An Objective System for Measuring Facial Attractiveness

Mounir Bashour, M.D., C.M., Ph.D.

Toronto, Ontario, Canada

Background: Research over the past 20 years has shown that judgments of facial attractiveness are universal; people from all cultures and backgrounds rank and rate faces for attractiveness the same. As such a model for objectively rating facial attractiveness is theoretically plausible, if designed, it would have many uses, including outcomes analysis in plastic surgery of the face. The authors tested a schematic facial composite/prototype mathematical model (the phi mask created by Dr. Stephen Marquardt) as a method for measuring facial attractiveness in an objective manner.

Methods: Thirty-seven male and 35 female faces of 18- to 30-year-old whites of European extraction were rated, as were 31 composite faces of each sex using both Internet and direct survey judges. The faces were tested against the phi mask model analyzing deviations of facial anthropometric points from corresponding phi mask nodal points using equivalent weightings, and weightings arrived at by way of multiple linear regression.

Results: The deviation from the phi mask significantly correlates with attractiveness, explaining from 25 to 75 percent of the variance in attractiveness judgments, depending on the methodology used.

Conclusions: The phi mask model supports averageness or prototypicality of the face as being the major component of the facial attractiveness gestalt and is a first step in producing an objective system for measuring facial attractiveness. (*Plast. Reconstr. Surg.* 118: 757, 2006.)

Surgeons have as their mission the restoration of function and health to the patient while minimizing morbidity and mortality, and maintaining or improving the final aesthetic outcome. In the past 20 years, the emphasis on aesthetic outcome has moved into the forefront, not just in plastic surgery but in all fields of surgery. As facial plastic surgeons, the most important factor to all our patients is the aesthetic outcome of the operation. This holds true when performing reconstructive and functional facial plastic surgery and not just when performing pure cosmetic surgery. Patients are very concerned about their facial appearance and, as surgeons (in the past few decades especially), we too have made aesthetic outcome a primary concern.

Coinciding with this new emphasis on aesthetic outcome, dramatic drives in cost containment and quality assurance have also been implemented in the health field. Administrators and managers using as their major tools various objective outcome measurement systems have helped bring these changes about. In tandem with them, physicians have found these same tools to be an excellent means of self-monitoring, allowing discrepancies and problems to be quickly picked up and permitting objective surveillance of outcomes as changes in practice and technique are implemented.

Mostly as a result of the fact that cosmetic surgery is an elective procedure not paid for by third-party payers, it has managed to largely escape the outcomes knife and its associated good and bad spinoffs. Surgical aesthetic outcome is currently evaluated by completely subjective methods,[1] and little has been done to quantify these qualitative results in an objective manner.[2–4] Patient and surgeon subjectively decide whether the aesthetic outcome is acceptable or not. The subjective nature of this evaluation makes statistical analysis of surgical outcomes impossible. Surgeons and patients and likely health administration and health payers would all benefit from an objective outcome measurement system.

From the University of Toronto.

Received for publication February 7, 2005; accepted April 11, 2005.

DOI: 10.1097/01.prs.0000207382.60636.1c

Recently, some investigators have begun to talk about the need for outcomes analysis in the aesthetic surgery field.[1] The approach they use is to create outcomes scales, and preliminary data indicate that they may have some utility.[2–4] However, the scales are still based on patient subjective evaluation, and although we agree that this is a very relevant measure, a purely objective outcome scale would be extremely desirable.

In a recent review by Ching et al.[1] of outcomes measurements for aesthetic surgery (facial and otherwise), 53 identifiable instruments were found in the literature extending back to 1961. These instruments were divided into four subtypes: satisfaction assessments (six found), objective assessments (five found), psychological assessments (34 found), and quality-of-life assessments (eight found). For satisfaction assessment, the most commonly used method of all 53 identified instruments is comparison of preoperative and postoperative photographs, usually by a surgeon or an independent observer. Ching et al.[1] feel that this method is limited "because there are no validated and reliable means to quantify results to make meaningful comparisons." For another identified satisfaction assessment instrument, facial halves comparison by Hamra,[5] where two halves of a face, one preoperative and one postoperative, are combined together in a photograph, Ching et al. state that its "evaluation is subjective, without a numerical assessment."

In their review of objective assessments, only five methods were found, which we ourselves have looked at in detail. These included Tapia et al.,[6] who looked at their results of 685 rhytidectomies (face lifts) by analyzing 4110 preoperative and postoperative digitized photographs. They created a scoring system on 12 aspects of facial aging; three surgeons visually and subjectively scored the preoperative photographs (average score, 9.75) and postoperative photographs (average score, 2.84) using this scoring system, and noted an average improvement of 6.91 points. Tapia et al. also looked at two objective measurements, the cervicomental angle (which improves an average of 20 degrees) and lifting of the eyebrows (the medial eyebrow average lift is 0.1275 cm, the central eyebrow average lift is 0.2259 cm, and the lateral eyebrow average lift is 0.2877 cm). Amazingly enough, no correlation between the subjective score and objective measurements was attempted; furthermore, no mention is made of how the change in cervicomental angle relates to the improvement in patient result, and for the eyebrow measurements, all that is said is "we noticed a clear relation of greater lifting of the eyebrow corresponding to more satisfactory overall final results." No data—statistical or otherwise—were provided.

The second objective assessment identified by Ching et al.[1] was in Pitanguy et al.[7] In fact, Pitanguy and his Brazilian colleagues designed an elegant objective system for modeling soft-tissue changes with aging, not an objective system for assessing aesthetic outcomes or one capable of measuring facial attractiveness. The study was conducted using 40 women who had photographs of their face obtained at least 5 years apart in time. These photographs were marked with 26 characteristic points (of interest to us is that 24 of these points were identical by chance to 24 of the 37 points we used in our study). These points were used to calculate various linear distances on the face in each photograph, and the change in the distance was normalized by dividing by the interpupillary distance for that subject (again, this is of interest to us, because although we did not normalize the faces using the interpupillary distance, we did place and resize the phi mask using the interpupillary distance). These normalized changes in anthropometric distances over time were fit by least squares using the second-order polynomial that produced the smallest error. Essentially, their method allows, after measuring and normalizing a photograph of a woman, prediction with a known amount of error, the appearance of that woman at a different age. In fact, their method has been used to create a warping (aging/deaging) program for facial photographs.

The third objective assessment identified was in Yousif et al.[8] Yousif et al. again did not directly create an objective system for assessing aesthetic outcomes, or one capable of measuring facial attractiveness. They looked at a very specific facial feature, the nasolabial fold, using photogrammetry (anthropometric measurements from photographs), noting that with "aging there is anterior, lateral, and inferior displacement of the cheek mass with a resultant deepening of the nasolabial fold, while relationships between the upper lip and the fold itself remain constant."[8] This evidence was used to support the theory that the nasolabial fold is created by loss of support of the cheek mass complex by gravitation and aging.

The fourth objective assessment identified was in Mishima et al.[9] This article essentially announces that the authors have created two software systems that allow them to (1) use a three-dimensional digitizer to automatically identify facial landmarks from a wire frame model of a plaster cast (only the area around the nose is described) and capture their three-dimensional coordinates and (2) allow automatic superimposition of a postoperative wire frame model and calculate the displacement of the three-dimensional coordinates. Again, this article does not directly describe an objective system for assessing aesthetic outcomes or one capable of measuring facial attractiveness. It is of interest in that it looks at displacement of an anthropometric model we describe in this article as a three-dimensional wire model to evaluate the face preoperatively and postoperatively, looking at the change in fit for an objective measure of improvement.

The fifth and final objective assessment identified was in Bhatia et al.[10] This article is similar to that by Mishima et al.[9] in that it announces another three-dimensional facial scanning and measurement system capable of measuring volume changes in the face with a facial operation. Again, the article does not directly describe an objective system for assessing aesthetic outcomes or one capable of measuring facial attractiveness. Furthermore, the procedure was tested not on a real facial surgery patient but on a volunteer whose face was injected with known amounts of saline solution subcutaneously.

To have an objective system, facial attractiveness must be able to be objectively measured. Is it possible to measure facial attractiveness objectively? Research findings from the psychology field not only show that judges strongly agree about facial attractiveness but also indicate that a universal standard of facial attractiveness does in fact likely exist.[11]

Cognitive visual psychology research in the field of facial attractiveness has made numerous profound discoveries in the past 20 years. A common view held by many contemporary scientists involved in the area, however, is that the field would greatly benefit from an objective measurement system for facial attractiveness.[12]

Two recent mathematical modeling systems have been developed for human facial attractiveness. One model we will not discuss directly here is the FacePrints model created by Victor S. Johnston.[13,14] The second model is a facial overlay system or mask variously called the phi, archetypal, golden, or golden ratio mask (Fig. 1). This mask has been claimed as being adaptable to the creation of an objective system for measuring facial attractiveness.[15–17]

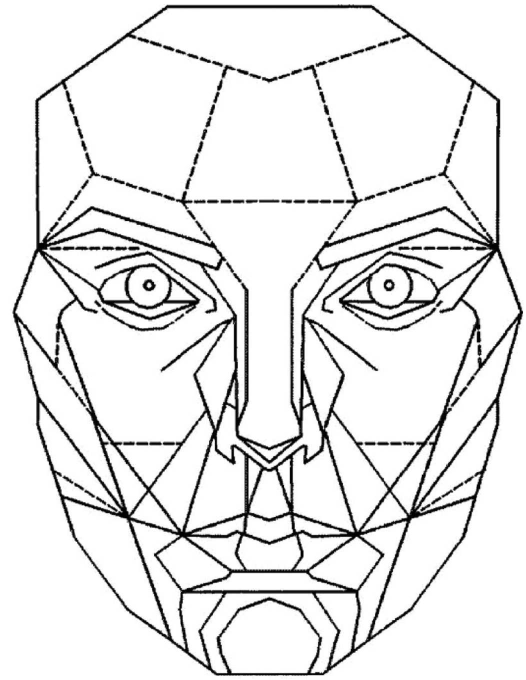


Fig. 1. Phi mask from Marquardt (published in patent: Marquardt, S. R. *Method and Apparatus for Analyzing Facial Configurations and Components*, in U.S. Patent and Trademark Office, 1997).

