Memory & Cognition
1998, 26 (1), 146-160

The perception of face gender: The role of stimulus structure in recognition and classification

ALICE J. O'TOOLE
University of Texas at Dallas, Richardson, Texas

KENNETH A. DEFFENBACHER
University of Nebraska, Omaha, Nebraska

and

DOMINIQUE VALENTIN, KAREN McKEE, DAVID HUFF, and HERVÉ ABDI
University of Texas at Dallas, Richardson, Texas

The perception of face gender was examined in the context of extending "face space" models of human face representations to include the perceptual categories defined by male and female faces. We collected data on the recognizability, gender classifiability (reaction time to classify a face as male/female), attractiveness, and masculinity/femininity of individual male and female faces. Factor analyses applied separately to the data for male and female faces yielded the following results. First, for both male and female faces, the recognizability and gender classifiability of faces were independent—a result inconsistent with the hypothesis that both recognizability and gender classifiability depend on a face's "distance" from the subcategory gender prototype. Instead, caricatured aspects of gender (femininity/masculinity ratings) related to the gender classifiability of the faces. Second, facial attractiveness related inversely to face recognizability for male, but not for female, faces—a result that resolves inconsistencies in previous studies. Third, attractiveness and femininity for female faces were nearly equivalent, but attractiveness and masculinity for male faces were not equivalent. Finally, we applied principal component analysis to the pixel-coded face images with the aim of extracting measures related to the gender classifiability and recognizability of individual faces. We incorporated these model-derived measures into the factor analysis with the human rating and performance measures. This combined analysis indicated that face recognizability is related to the distinctiveness of a face with respect to its gender subcategory prototype. Additionally, the gender classifiability of faces related to at least one caricatured aspect of face gender.

Human faces provide us with a plethora of information that is valuable and necessary for social interaction. When we encounter a face, we can quickly and efficiently decide whether it is one we know. For faces of persons we know, we can often retrieve semantic and identity information about the person. Additionally, from both familiar and unfamiliar faces we can make judgments about the gender, approximate age, and race of the person. The information we use to accomplish these latter judgments has been referred to by Bruce and Young (1986) in their model of face processing as "visually derived semantic" information. The importance of visually derived semantic information for understanding human performance on face processing tasks has become increasingly evident in recent attempts to bridge the gap between the perceptual and memory-based components of the face processing system (Hancock, Burton, & Bruce, 1996; O'Toole, Abdi, Deffenbacher, & Valentin, 1995; O'Toole, Deffenbacher, Valentin, & Abdi, 1994).

In the present study, we concentrate on the categorical dimension of face gender. We begin by reviewing briefly the basic psychological findings supporting a prototype-based, face space conceptualization of human face processing. This abstract representational framework provides a parsimonious account of several well-established psychological findings concerning face typicality (Light, Kayra-Stuart, & Hollander, 1979; Valentine, 1991; Valentine & Bruce, 1986). We then consider the implications of extending this representational framework to include natural, perceptual categories of faces, such as face gender. These categories share the configural base of a face prototype, but differ in the nature of the visually derived semantic information that specifies the subcategorical face configurations.

This work was supported in part by NIMH Grant MH51765-02 to A.J.O. Thanks are due to June Chance and Al Goldstein for providing the faces used in the present experiments and simulations and to James C. Bartlett, Shimon Edelman, Heinrich H. Bülthoff, George Wolford, and two anonymous reviewers for very helpful comments and suggestions on an earlier version of this manuscript. Correspondence should be addressed to A. J. O'Toole, School of Human Development, GR4.1, University of Texas at Dallas, Richardson, TX 75083-0688 (e-mail: otoole@utdallas.edu).

—**Accepted by previous editor, Geoffrey R. Loftus**

Face Spaces, Typicality, and Gender Categories

The internal representation of a facial prototype, average face, or common facial configuration has played a prominent role in many theories of human face processing and has been called variously a "facial prototype" (Valentine & Bruce, 1986), "CONSPEC" (Morton & Johnson, 1991), and a "face schema" (Goldstein & Chance, 1980). The psychological evidence supporting a face prototype comes from the well-established relationships reported between facial ratings and human performance on face processing tasks. For example, it is well known that faces judged to be typical are less accurately recognized than are faces judged to be unusual (Light et al., 1979). This occurs under the assumption that the theoretical face space is more "crowded" close to the prototype, and so typical faces are more confusable with other faces than are distinctive faces. Additionally, faces judged to be typical are classified as faces more quickly than are faces judged to be unusual (Valentine & Bruce, 1986). This result occurs under the assumption that typical faces are closer to the prototype than are unusual faces, and so can be compared to the prototype more quickly than can unusual faces (Valentine & Bruce, 1986).

