors to overall physical attractiveness (e.g., Alicke et al., 1986; Furnham, Lavancy, & McClelland, 2001), attractiveness of the body matters as well. Furthermore, it is of course possible that the relative contributions of private and shared taste are somewhat different with respect to

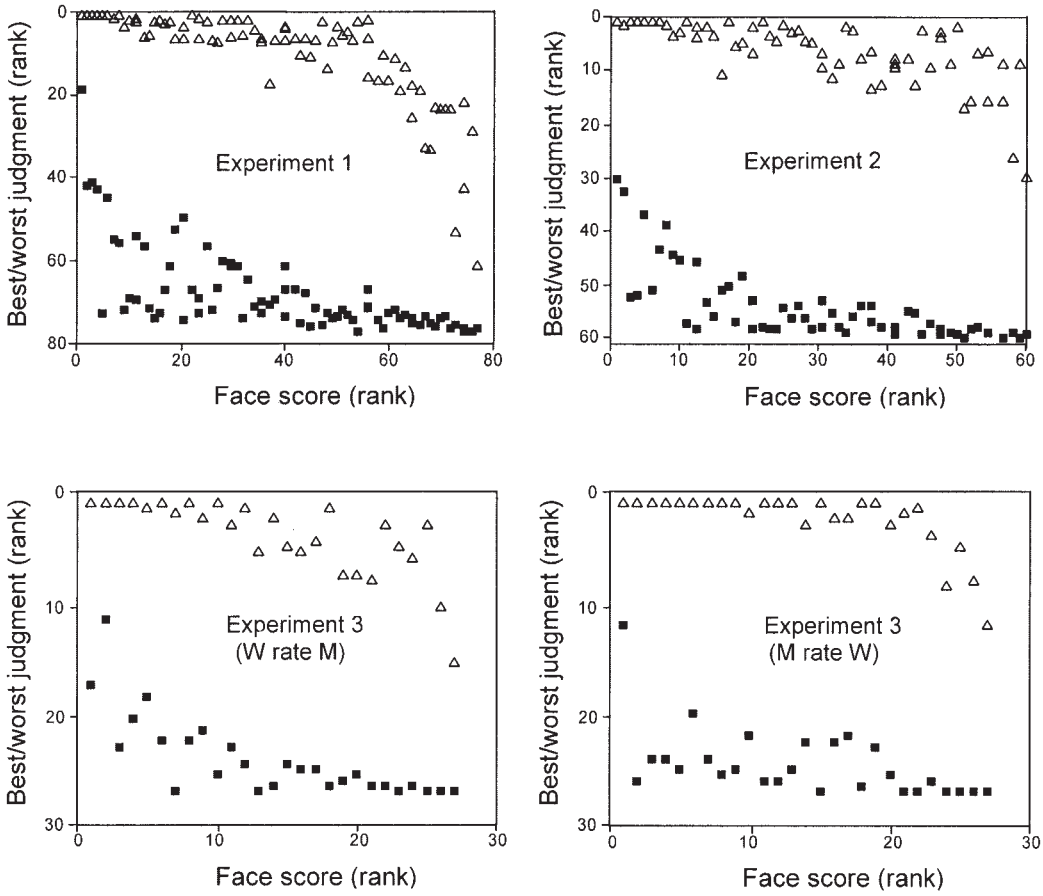

Figure 2. The impact of private taste. For each rater, the rank order of faces on the basis of both evaluations of each face was computed. The figures show each face's most (triangles) and least (squares) favorable evaluations. As an example, take the leftmost triangle and square from Experiment 2: Both pertain to the face with the most favorable face score in Experiment 2. The triangle indicates that there was at least one judge who liked this picture best; the square indicates that this picture obtained the 30th rank in the preference order of the judge who liked this face least. As can be seen, judge samples of only moderate size (30, 31, and 50 judges, respectively) already yield extreme evaluation differences for almost all faces.

overall physical attractiveness. However, typical claims that judges agree about the attractiveness of others have mostly stemmed from studies on facial attractiveness. Nonetheless, future research is clearly needed here. Second, it is regrettable that no confidence intervals on *bi* could be given. However, the fact that all three experiments showed comparable results largely makes amends for this deficit. Finally, a reanalysis of the data presented here showed that any answer to the question of how much of beauty is in the eye of the beholder heavily depends on the attractiveness homogeneity of the faces used. Thus, it is not possible to provide a universal answer to this question. However, because the samples used had acceptable ecological validity, the findings of the present experiments—indicating that private and shared taste are roughly equally important—may be regarded as a valid answer.