The phi mask is based on the golden ratio phi (the ratio obtained when a line ABC is cut such that AB/AC = BC/AB), first derived by the ancient Greeks but certainly in use since even more ancient times (e.g., in Egyptian art and architecture) and possessing many fascinating mathematical properties not least of which are the Fibonacci sequence and the logarithmic spiral.[18] This ratio appears almost ubiquitously in nature, including in the basic geometric shapes of the pentagon, decagon, and dodecagon; in the phyllotaxis or leaf arrangement of a vast number of plants and flowers; in the spiral of seashells such as the nautilus; in the human mandible and its growth rate; in the human figure and face; and even in the spiral of DNA.[19,20] The growth rate of an organism has been found to be proportional to the size of the organism and follows the pattern of the logarithmic spiral—this growth pattern can even be observed in the unfolding of the human embryo as it gestates. The phi ratio has been found in the faces and figures of statues dating from the ancient Egyptian and Greek periods and was first introduced into the modern medical literature by Ghyka[21] and Seghers et al.[22] and popularized by Ricketts.[23–27] Marquardt builds on the findings of these researchers by theorizing that an archetypal or prototypical face can be built entirely by using the ratio phi, an idea he supports by the ubiquitous presence of the ratio in nature, particularly in DNA and gestational growth (perfect unperturbed growth = prototypical = perfect adherence to the golden ratio). The mask is based on the use of multiple variously sized pentagram complexes as Marquartdt terms them. These pentagon complexes are in fact variously sized golden or regular decagons, which can be created by superimposing two same sized golden or regular pentagons pointing in opposite directions. You can see how the frontal repose mask fits into the golden decagon (or "pentagram complex" per Marquardt) in Figure 2.

Marquardt uses the primary pentagonal complex to form the basic framework of the mask, using specific lines, line segments, and points to construct the component lines and points of the mask. Various sized secondary decagons or pentagonal complexes related mathematically to the primary complex (the size of the subsidiary pentagon complexes is derived using the formula PC[subject] = PC[reference] $\times$ $(1/\varphi)^{n}$ $\times$ Z, where PC[subject] stands for the main pentagon complex shown in Figure 2, $n = 0$, Z = 1, φ is the golden ratio, and n and Z are variables used) are used to derive the remaining component lines and points of the mask: [n = 6; Z = 1] is used twice for the two iris complexes; [n = 5; Z = 1] is used three times, for the nasal tip complex, the internal lip complex, and the internal nares complex; [n = 5; Z = $\varphi^{1/3}$] is used once as the inner nasal tip halo complex; [n = 5; Z = $2/\varphi$] is used once as the outer nasal tip halo complex; [n = 4, Z = 1] is used four times, for the nasal pentagon complex, the chin button pentagon complex, and the two eye pentagon complexes; [n = 3, Z = 1] is used 14 times, for the nose/mouth complex, the mouth/chin complex, chin inferior border complex, chin complex, right and left sided chin complexes, right and left eye/cheek complexes, right and left eyebrow complexes; right and left cheek complexes, and right and left nose/mouth complexes; [n = 3, Z = $\varphi^{1/3}$] is used twice for the right and left eyebrow/cheek complexes; [n = 2, Z = 1] is used once for the frontal repose smile complex; and [n = 1, Z = 1] is used once for the internal facial pentagon system.

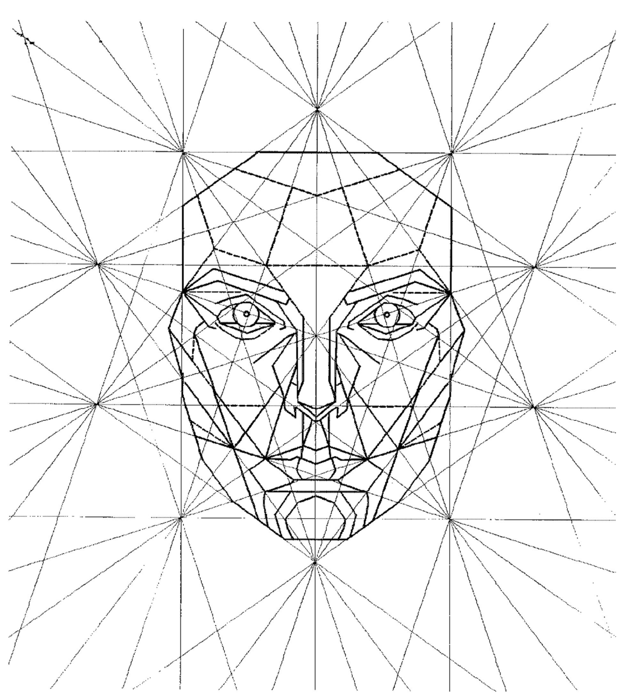


Fig. 2. The phi mask framed by the reference pentagonal complex.

It is of note that Marquardt derived the placement of the particular component lines and points of the mask by applying the pentagon complexes to the faces of female subjects (specifically, to cut-out magazine pictures of various female models). Further details of the mathematical derivation and properties of phi mask may be found in its patent documents.[15,16]

This mathematical model, in addition to two U.S. patents,[15,16] has received tremendous publicity and media attention[17] and has become implanted in the minds of the general public, but to our knowledge it has never been scientifically tested. Are the claims made of it by its creator in the media in fact true? If they are, then this is a potentially powerful tool worthy of scientific attention. If the claims are false, we feel that, given the amount of public exposure that the model has already had, a note in the public record indicating the claims are false is important. Finally, if the claims have some validity but are exaggerated, an exact understanding of the model's capabilities and its limitations would be useful to allow researchers to build on and possibly improve or create a new model that does meet the original claims if at all possible. Furthermore, as has been the experience of many researchers, the simple act of testing a model often creates new avenues of research by itself. Thus, the central hypothesis of this article is that the phi mask can be used to create an objective measurement system for facial attractiveness.

If that is indeed the case, an objective quantitative system should be devisable that would have at least the same correlation with measures by various panels of judges (at least an $r > 0.80$ and preferably an $r > 0.90$).[11] To date, the details of such a system or even of its scale have not been published as far as we are aware.

The specific definition of facial attractiveness we use is as follows: *The visual properties of a face that are pleasing to the visual sense of an observer.* This is as opposed to beauty, which we define as *the assemblage of graces or properties pleasing to the eye, the ear, any or all of the senses, the intellect, the aesthetic faculty, and/or the moral sense.*

In fact, what is measured in this and many other studies is a more scientifically precise component of the "beauty" gestalt: *full frontal repose static two-dimensional photographic facial attractiveness.* This can be precisely defined as *the time-static visual properties of a face in a photographic two-dimensional frontal repose image that are pleasing to the visual sense of an observer.*

SUBJECTS AND METHODS

Thirty-seven male subjects and 35 female subjects aged 18 to 30 years (mean age for male subjects, 22.8 ± 3.27 years; mean age for female subjects, 21.2 ± 2.92 years) were recruited from the student body at the University of Toronto. All subjects were of white European extraction. The subjects were financially compensated for their time and the research was approved by the Ethics Board of the University of Toronto.

The subjects were digitally photographed using a Kodak DCS-560 Camera (Eastman-Kodak, Rochester, N.Y.) [a digital Canon EOS-1N SLR camera (Canon, Tokyo, Japan) that produces very-high-resolution, uncompressed, 18-MB pictures]. Subjects were without makeup or adornments (e.g., earrings), and male subjects were clean shaven. All individuals wore hair off the forehead, head position was standardized to that prescribed by Dr. Marquardt,[15–17,28] and subjects were told to adopt a neutral facial expression with closed mouth while sitting for the photographs. Photographs were taken using a standardized photogrammetry technique[29,30] with standardized lighting conditions against a common background using the same distances for the faces from the focal plane of the 100-mm lens (chosen to produce the least lens distortion in face photography).[29] All face images were kept at their respective relative sizes and not normalized. All images were finally cropped to reduce visibility of hair, ears, and neck, as only the face itself was of relevance in this experiment[31–33] (Figs. 3 and 4).

For analysis purposes and to allow comparison to previous studies, a series of 16 two-face composites (Av2), eight four-face composites (Av4), four eight-face composites (Av8), two 16-face composites (Av16), and one 32-face composite (Av32) were also created from 32 of the original face pictures for each sex using the same technique previously described by other researchers[31,34–38] (Figs. 5 and 6). In total, there were 68 male faces and 66 female faces evaluated by raters in the next stage (31 for each sex being the above-described composites).

In creating the composites, the shape of the major facial features of each face was defined by manually marking 224 predefined feature points (e.g., left corner of the mouth) for each digital face image.[31,39–41] Points were allocated to capture the distinctive shape of individual facial features while maintaining an equivalent spacing on the left and right sides of the face.

All the faces (including composites) were rated using a program custom designed for multiple face rating written using Visual Basic Script in Microsoft Access (Microsoft Corp., Redmond, Wash.). The program is capable of being used either directly on a desktop or notebook system or by means of the Internet while being based on a server.

The faces were rated both directly by students at the University of Toronto and by patients at Lasik M.D. in Montreal in one arm of the rating process (the survey arm) and by random Internet users in the other arm of the rating process (the Internet arm). In the survey arm, the mean age was 25.8 ± 10.8 years, with an age range of 10 to 52 years, and consisted of 25 male and 25 female judges. In the Internet arm, the mean age was 21.6 ± 9.8 years, with an age range of 10 to 52 years, and also consisted of 25 male and 25 female judges. Combining the two arms of the study, we get a combined mean age of 23.7 ± 10.3 years.