In addition to the basic findings concerning face typicality, human observer ratings of facial attractiveness have also been shown to vary inversely with face recognizability (Light, Hollander, & Kayra-Stuart, 1981), indicating, by implication, that attractive faces may in some ways be "average." Data from an earlier study by Shepherd and Ellis (1973), however, are not entirely consistent with the results of Light et al. (1981). Shepherd and Ellis examined the effects of attractiveness (high, medium, and low) on recognizability at three delay periods (1, 6, and 35 days). In the short and intermediate delay conditions, they found no effects of attractiveness on recognizability. In the 35-day delay condition, however, they found a U-shaped relationship between attractiveness and recognizability, with very attractive and very unattractive faces better recognized than moderately attractive faces. An important difference between the study of Light et al. and that of Shepherd and Ellis, however, is the gender of faces used as stimuli; Light et al. used only male faces, whereas Shepherd and Ellis used only female faces.

The relationship between perceived attractiveness and computationally defined facial averages has been debated vigorously in a number of recent papers (Alley & Cunningham, 1991; Langlois & Roggman, 1990; Langlois, Roggman, & Mussleman, 1994; Langlois, Roggman, Mussleman, & Acton, 1991; Pittenger, 1991). Langlois and Roggman (1990) found that *composite* faces, created by arithmetically averaging the images of several male or female faces, were judged to be more attractive than almost any single male or female face. On the other hand, Perrett, May, and Yoshikawa (1994) found that composites of faces judged to be "attractive" were themselves judged to be more attractive than composites made of an equal number of faces chosen randomly from the sample. Combined, these data suggest that although the "averageness" or (proto)typicality of a face may relate to its attractiveness, it is not likely to be the only determining factor.

Despite the central importance of the average/prototype face to theories of human face processing, little is known about how face configurations/prototypes specific to subcategories of faces, such as male or female faces, or faces of different races, relate to this theoretical construct. In addition to the common configuration that all faces share, there exist several subcategories of faces with somewhat different configural bases. These include the visually derived semantic subcategories associated with race, gender, and perhaps age.[1] Although faces within these subgroups share the general face configuration (i.e., the relative position of eyes, nose, and mouth), different visually derived semantic subgroups can be distinguished from one another by normative and variational differences in (1) feature-based information, (2) "second-order" configural information (see Rhodes, 1988), or (3) in some combination of both. For example, faces of different races differ in the norm and variability of features like eye color, hair color, and eye shape, and may also differ in norms related to general face shape such as the degree of protrusion of the facial features. Likewise, male and female faces differ normatively in feature-based information such as the size of the nose and prominence of the brow and also in more global facial shape characteristics such as "fleshiness" (Enlow, 1982). We will refer to these normative differences between male and female faces as "stimulus structure differences."

Stimulus Structure Differences Between Male and Female Faces

Psychological studies. Questions concerning the nature of stimulus structure differences between male and female faces can be considered both from a psychological and a computational perspective. From a psychological perspective, in recent years there has been an intense interest in determining the information human observers use to determine the gender of a face (e.g., Brown & Perrett, 1993; Bruce et al., 1993; Bruce & Langton, 1994; Burton, Bruce, & Dench, 1993; Chronicle et al., 1995; Roberts & Bruce, 1988; Yamaguchi, Hirukawa, & Kanazawa, 1995). These researchers have measured or manipulated facial aspects/features potentially relevant for determining the gender of a face and have related these measures or manipulations to human performance in classifying faces by gender. Several approaches have been taken, including (1) relating human gender classification performance to geometrically based "facial features"—that is, 2-D and 3-D[2] distances and ratio measures among facial landmarks (Burton et al., 1993); (2) examining the importance of individual discrete features (e.g., noses) in the gender decision (Brown & Perrett, 1993; Chronicle et al., 1995; Yamaguchi et al., 1995); and (3) varying the mode of presentation of information in faces (e.g., by presenting photographic negatives of faces, inverted faces, and 3-D head data from laser scans, Bruce & Langton, 1994).

Combined, these approaches have indicated the difficulty of reducing gender-relevant features to simple geometrically defined interlandmark facial distances (Burton et al., 1993) and have highlighted the importance of a broad range of shape and image intensity facial cues to human face gender judgments (Bruce & Langton, 1994). The work with discrete features has also indicated a special role for features like the nose (Chronicle et al., 1995), eyebrows and facial outline (Yamaguchi et al., 1995), and jaw (Brown & Perrett, 1993). Additionally, it has been suggested that gender-specific features may not be completely constant across different races of faces (Yamaguchi et al., 1995; see also, O'Toole, Peterson, & Deffenbacher, 1996, who demonstrated an "other-race effect" for classifying faces by gender).