What about the prevailing message in the scholarly literature that "standards of beauty are widely shared" (Rhodes, Zebrowitz, et al., 2001, p. 31)? In light of the data presented here, a statement like this is not "wrong," but it is very likely to bring about a wrong notion about facial attractiveness. It is a bit like telling a little girl that a *zoo* is a place where many children eat ice cream and have much fun; in saying as much, one says nothing wrong, but the girl will acquire a queer concept of a zoo. After all, it is not less true to say that standards of beauty are widely private. Because both statements are true, seemingly militating phenomena can peacefully coexist: Some people can make a fortune with their good looks because they appeal to a broad public, *and* friends can endlessly debate about who is attractive and who is not.

Private taste substantially contributes to judgments of facial attractiveness. Some important consequences of this finding have been laid out above: First, researchers should recognize interindividual differences in attractiveness judgments as a potential field of investigation and should not simply overlook them. Some promising work has recently addressed such differences (Little & Perrett, 2002). Future research should also address differences in judge scores. It seems reasonable to assume that people with above average mate-search motivation (e.g., singles and people with a promiscuous sociosexual orientation) will also have above average judge scores. Second, scientists who want to corroborate that certain preferences reflect an adaptation should carefully consider whether face scores or judgments of individual observers are the appropriate unit of analysis. Because of the strong impact of private taste on attractiveness judgments, researchers cannot take for granted that any preference evident in face scores also holds for the large majority of raters. And finally, on days when we dislike our image in the mirror, two thoughts may comfort us: Self-perception of attractiveness is only loosely related to others' perception ($r = .24$, on average; Feingold, 1988), and (almost) all people are good-looking—at least to some.

References

Alicke, M. D., Smith, R. H., & Klotz, M. L. (1986). Judgment of physical attractiveness: The role of faces and bodies. *Personality and Social Psychology Bulletin, 12,* 381–389.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Burdick, R. K., & Graybill, F. A. (1992). *Confidence intervals on variance components.* New York: Marcel Dekker.

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences, 12,* 1–49.

Buss, D. M. (2003). *Evolutionary psychology: The new science of the mind* (2nd ed.). Boston: Allyn & Bacon.

Buss, D. M., & Barnes, M. (1986). Preferences in human mate selection. *Journal of Personality and Social Psychology, 50,* 559–570.

Chen, A. C., German, C., & Zaidel, D. W. (1997). Brain asymmetry and facial attractiveness: Facial beauty is not simply in the eye of the beholder. *Neuropsychologia, 35,* 471–476.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Cosmides, L., & Tooby, J. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York: Oxford University Press.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52,* 177–193.

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,* 61–72.

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C.-H. (1995). "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology, 68,* 261–279.

Dion, K. K. (2002). Cultural perspectives on facial attractiveness. In G. Rhodes & L. A. Zebrowitz (Eds.), *Facial attractiveness: Evolutionary, cognitive, and social perspectives* (pp. 239–260). Westport, CT: Ablex.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but . . .: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110,* 109–128.

Feingold, A. (1988). Matching for attractiveness in romantic partners and same-sex friends: A meta-analysis and theoretical critique. *Psychological Bulletin, 104,* 226–235.

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin, 111,* 304–341.

Fink, B., Grammer, K., & Thornhill, R. (2001). Human (*Homo sapiens*) facial attractiveness in relation to skin texture and color. *Journal of Comparative Psychology, 115,* 92–99.

Forstmeier, W., & Birkhead, T. R. (2004). Repeatability of mate choice in the zebra finch: Consistency within and between females. *Animal Behaviour, 68,* 1017–1028.

Furnham, A., Lavancy, M., & McClelland, A. (2001). Waist to hip ratio and facial attractiveness: A pilot study. *Personality and Individual Differences, 30,* 491–502.

Geldart, S., Maurer, D., & Carney, K. (1999). Effects of eye size on adults' aesthetic ratings of faces and 5-month-olds' looking times. *Perception, 28,* 361–374.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

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

Hönekopp, J., Becker, B. J., & Oswald, F. L. (2006). The meaning and suitability of various effect sizes for structured Rater × Ratee designs. *Psychological Methods, 11,* 72–86.

Hönekopp, J., Rudolph, U., Beier, L., Liebert, A., & Müller, C. (2006). *Physical attractiveness of face and body as indicators of physical fitness.* Manuscript submitted for publication.