The rating scale used ranged from 1 to 10 and was classified as follows: 10, extremely attractive; 9, very attractive; 8, attractive; 7, mildly attractive; 6, neutral plus; 5, neutral minus; 4, mildly unattractive; 3, unattractive; 2, very unattractive; and 1, extremely unattractive. The means of the ratings for each face were calculated for each of the Internet and survey arms and for the combination of the two arms and was termed an attractiveness quotient, thus giving us three measures: survey

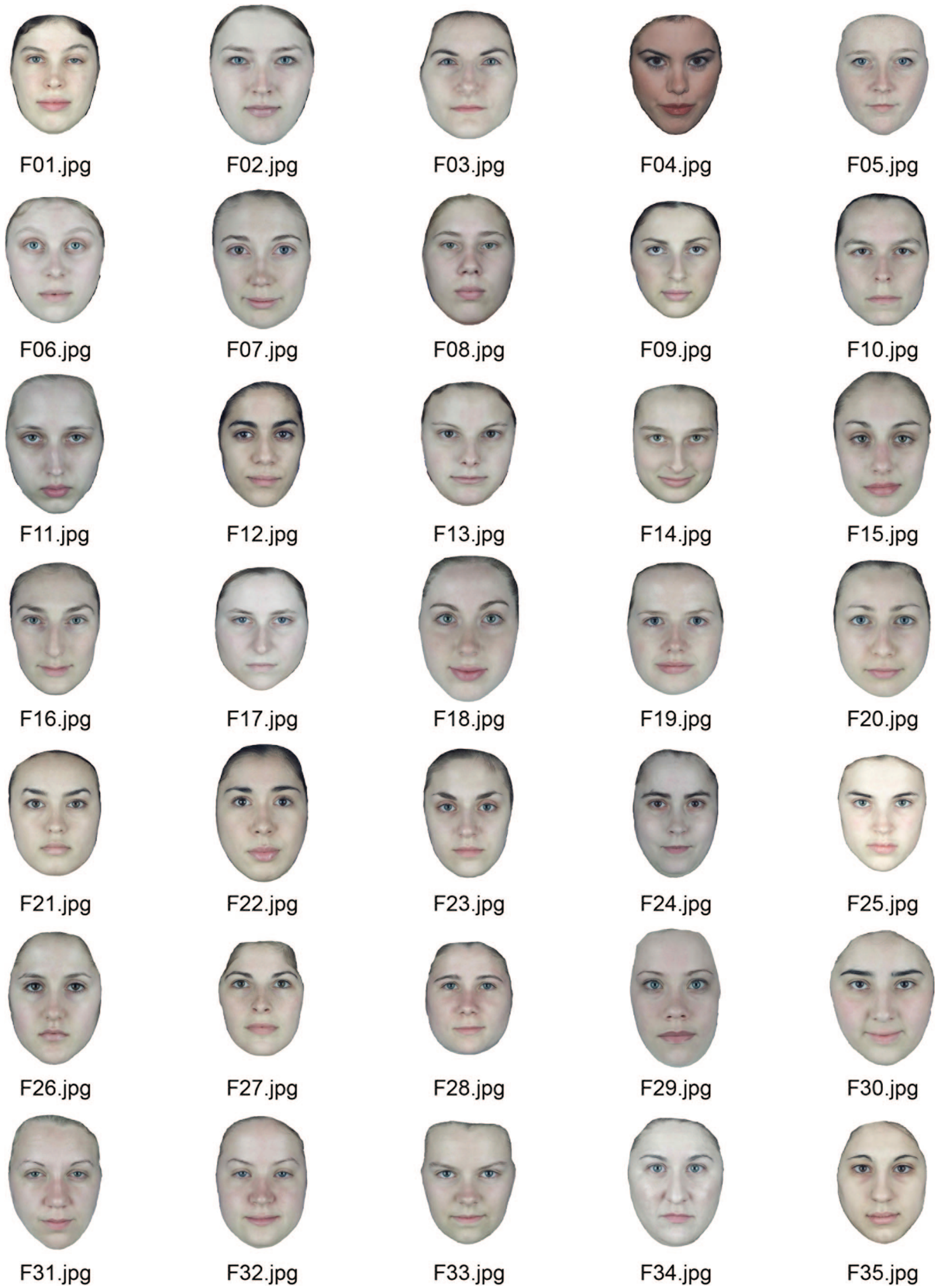


Fig. 3. Female faces normalized and cropped.

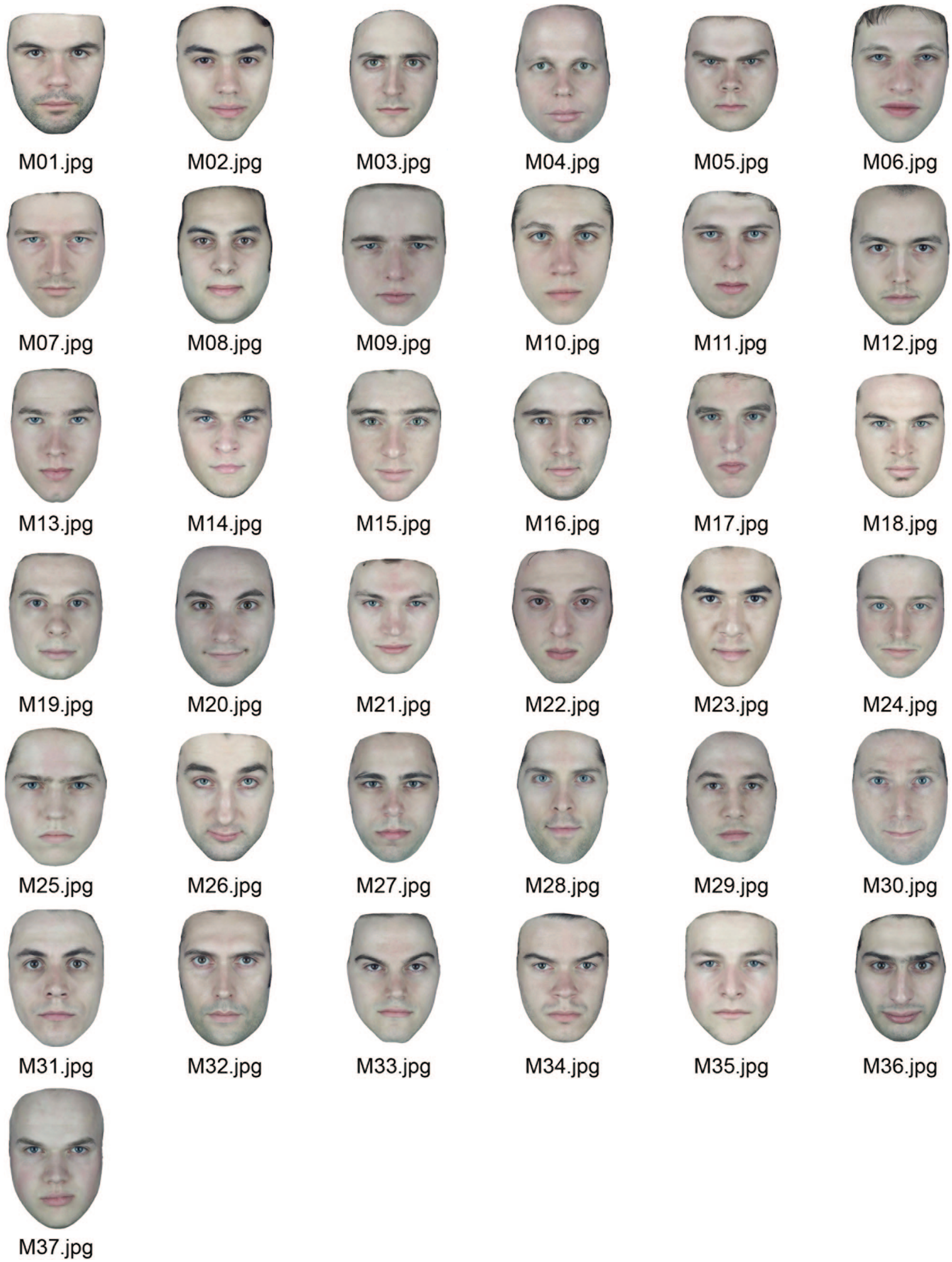


Fig. 4. Male faces normalized and cropped.