Computational studies. The recent efforts to determine the features used by human observers to classify faces by gender have been complemented by equally intense computational efforts aimed at developing computer models that can classify faces by gender (e.g., Abdi, Valentin, Edelman, & O'Toole, 1995; Cottrell & Fleming, 1990; Golomb, Lawrence, & Sejnowski, 1991; Gray, Lawrence, Golomb, & Sejnowski, 1995; O'Toole, Vetter, Troje, & Bülthoff, 1997). In contrast to the psychological studies, which begin by postulating a priori a set of gender-specific features, most computational studies have applied statistical pattern recognition procedures to relatively raw or unprocessed 2-D image or 3-D shape data about faces. These procedures have been implemented frequently with connectionist networks, but are usually equivalent to standard statistical analyses (frequently, principal component analysis [PCA]) of the raw face image or shape data.

In the present study, we used a PCA model because it has been applied most commonly to faces and has been shown to relate reliably to human recognition performance and typicality ratings of faces (Hancock et al., 1996; O'Toole et al., 1994). The purpose of applying PCA to faces is to derive a set of independent or orthogonal dimensions (principal components, eigenvectors[3]) with which faces can be described efficiently and completely. As such, PCA models can been used to quantify the statistical structure of the information in faces, including aspects of the visually derived semantic structure. Individual faces in this model are represented with "features"—that is, principal components (PCs) or eigenvectors, derived from a *set* of face images. When both male and female faces are included in the set, individual PCs have been shown to capture information useful for determining the gender of a face (O'Toole, Abdi, Deffenbacher, & Valentin, 1993). Additionally, simple face representations based on combinations of the PCs have been shown to support excellent gender classification performance when input to a simple linear classifier network (Abdi et al., 1995; O'Toole et al., 1997).

Despite the rather nontraditional nature of PCs as features, this kind of representation fits easily into the basic conceptual structures posited in face space models. Specifically, PC-based representations are founded on the concept of a multidimensional space and can accommodate a prototype. This representational framework simply supplements abstract psychological theories, which have not generally been specific about the dimensions of the face space, with a set of concrete, quantifiable (analyzable) dimensions derived from a set of faces. More formally, representing faces via their coordinates on these dimensions defines a face space. Thus, PCA provides one possible instantiation of a face space model that yields a set of *stimulus-derived* dimensions (see Hancock et al., 1996; O'Toole et al., 1995). PCA can be thought of, therefore, as a perceptual front-end for more abstract models of face processing (O'Toole et al., 1995).

Extending the concepts of a face space and face prototype to accommodate natural face categories such as gender raises a number of interesting issues concerning the nature of face typicality and its effects on human performance in recognizing and categorizing human faces. The structure of a face space that accommodates visually derived semantic subcategories of faces is substantially different from that resulting from a face space accommodating only a single homogeneous set of faces. For example, imagine individual faces represented by points in an *n*-dimensional face space, with the distances between any two points being a measure of the perceived similarity between the two faces. When applied to a single homogeneous group of faces (e.g., young adult Caucasian male faces), the points are likely to form a single cluster. Applied to both male and female faces, it is likely that two gender-based clusters will result. Accordingly, the simple assumption that the face space close to the average face is "crowded" is not likely to be true when both male and female faces are included in the space. Rather, the face space close to the average male and average female faces should be crowded, with relatively few faces around the overall average face. In this case, the recognizability or confusability of a face should be most related to its distance from the subcategory average. Likewise, if the classification of a face as an exemplar of a gender subcategory involves a comparison to the subcategory prototype, the recognizability of a face should be inversely related to the time required to classify it as male or female.

Alternatively, the addition of subcategorical structures to a face space raises psychological issues concerning the importance of the "contrastive" nature of gender categories in a space. Although the average male or female face may be considered to be the most typical version of these categories, there is some evidence to suggest a special psychological role for subcategory caricatures that express maximally contrastive aspects of categories (Rowland & Perrett, 1995; Yamaguchi et al., 1995). Highly feminine faces are likely to be faces that are most different from male faces, and vice versa. In the simple face space conceptualization, feminine faces might be represented by the points that are farthest from the male subcategory prototype. If these contrastive aspects of the categories of male and female are perceptually important for the categorization task, then we might expect to see a dissociation of face recognizability and gender classifiability, with the latter tied more to the perceived femininity/masculinity of the face.