Howells, D. J., & Shaw, W. C. (1985). The validity and reliability of

ratings of dental and facial attractiveness for epidemiologic use. *American Journal of Orthodontics, 88,* 402–408.

Hume, D. K., & Montgomerie, R. (2001). Facial attractiveness signals different aspects of "quality" in women and men. *Evolution and Human Behavior, 22,* 93–112.

Jennions, M. D., & Petrie, M. (1997). Variation in mate choice and mating preferences: A review of causes and consequences. *Biological Review, 72,* 283–327.

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,* 251–267.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis.* New York: Guilford Press.

Kerr, N. L., & Kurtz, S. T. (1978). Reliability of "the eye of the beholder": Effects of sex of the beholder and sex of the beheld. *Bulletin of the Psychonomic Society, 12,* 179–181.

Kowner, R. (1996). Facial asymmetry and attractiveness judgment in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 662–675.

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,* 390–423.

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

Little, A. C., Penton-Voak, I. S., Burt, D. M., & Perrett, D. I. (2002). Evolution and individual differences in the perception of attractiveness: How cyclical hormonal changes and self-perceived attractiveness influence female preferences for male faces. In G. Rhodes & L. A. Zebrowitz (Eds.), *Facial attractiveness: Evolutionary, cognitive, and social perspectives* (pp. 59–90). Westport, CT: Ablex.

Little, A. C., & Perrett, D. (2002). Putting beauty back in the eye of the beholder. *Psychologist, 15,* 28–32.

Lucker, G. W., Beane, W. E., & Guire, K. (1981). The idiographic approach to physical attractiveness research. *Journal of Psychology, 107,* 57–67.

Marcus, D. K., & Miller, R. S. (2003). Sex differences in judgments of physical attractiveness: A social relations analysis. *Personality and Social Psychology Bulletin, 29,* 325–335.

McArthur, L. Z., & Apatow, K. (1983–1984). Impressions of baby-faced adults. *Social Cognition, 2,* 315–342.

Mealey, L. Bridgstock, R., & Townsend, G. C. (1999). Symmetry and perceived facial attractiveness: A monozygotic co-twin comparison. *Journal of Personality and Social Psychology, 76,* 151–158.

Mendelson, M. J., Mendelson, B. K., & Andrews, J. (2000). Self-esteem, body esteem, and body-mass in late adolescence: Is a Competence × Importance model needed? *Journal of Applied Developmental Psychology, 21,* 249–266.

Mita, T. H., Dermer, M., & Knight, J. (1977). Reversed facial images and the mere-exposure hypothesis. *Journal of Personality and Social Psychology, 35,* 597–601.

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology, 22,* 103–122.

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994, March 17). Facial shape and judgements of female attractiveness. *Nature, 368,* 239–242.

Perrett, D. I., Penton-Voak, I. S., Little, A. C., Tiddeman, B. P., Burt, D. M., Schmidt, N., et al. (2001). Facial attractiveness judgments reflect learning of parental age characteristics. *Proceedings of the Royal Society London, Series B, 269,* 873–880.

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

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

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

Scheib, J. E., Gangestad, S. W., & Thornhill, R. (1999). Facial attractiveness, symmetry and cues of good genes. *Proceedings of the Royal Society London, Series B, 266,* 1913–1917.

Secord, P. F., & Jourard, S. M. (1953). The appraisal of body-cathexis: Body-cathexis and self. *Journal of Consulting Psychology, 17,* 343–347.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Statistisches Bundesamt. (n.d.). *E 03 Eheschließende 2000 nach Alter der Eheschließenden und dem Altersabstand* [Marriage partners in 2000 according to their age and the age difference between them] [Data file]. Wiesbaden, Germany: Author.

Symons, D. (1995). Beauty is in the adaptations of the beholder: The evolutionary psychology of human female sexual attractiveness. In P. R. Abramson & S. D. Pinkerton (Eds.), *Sexual nature, sexual culture.* Chicago: University of Chicago Press.

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

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

Thornhill, R., & Grammer, K. (1999). The body and face of woman: One ornament that signals quality? *Evolution and Human Behavior, 20,* 105–120.

Tobiasen, J. M. (1987). Social judgments of facial deformity. *Cleft Palate Craniofacial Journal, 24,* 323–327.

Walster, E., Aronson, V., Abrahams, D., & Rottman, L. (1966). Importance of physical attractiveness in dating behavior. *Journal of Personality and Social Psychology, 4,* 508–516.

Received February 4, 2005
Revision received May 25, 2005
Accepted May 31, 2005 ■