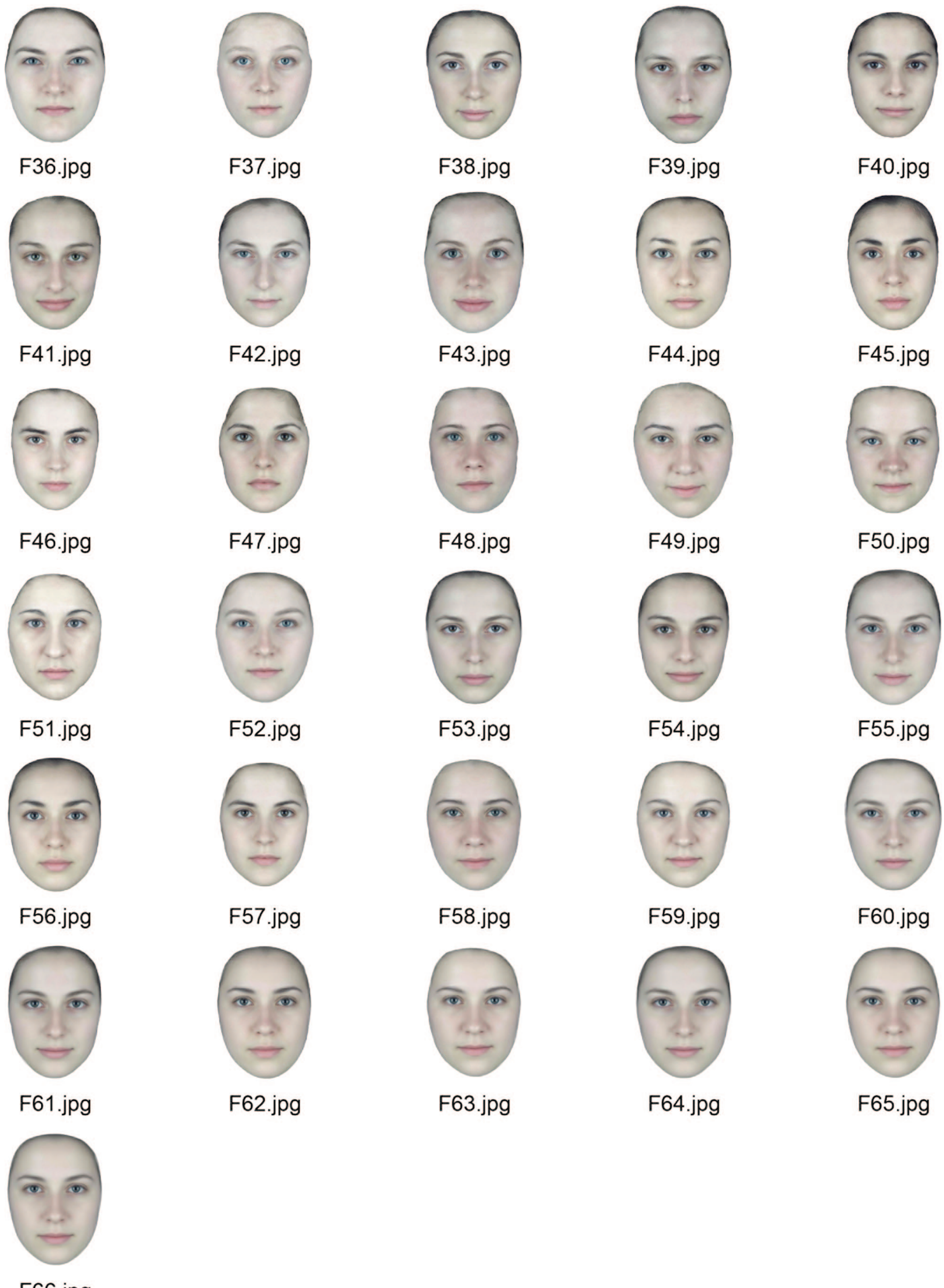


Fig. 5. Female composite faces.

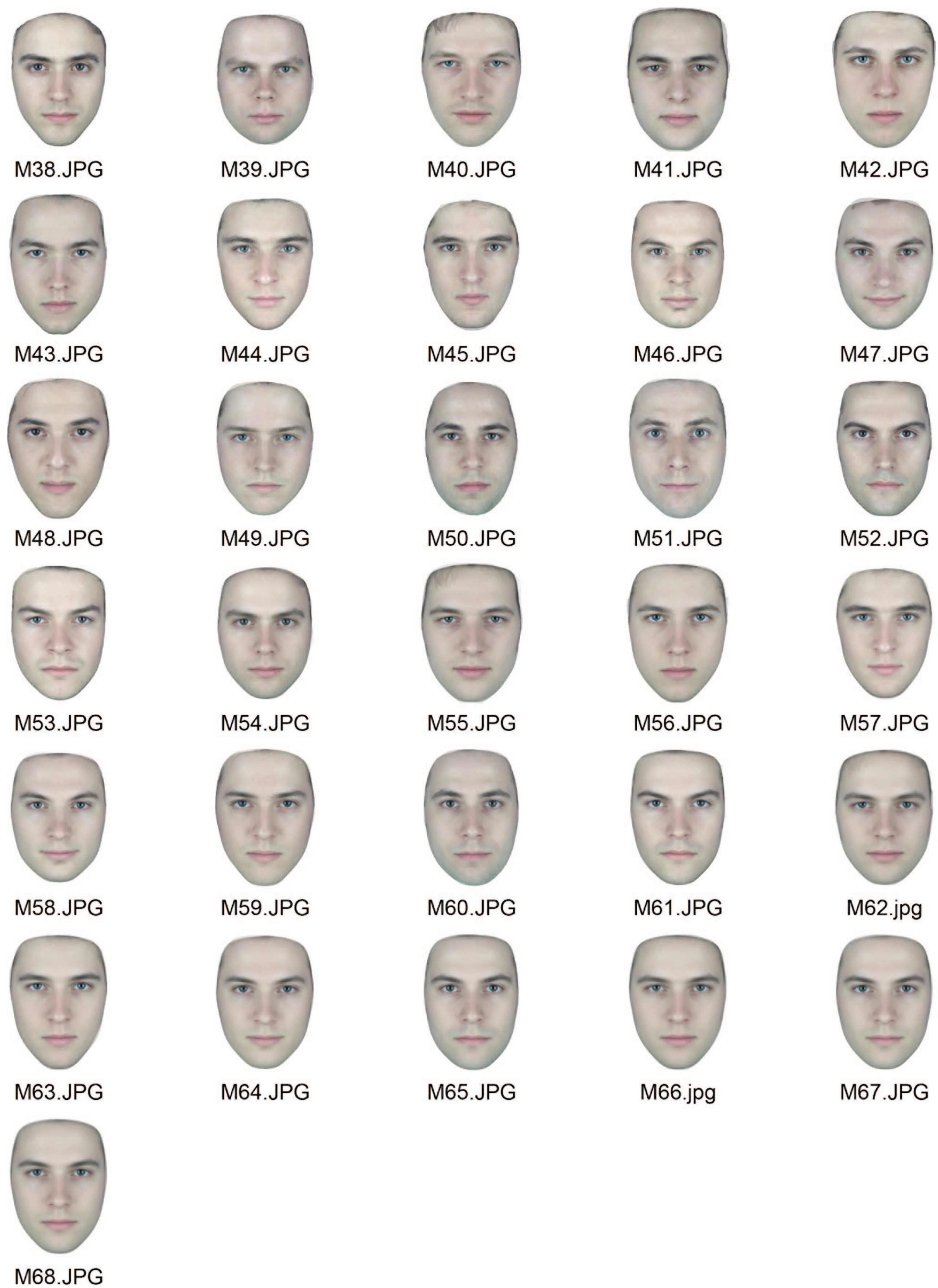


Fig. 6. Male composite faces.

attractiveness quotient, Internet attractiveness quotient, and combined attractiveness quotient.

Using this phi mask model as a template for "attractiveness," we created a quantitative system by measuring the numerical divergence of the real anthropometric landmarks from their equivalent mask nodal points. Two different methods of weighting the landmarks/nodal points as more or less important for attractiveness are modeled: equivalent weightings and weightings arrived at by way of multiple linear regression.

Each of the faces had the phi mask applied to it using Adobe Photoshop 8.0 (considered the standard for photograph manipulation) (Adobe Systems, San Jose, Calif.) and in consensus with the general body of the facial attractiveness literature,[33,36–38,40,42–46] the mask was sized and placed using only the interpupillary distance as the reference line (Fig. 7). Because of the standardized photogrammetry technique used in taking the photographs, we were able to have all the faces imaged as their real size on the computer monitor and were able to measure distances in metric units either by hand or using computer software. We elected to use Image Pro Plus 4.5 (considered the standard for scientific image analysis) to objectively and reproducibly measure the deviation of the mask focal points from their equivalent points on the image of the face. Thirty-seven nodal points were selected that were present on the mask and that we were able to accurately and reproducibly identify on the image of the face as anthropometric landmarks (Fig. 8). The deviation of mask to face nodal points was measured in centimeters, and a total deviation termed the mask deviation score was used as a measure of how far a face deviated from the phi mask (Fig. 9). Our working hypothesis was that we expected to find a significant negative correlation between this mask

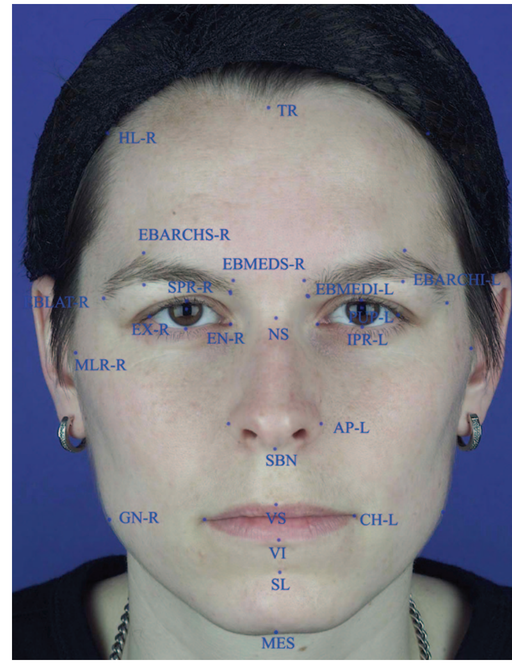


Fig. 8. Measurement focal points illustrated on F10.

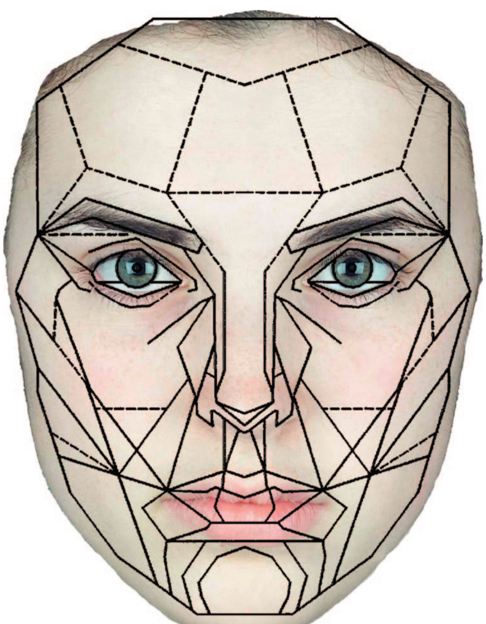


Fig. 7. An example of mask sitting on F25.

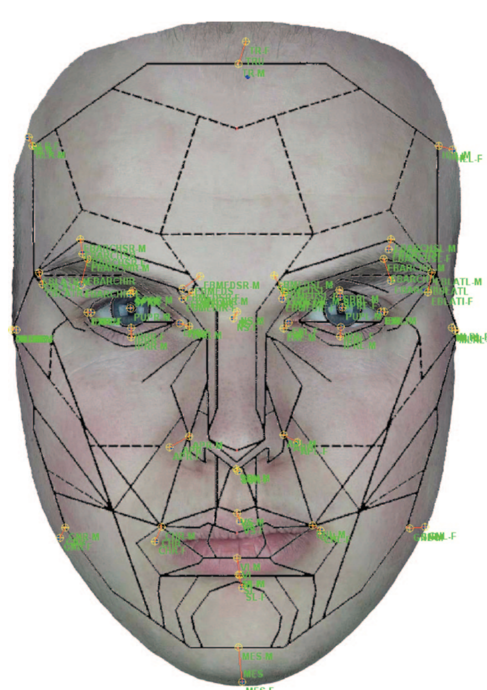


Fig. 9. An example of measurement system on M37.

deviation score and the attractiveness quotient found by judges.

This direct correlation of mask deviation score versus attractiveness quotient is the most basic system of analysis, and it essentially assigns equal weightings to deviations from each of the anthropometric landmarks. Multiple linear regression for 35 nodal point deviations (excluding the two pupil nodal points because they are the fixation points for the mask and have zero deviation) was also undertaken as a statistical method of weighting the nodal points.

RESULTS

Reliability of the attractiveness ratings was assessed for the male and female sets of faces using Cronbach's coefficient alpha, which was 0.97 for male faces, 0.96 for female faces, and 0.98 when both sexes were combined for the survey arm of the judges. The reliability of the male and female raters was equal at 0.96 for the survey arm. For the Internet arm, Cronbach's coefficient alpha was 0.91 for male faces, 0.96 for female faces, and 0.95 when both sexes were combined. The reliability of the male and female raters was equal at 0.90 and 0.95, respectively, for the Internet arm.

Attractiveness Quotient: Internet Arm versus Survey Arm

Mean attractiveness quotients ± SD for the Internet and survey arms were 4.05 ± 1.08 and 4.76 ± 1.27, respectively, for both sex faces; 4.30 ± 1.22 and 4.80 ± 1.36 for female faces; and 3.80 ± 0.86 and 4.73 ± 1.20 for male faces. The t test for the attractiveness quotient means obtained by Internet and survey arms yielded a value of $t = 15.39$ ($p < 0.0001$) for both sexes. For male sex, $t = 9.59$, and for female sex, $t = 13.79$ ($p < 0.0001$ for both). However, linear regression (Fig. 10) showed $r = 0.91$ ($F = 626.56$, $p < 0.0001$) for both sexes and $r = 0.95$ ($F = 631$, $p < 0.0001$) for female and $r = 0.91$ ($F = 302$, $p < 0.0001$) male faces.

Individual versus Composite Faces

The means for the faces were submitted to separate analyses of variance, with composite level as a repeated measures factor. The analysis of variance comparing images of individual male faces with Av2, Av4, Av8, Av16, and Av32 composite images revealed a significant effect of the number of faces: $F(5,67) = 17.78$, 25.28, and 24.61 ($p < 0.0001$) for the Internet, survey, and combined

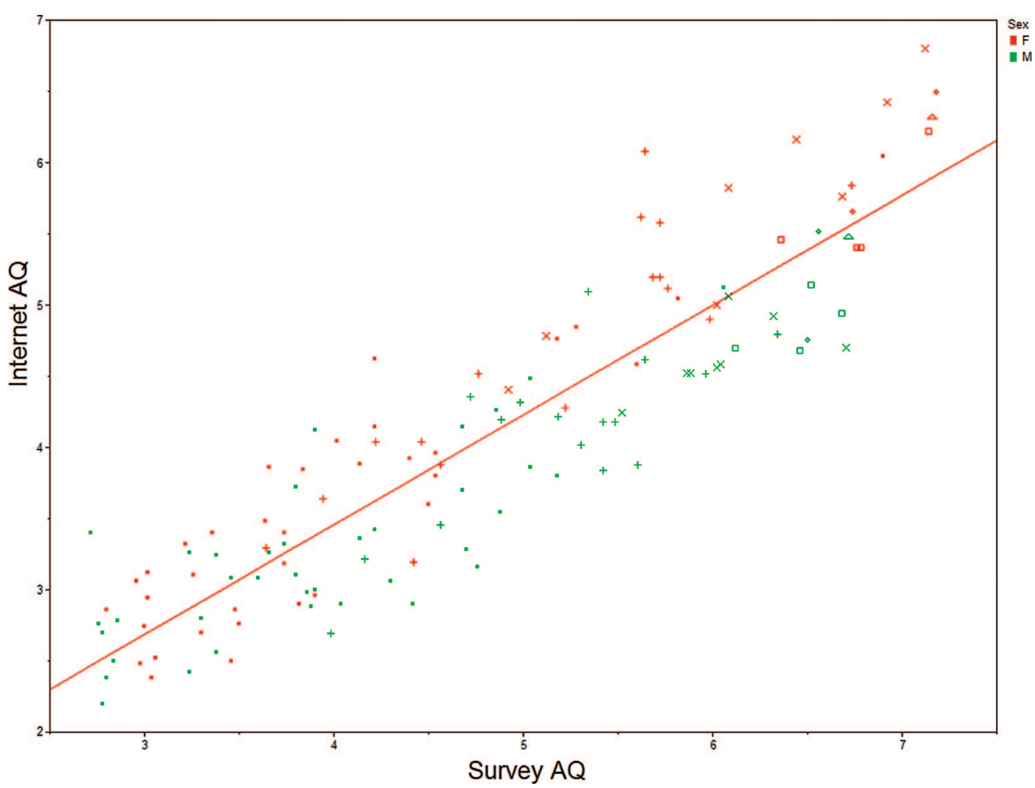


Fig. 10. Linear regression line of Internet attractiveness quotient versus survey attractiveness quotient. As can be seen from the tight scatter of the points around the line of regression, there is a tight linear correlation between the two survey methods. Number of faces: 1 (■), 2 (+), 4 (×), 8 (□), 16 (◇), 32 (Δ).

arms, respectively. Planned comparisons by two-sample t test with separate variance estimates showed that the Av32, Av16, Av8, Av4, and Av2 composites were all rated significantly more attractive than their corresponding individual male faces at $p < 0.0001$ (Bonferroni corrected significance level for multiple comparisons is 0.006).

The analysis of variance comparing images of individual female faces with Av2, Av4, Av8, Av16, and Av32 composite images revealed a significant effect of the number of faces: $F(5,65) = 15.20$, 19.41, and 18.03 ($p < 0.0001$) for the Internet, survey, and combined arms, respectively. Planned comparisons by two-sample t test with separate variance estimates showed that the Av32, Av16, Av8, Av4, and Av2 composites were all rated significantly more attractive than their corresponding individual female faces at $p < 0.0001$ (Bonferroni corrected significance level for multiple comparisons is 0.007).

The analysis of variance comparing images of the faces of both sexes combined with Av2, Av4, Av8, Av16, and Av32 composite images (Fig. 11) revealed a significant effect of the number of faces $F(5133) = 16.34$, 45.36, and 21.63 ($p < 0.0001$) for the Internet, survey, and combined arms, respectively. Planned comparisons by two-sample t test with separate variance estimates showed that the Av32, Av16, Av8, Av4, and Av2 composites were all rated significantly more attractive than their corresponding individual faces at $p < 0.0001$ (Bonferroni corrected significance level for multiple comparisons is 0.007).

Mask Deviation Score versus Attractiveness Quotient

Analysis of variance of mask deviation score with sex and level of composites as factors showed a significant effect for both the number of faces in the composite ($F = 7.98$, $p < 0.0001$) and for sex ($F = 13.84$, $p < 0.001$) (Fig. 12) (Bonferroni corrected significance level for multiple comparisons is 0.006).

For faces of both sexes (combined), Pearson's correlations between mask deviation score and survey attractiveness quotient, Internet attractiveness quotient, and combined attractiveness quotient were −0.49, −0.48, and −0.50 (all $p < 0.0001$), respectively. For male faces alone, Pearson's correlations between mask deviation score and survey attractiveness quotient, Internet attractiveness quotient, and combined attractiveness quotient were −0.53, −0.46, and −0.51 (all $p < 0.0001$), respectively. For female faces, Pearson's correlations between mask deviation score and survey attractiveness quotient, Internet attractiveness quotient, and combined attractiveness quotient were −0.51, −0.47, and −0.49 (all $p < 0.0001$), respectively.

Multiple Linear Regression of Nodal Point Deviations versus Attractiveness Quotient

Giving higher weightings to the more significant nodal point deviations and lower weightings to the less significant nodal point deviations could produce a phi mask model with better fit. In an effort to demonstrate this, a multiple linear regression analysis of the nodal point deviations (as independent variables) against the average facial attractiveness score obtained for each face was conducted for all arms of the survey (Internet, survey, and combined attractiveness quotient are the dependant variables).

For phi mask placed on faces of both sexes:

For Internet attractiveness quotient,
$R = 0.74$,
$R^2 = 0.55$, adjusted
$R^2 = 0.39$,
$F = 2.44$
($p < 0.0001$).

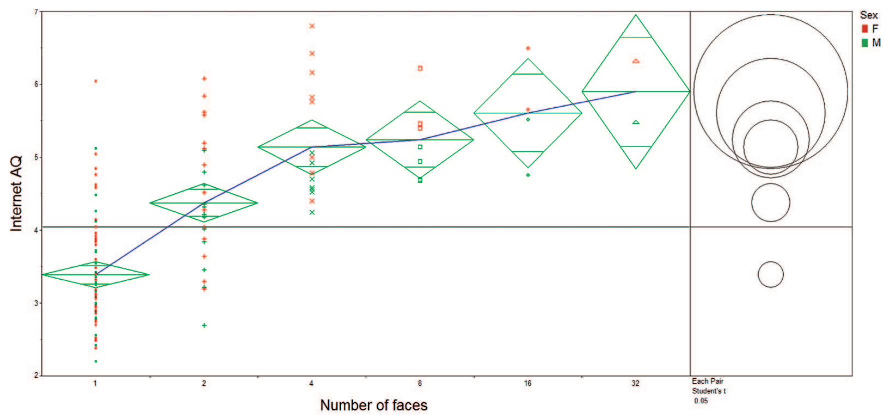
For survey attractiveness quotient,
$R = 0.73$,
$R^2 = 0.53$, adjusted
$R^2 = 0.36$,
$F = 3.28$
($p < 0.0001$).

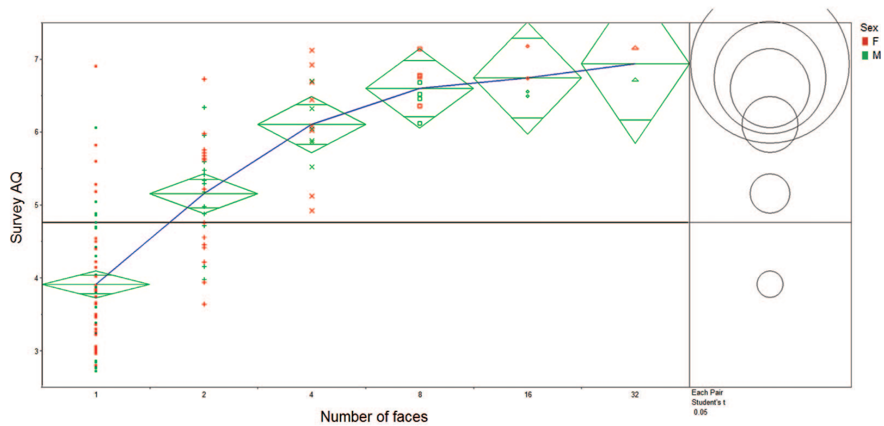
For combined attractiveness quotient,
$R = 0.74$,
$R^2 = 0.55$, adjusted
$R^2 = 0.38$,
$F = 2.76$
($p < 0.0001$).

When placed on female faces alone:

For Internet attractiveness quotient,
$R = 0.93$,
$R^2 = 0.87$, adjusted
$R^2 = 0.71$,
$F = 5.59$
($p < 0.0001$).

For survey attractiveness quotient,
$R = 0.93$,
$R^2 = 0.87$, adjusted
$R^2 = 0.72$,
$F = 5.75$
($p < 0.0001$).





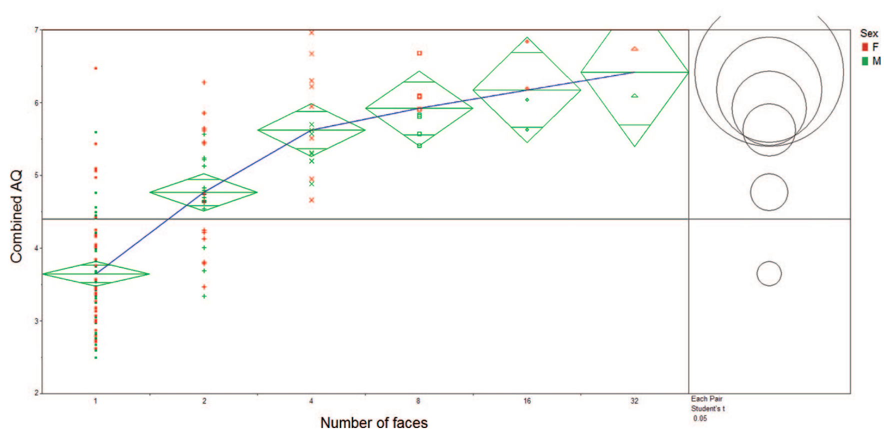


Fig. 11. Visual comparison of mean attractiveness quotient by number of faces (Internet, survey, and combined arms). The illustrations are visual comparisons of group means and show the alignment of comparison circles with the confidence intervals of their respective group means for the *t* test comparison. Each means diamond illustrates a sample mean and 95 percent confidence interval, The line across each diamond represents the group mean. The vertical span of each diamond represents the 95 percent confidence interval for each group. Overlap marks are drawn at [($\sqrt{}$2CI)/2] above and below the group mean. Overlap marks in one diamond that are closer to the mean of another diamond than that diamond's overlap marks indicate that those two groups are not different at the 95 percent confidence level. Each pair of group means can be compared visually by examining how the comparison circles intersect. The outside angle of intersection indicates whether group means are significantly different. Circles for means that are significantly different either do not intersect or intersect slightly so that the outside angle of intersection is less than 90 degrees. If the circles intersect by an angle of more than 90 degrees or if they are nested, the means are not significantly different. Number of faces: 1 (■), 2 (+), 4 (×), 8 (□), 16 (◇), 32 (Δ).

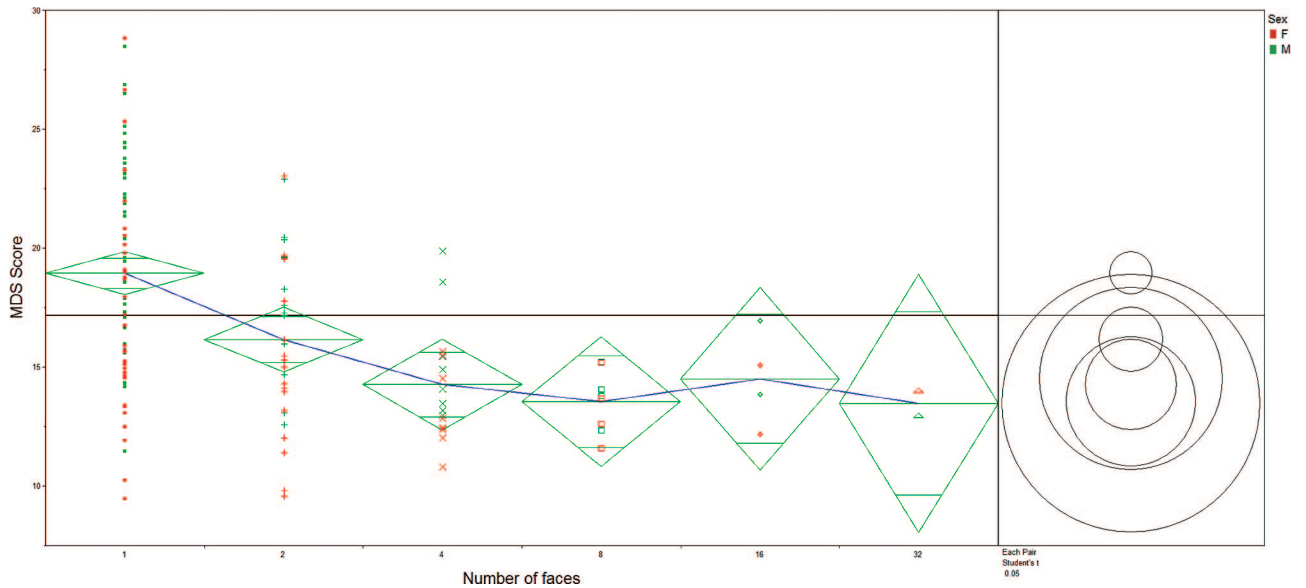


Fig. 12. Visual comparison of mean mask deviation score by number of faces. For male faces, mask deviation score had a mean ± SD of 18.39 ± 4.17 (range, 11.47 to 28.46), and for female faces, mask deviation score had a mean ± SD of 15.96 ± 4.15 (range, 9.46 to 28.81). The *t* test for the means showed a significant difference between male mask deviation score scores and female mask deviation scores, with $t = 3.38$ and $p < 0.001$. Number of faces: 1 (■), 2 (+), 4 (×), 8 (□), 16 (◇), 32 (Δ).

For combined attractiveness quotient,
$R = 0.94$,
$R^2 = 0.88$, adjusted
$R^2 = 0.74$,
$F = 6.35$
($p < 0.0001$).

When placed on male faces alone:

For Internet attractiveness quotient,
$R = 0.78$,
$R^2 = 0.60$, adjusted
$R^2 = 0.16$,
$F = 1.38$
($p = 0.1825$).

For survey attractiveness quotient,
$R = 0.79$,
$R^2 = 0.63$, adjusted
$R^2 = 0.22$,
$F = 1.54$
($p = 1.087$).

For combined attractiveness quotient,
$R = 0.79$,
$R^2 = 0.62$, adjusted
$R^2 = 0.21$,
$F = 1.50$
($p = 0.1265$ (Fig. 13).

DISCUSSION

This study is the first to compare direct survey rating of facial attractiveness and Internet-based rating of facial attractiveness. The only other published Internet research into attractiveness only looked at female body attractiveness from line drawings.[47] Although the agreement ($r = 0.91$) between the two rating arms is as high or higher than that found between most facial attractiveness studies (which range from $r = 0.80$ to $r = 0.95$),[11] allowing us to combine both arms, some interesting points do arise. They are as follows: Internet judges rate male faces much more harshly than female faces and judge all faces more harshly than their counterparts in the direct survey arm. This is probably because Internet judges feel more anonymous,[48] not as directly involved in the process, and not as responsible for the outcome, compared with judges directly asked by the investigators for their participation.[48–50] Also, they may have been more bothered by the cutoff aspect of the face, when they were probably expecting normal full faces as are commonly seen in online attractiveness Internet sites such as www.hotornot.com. The results of the Internet arm are still reliable ($r = 0.91$), and this is an important finding that will allow researchers in the future to increase their efficiency by harnessing the power of the Internet.[51,52] Our findings that attractiveness increases with the number of component faces in a composite are not new and have been published previously.[36,37,45,46,53] Our results are perhaps even stronger than those published previously because all the composites including the two-face compos-

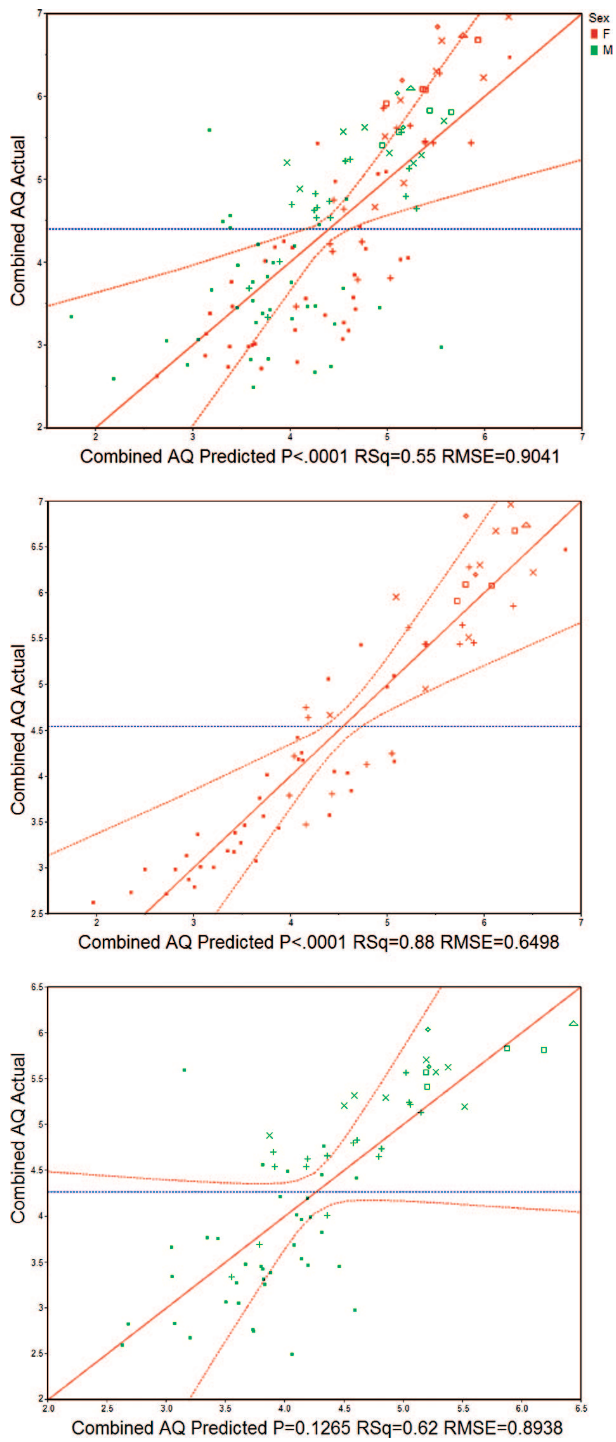


Fig. 13. Plots of multiple linear regression predicted versus actual combined attractiveness quotient for both sexes (*above*), female (*center*), and male sex (*below*). These plots show how the data fit the model predicted by the multiple linear regression. In the plots, the distance from each point to the line of fit is the error or residual for that point; the distance from each point to the horizontal line is what the error would be if effects in the model were removed. Thus, strength of the effect is shown by how strongly the line of fit is suspended away from the horizontal by the points; 95 percent confidence curves are on the graph so it can be seen a glance whether an effect is significant. In each plot, if the 95 percent confidence curves cross the horizontal reference line, the effect is significant; if the curves do not cross, it is not significant (at the 5 percent level). Number of faces: 1 (■), 2 (+), 4 (×), 8 (□), 16 (◇), 32 (Δ).

ites (Av2) were more attractive than almost any of their component faces. Previous studies have shown that the attractiveness of these composites is not attributable solely to their morphed textures (which give smoother complexions and soft-focus appearance)[37,54] or to their increased symmetry or conveyed youthfulness.[37] Studies have also found that these composite faces as they increase in the number of component faces become less variable in judged attractiveness ratings and look increasingly alike by visual observation (these findings are mirrored in our study). Thus, the mathematical averaging procedure of morphing faces results in more typical and less unusual faces. It is hypothesized that these "averaged" or typical faces are preferred in attractiveness by both adults and infants alike because they are perceived as prototypes of the face—they are more facelike. Both evolutionary psychology theory and cognitive psychology theory support the notion that prototypical faces should be viewed as more attractive. Evolutionary theory states that the preference is innate because of evolutionary sexual and natural selection pressure; cognitive theory states that the preference is acquired very early in infancy through learned exposure to category exemplars, in this case, faces.[37] These findings are robust and have been supported in multiple cross-cultural studies showing strong agreement, which is the hallmark of biologically based preferences.

Direct correlation of mask deviation score to attractiveness quotient assigning equal weightings to deviations from each of the anthropometric landmarks shows that the phi mask does indeed work to give an objective measure of facial attractiveness at the most rudimentary level of weighting and explains more than 25 percent of the variance in attractiveness. Studies have shown that certain features are more important in influencing attractiveness ratings than others.[55] As such, we expect that deviations from more important featural anthropometric points should have more of an influence on attractiveness than deviations from other less important points.

Multiple linear regression is in essence a statistical method of weighting the nodal points. From our results, our contention that weighting the nodal points would allow us to achieve a closer correlation to judged attractiveness appears to be true. As expected, the correlation to female faces is much higher than to male faces, in fact it is in the same range as correlations obtained between judges. Of note are the adjusted R^2 values for the multiple linear regressions. An inherent problem with multiple linear regression is that as you add variables, you increase the correlation, often artificially. The adjusted R^2 gives a sense of a more accurate correlation when this is taken into account. However, even the adjusted R^2 for female faces is still 0.74. Interestingly, looking at effect tests that show the variables that have the most significance on the multiple linear regression, the variables that were most consistently found significant were SPR (the position of the upper lid) and EBARCHS (the position of the eyebrow arch superiorly), followed by GN (essentially the width of the jaw), and AP (essentially the width of the nose). This is consistent with other studies showing the eyes, jaw, and nose to be the most significant features in attractiveness.[33]

The much stronger correlation of the mask to female faces over male faces suggests that Dr. Marquardt is mistaken in his belief that the phi mask as designed is ideal for both male and female faces.[15,16] Marquardt's phi mask was essentially derived from studying and averaging multiple images of very attractive female models, mostly from the covers of fashion magazines.[28] The use of the golden ratio as a method to give mathematical integrity to the pursuit of a system for objectifying facial attractiveness had been first suggested by Seghers in 1964[22] and gained popularity because of the work of Ricketts afterward.[23–27] We feel that the phi mask is essentially a schematic of a high-component-number composite consisting of highly attractive female faces that has been given mathematical credence by fitting the golden ratio to it. Is the mathematical association of the golden ratio necessary for the schematic mask to correlate to attractive ratings for faces? Most likely it is not, and we believe any schematic derived from a high component composite would work just as well but would not have the precise reproducibility that only a mathematically derived model can have.

CONCLUSIONS

Our observational and experimental research has led us to several conclusions regarding the ability to measure facial attractiveness objectively. The general conclusion is that indeed it seems possible that at some point in time researchers will be able to devise a highly accurate method for measuring facial attractiveness objectively. The phi mask used in the most rudimentary way (with equal weightings being given to deviations from anthropometric landmarks) yields a statistically significant correlation to measures of facial attractiveness obtained from the current standard measure "truth by consensus." Although the basic analysis using equivalent weightings shows a significant statistical correlation, it is not very high, at most explaining only 25 percent of the variance in facial attractiveness.

This result in itself brings us to some other conclusions, the first being that the phi mask method relies on the "attractiveness is averageness" or the "attractiveness is prototypicality" hypothesis. The facial attractiveness literature shows us that although the averageness hypothesis is extremely strongly supported by many studies, averageness is most likely not the only cue that humans use to determine facial attractiveness. Different studies have shown that other cues, including symmetry, youthfulness or neoteny, and sexual dimorphism, have an impact on human judgment of facial attractiveness.

The phi mask model in its current state partially factors in symmetry because the model is itself perfectly symmetrical but as tested does not directly take into account featural fluctuating asymmetry. This could be possibly achieved by adding a weighting factor to instances of featural fluctuating asymmetry. For instance, if it is determined that asymmetry in the eyes is especially important (let us say, twice as important as any other featural asymmetry measure), we could multiply the nodal displacement measures for the eyes by the fluctuating asymmetry measure for that feature and finally by the importance weighting (two in our example for eyes).

The phi mask model does not take into account the age of the face in any way except for the fact that the prototypical model was created using faces in the young adult range (18 to 30 years old), and we could hypothesize that some of the deviation away from the mask of a particular face could be related to age changes, but we could not know whether that was the case for any individual face. Even though this has no bearing on the results of our own research because the faces we used for testing are all within the same age range of the faces that went into creating the model, we still feel that for a model to achieve its purpose it must in some way factor in age of the face. This can be accomplished by creating prototypical models for various age ranges (e.g., 30 to 40, 40 to 50, 50 to 60, 60 to 70) and/or multiplying the result by an age factor (e.g., 0.8 for 30 to 40, 0.7 for 40 to 50, 0.5 for 50 to 60, 0.4 for 60 to 70). There is strong evidence as well that the age cue affects male and female faces differently, with women aging beyond age 25 resulting in a far more precipitous drop in attractiveness than for similarly aged men. This sex difference may also be considered as partially belonging to the following section.

Sexual dimorphism has been shown to have a major impact on judgment of facial attractiveness, with hyperfeminine features and hypermasculine features both yielding increased facial attractiveness scores in their respective sex faces (the masculinity effect has also been shown to be correlated to the day of the menstrual cycle of female judges). This sexual cue is also not taken into account by the mask in its present form. Again, as for the aging face, separate prototypical models for each sex could be created. In fact, Dr. Marquardt has since created a male variant of the phi mask. Dr. Marquardt notes that the differences in the male variant mask in the repose frontal view are: "1] prominent supra-orbital (brow) ridges (frontal bossing) resulting in deep set appearing eyes; 2] flatter and narrower eyebrows; 3] slightly narrower eyes; 4] eyes less "wide open" (eye lids slightly closed); 5] slightly longer and/or wider nose; 6] slightly thinner lips (especially upper lip); 7] square/angled and or larger jaws."[17] Clearly, Dr. Marquardt has departed from the original claim that he made in his patents that the phi mask in its original state could be used to measure facial attractiveness in both sexes.

Other cues have also been shown to have an effect on facial attractiveness as well (e.g., skin quality and body mass index[56]), and it is theoretically possible to provide weightings for these in constructing a facial attractiveness measurement model. For example, an equation such as the one below may be constructed to arrive at a final number for objective attractiveness: Attractiveness quotient = Weighted deviation from sex prototypical age face × Sex featural fluctuating asymmetry weightings × Sex-specific age factor × Sex skin quality factor × Sex body mass index factor.

The multiple linear regressions for the 35 nodal points allowed the phi mask to explain 70 percent of the variance in male facial attractiveness and almost 90 percent of the value in female facial attractiveness. These values are in the same ballpark as correlations obtained between judges; therefore, our contention that weighting the nodal points would allow us to achieve a closer correlation to judged attractiveness not only appears to be true, but we seem to have devised an objective quantitative system that has "at least the same correlation with measures by various panels of judges (at least an $r > 0.80$ and preferably an $r > 0.90$)."[11]

Thus, we can add to our list of general conclusions that the phi mask fits female faces more closely than male faces; that weighting the deviation of nodal points improves the performance of the phi mask model; and that the phi mask model can potentially explain perhaps as much as 80 to 90 percent of the variance in facial attractiveness when tweaked appropriately in the future. Finally, the phi mask model supports averageness or prototypicality of the face as being the major component of the facial attractiveness gestalt and is a first step in producing an objective system for measuring facial attractiveness.

Mounir Bashour, M.D., C.M., Ph.D.
4175 Sainte Catherine Street West
Montreal, Quebec H3Z 3C9, Canada
bashour@email.com

REFERENCES

1. Ching, S., Thoma, A., McCabe, R. E., and Antony, M. M. Measuring outcomes in aesthetic surgery: A comprehensive review of the literature. *Plast. Reconstr. Surg.* 111: 469, 2003.
2. Alsarraf, R., Larrabee, W. F., Jr., Anderson, S., et al. Measuring cosmetic facial plastic surgery outcomes: A pilot study. *Arch. Facial Plast. Surg.* 3: 198, 2001.
3. Alsarraf, R., and Larrabee, W. F., Jr. Outcomes research in facial plastic surgery. *Arch. Facial Plast. Surg.* 3: 7, 2001.
4. Alsarraf, R. Outcomes research in facial plastic surgery: A review and new directions. *Aesthetic Plast. Surg.* 24: 192, 2000.
5. Hamra, S. T. The role of orbital fat preservation in facial aesthetic surgery: A new concept. *Clin. Plast. Surg.* 23: 17, 1996.
6. Tapia, A., Etxeberria, E., Blanch, A., et al. A review of 685 rhytidectomies: A new method of analysis based on digitally processed photographs with computer-processed data. *Plast. Reconstr. Surg.* 104: 1800, 1999.
7. Pitanguy, I., Pamplona, D., Weber, H. I., et al. Numerical modeling of facial aging. *Plast. Reconstr. Surg.* 102: 200, 1998.
8. Yousif, N. J., Gosain, A., Sanger, J. R., et al. The nasolabial fold: A photogrammetric analysis. *Plast. Reconstr. Surg.* 93: 70, 1994.
9. Mishima., K., Sugahara, T., Mori, Y., and Sakuda, M. Application of a new method for anthropometric analysis of the nose. *Plast. Reconstr. Surg.* 98: 637, 1996.
10. Bhatia, G., Vannier, M. W., Smith, K. E., et al. Quantification of facial surface change using a structured light scanner. *Plast. Reconstr. Surg.* 94: 768, 1994.
11. Langlois, J. H., Kalakanis, L., Rubenstein, A. J., et al. Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychol. Bull.* 126: 390, 2000.
12. Patzer, G. *The Physical Attractiveness Phenomenon.* New York: Plenum, 1985.
13. Johnston, V. S., Solomon, C. J., Gibson, S. J., et al. Human facial beauty: Current theories and methodologies. *Arch. Facial Plast. Surg.* 5: 371, 2003.
14. Johnston, V. S. Method and Apparatus for Generating Composites of Human Faces, in U.S. Patent and Trademark Office. USA, 1994.

15. Marquardt, S. R. Method and Apparatus for Analyzing Facial Configurations and Components, in U.S. Patent and Trademark Office. USA, 1997.
16. Marquardt, S. R. Method and Apparatus for Analyzing Facial Configurations and Components, in U.S. Patent and Trademark Office. USA, 1999.
17. Marquardt, S. R. www.beautyanalysis.com. In S. R. Marquardt, 2001.
18. Wolfram Research. Available at: http://mathworld.wolfram.com/GoldenRatio.html, in Wolfram Mathworld, 2005.
19. Thompson, D. A. W. *On Growth and Form.* New York: Dover, 1992.
20. Cook, T. A. *The Curves of Life.* New York: Dover, 1978.
21. Ghyka, M. *The Geometry of Art and Life.* New York: Dover, 1977.
22. Seghers, M., Longacre, J., and deStefano, G. The golden proportion and beauty. *Plast. Reconstr. Surg.* 34: 241, 1964.
23. Ricketts, R. M. Esthetics, environment, and the law of lip relation. *Am. J. Orthod.* 54: 272, 1968.
24. Ricketts, R. M. The golden divider. *J. Clin. Orthod.* 15: 752, 1981.
25. Ricketts, R. M. Perspectives in the clinical application of cephalometrics: The first fifty years. *Angle Orthod.* 51: 115, 1981.
26. Ricketts, R. M. Divine proportion in facial esthetics. *Clin. Plast. Surg.* 9: 401, 1982.
27. Ricketts, R. M. The biologic significance of the divine proportion and Fibonacci series. *Am. J. Orthod.* 81: 351, 1982.
28. Marquardt, S. R. Personal communication with Dr. S. Marquardt, in Bashour, M. (Ed). California, 2000.
29. Farkas, L. G. *Anthropometry of the Head and Face.* New York: Raven, 1994.
30. Farkas, L. G., Bryson, W., and Klotz, J. Is photogrammetry of the face reliable? *Plast. Reconstr. Surg.* 66: 346, 1980.
31. Perrett, D. I., May, K. A., and Yoshikawa, S. Facial shape and judgements of female attractiveness. *Nature* 368: 239, 1994.
32. O'Doherty, J., Winston, J., Critchley, H., et al. Beauty in a smile: The role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia* 41: 147, 2003.
33. Brown, E., and Perrett, D. I. What gives a face its gender? *Perception* 22: 829, 1993.
34. Alley, T. R., and Cunningham, M. R. Averaged faces are attractive, but very attractive faces are not average. *Psychol. Sci.* 2: 123, 1991.
35. Perrett, D. I., Lee, K. J., Penton-Voak, I., et al. Effects of sexual dimorphism on facial attractiveness. *Nature* 394: 884, 1998.
36. Langlois, J. H., and Roggman, L. A. Attractive faces are only average. *Psychol. Sci.* 1: 115, 1990.
37. Langlois, J. H., Roggman, L. A., and Musselman, L. What is average and what is not average about attractive faces? *Psychol. Sci.* 5: 214, 1994.
38. Langlois, J. H., Roggman, L. A., Musselman, L., et al. A picture is worth a thousand words: Reply to "On the difficulty of averaging faces." *Psychol. Sci.* 2: 354, 1991.
39. Benson, P. J., and Perrett, D. I. Synthesising continuous-tone caricatures. *Image Vision Comput.* 9: 123, 1991.
40. Benson, P. J., and Perrett, D. I. Extracting prototypical facial images from exemplars. *Perception* 22: 257, 1993.
41. Rowland, D. A., and Perrett, D. I. Manipulating facial appearance through shape and color. *I.E.E.E. Comput. Graph. Appl.* 15: 70, 1995.
42. Burson, N., Carling, R., and Kramlich, D. *Composites: Computer Generated Portraits.* New York: Beach Tree, 1986.
43. Grammer, K., and Thornhill, R. Human (*Homo sapiens*) facial attractiveness and sexual selection: The role of symmetry and averageness. *J. Comp. Psychol.* 108: 233, 1994.
44. Little, A. C., and Hancock, P. J. The role of masculinity and distinctiveness in judgments of human male facial attractiveness. *Br. J. Psychol.* 93: 451, 2002.
45. Rhodes, G., Sumich, A., and Byatt, G. Are average facial configurations attractive only because of their symmetry? *Psychol. Sci.* 10: 52, 1999.
46. Rhodes, G., and Tremewan, T. Averageness, exaggeration, and facial attractiveness. *Psychol. Sci.* 7: 105, 1996.
47. Krantz, J. H., Ballard, J., and Scher, J. Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behav. Res. Methods Instrum. Comput.* 29: 264, 1997.
48. Hewson, C. M., Laurent, D., and Vogel, C. M. Proper methodologies for psychological and sociological studies conducted via the Internet. *Behav. Res. Methods Instrum. Comput.* 28: 186, 1996.
49. Buchanan, T., and Smith, J. L. Research on the Internet: Validation of a World-Wide Web mediated personality scale. *Behav. Res. Methods Instrum. Comput.* 31: 565, 1999.
50. O'Neil, K. M., and Penrod, S. D. Methodological variables in Web-based research that may affect results: Sample type, monetary incentives, and personal information. *Behav. Res. Methods Instrum. Comput.* 33: 226, 2001.
51. Reips, U. D. The Web Experimental Psychology Lab: Five years of data collection on the Internet. *Behav. Res. Methods Instrum. Comput.* 33: 201, 2001.
52. Buchanan, T. Internet research: Self-monitoring and judgments of attractiveness. *Behav. Res. Methods Instrum. Comput.* 32: 521, 2000.
53. Rhodes, G., and Zebrowitz, L. A. (Eds.). *Advances in Visual Cognition: Facial Attractiveness. Evolutionary, Cognitive, Cultural, and Motivational Perspectives.* Westport, Conn.: Ablex, 2001.
54. Rhodes, G., Harwood, K., Yoshikawa, S., et al. The attractiveness of average faces: Cross cultural evidence and possible biological basis. In G. Rhodes and L. A. Zebrowitz (Eds.), *Advances in Visual Cognition: Facial Attractiveness. Evolutionary, Cognitive, Cultural, and Motivational Perspectives.* Westport, Conn.: Ablex, 2001.
55. Cunningham, M. R. Measuring the physical in physical attractiveness: Quasi-experiments on the sociobiology of female facial beauty. *J. Pers. Soc. Psychol.* 50: 925, 1986.
56. Hume, D. K. *Asymmetry and Facial Attractiveness in Humans.* Kingston, Ontario: Queens University, 2001. P. 212.