In the present study, we have undertaken a systematic exploration of human performance in a more realistic face space containing gender categories. As in previous work, in which the structure of human face space representations has been probed by establishing relationships between facial rating data and human performance on individual faces (e.g., Light et al., 1979; Valentine & Bruce, 1986), we began by collecting these data for a large number of individual male and female faces. Because of the well-established and complex interrelationships among these rating and performance measures (Hancock et al., 1996; Light et al., 1979; O'Toole et al., 1994; Valentine & Bruce, 1986; Vokey & Read, 1992), we have applied a factor analysis to describe the *structure* or *pattern of interaction* among a set of human rating and performance variables collected on individual faces. This approach has been taken successfully in several recent papers and has yielded insight into the multidimensional structure of facial ratings and human performance measures (see Hancock et al., 1996; O'Toole et al., 1994; Vokey & Read, 1992). Such structure would not be evident from only pairs of correlated variables.

The variables assessed in the present study consisted of (1) two facial ratings potentially related to the gender appearance of the faces (femininity/masculinity and attractiveness) and (2) the human performance measures of face recognizability and gender classifiability (i.e., reaction time to classify a face by gender). As noted, although attractiveness ratings have been shown to relate to average faces, there are still questions surrounding the biasing role of face gender in these judgments. We included measures of the perceived femininity of female faces and the perceived masculinity of male faces, which we considered (tentatively) as caricatures of the subgroups of female and male faces.[4] Factor analyses of these ratings and performance measures indicated a surprising independence of the recognizability and gender classifiability of faces.

We next compared the consistency of these findings for male and female faces and found important differences in the structure of the rating/performance space as a function of face gender, especially with respect to the attractiveness rating. Finally, we anchored the human measures to stimulus structural properties of the face categories by adding face measures extracted from a PCA of the face images to the factor analysis on human measures. These computational model measures contained information that related reliably to face gender and face recognizability. This combined analysis of the model and human face measures gave insight into the nature of the facial information underlying some of the human rating and performance data.

We have carried out three human experiments, a combined analysis of these experiments, and a computer simulation. We present the experiments first. We then present a factor analysis of the faces using the variables gathered in the experiments. Finally, we present the computational model and incorporate gender-related model measures directly into the factor analysis with the human judgment and performance data.

EXPERIMENT 1
Reaction Time to Classify Faces by Gender

Method

Observers. Eighteen observers (8 males and 10 females) from the University of Texas at Dallas (UTD) undergraduate population were recruited in exchange for a core psychology course research credit.[5]

Stimuli. One hundred and fifty-two (half male and half female) Caucasian faces were digitized from slides to a 150- × 225-pixel image with a resolution of 16 gray levels using a digitizer attached to a PC with a TARGA board (True Vision). Faces were of young adults, without facial hair or glasses, and were photographed in front of a homogeneous light background. All of the face images were aligned with each other by eye height and by the center point between the eyes. Face images were not normalized explicitly for size, but were all taken from the same camera distance and thus were roughly equal in size. These stimuli were used in all experiments and in the simulation.[6]

Procedure. Observers were instructed that the purpose of the study was to determine the speed with which they could accurately determine whether a face was that of a male or a female. Each observer read a short description of the experiment explaining that faces would appear on a computer screen one at a time and would remain visible until a response was made by pressing a button on a three-button computer mouse. Observers pressed the left-most button for one gender and the right-most button for the other gender. The assignment of left/right to gender categorization was counterbalanced across observers and was labeled appropriately in all cases. When the observer responded, the face disappeared and a computer prompt appeared instructing observers to rate their certainty (1 = *very sure*, 2 = *moderately sure*, and 3 = *guessing*). They again indicated their rating using the three-button mouse, labeled with a paper overlay above the "male" and "female" button labels. All observers participated in a short practice session using Japanese faces in order to acquaint themselves with the task and with the equipment.

Results

The primary purpose of this study was to obtain measures of the speed with which individual faces are categorized by gender. However, for comparison with other related studies, we present a standard analysis of variance (ANOVA) on the observer data in this experiment. Similar analyses are presented also in Experiments 2 and 3.

Mean reaction times (RTs) to classify male and female faces were computed individually for male and female observers. The group means appear in Table 1.[7] These data were analyzed using a two-factor ANOVA, with gender of observer as a between-subjects factor and gender of face as a within-subjects factor. The analysis revealed

Table 1
Human Rating and Recognition Performance of Observers as a Function of Gender and Faces

	Observers			
	Male		Female	
	Male Faces	Female Faces	Male Faces	Female Faces
Experiment 1, Speeded Gender Classification				
Reaction time to classify by gender (msec)	1,141.31	1,286.80	1,161.08	1,440.06
Accuracy (% correct)	94.9	94.9	98.7	95.3
Experiment 2, Ratings				
Attractiveness	.69	.65	.56	.71
Femininity	—	−0.96	—	1.05
Masculinity	.74	—	.82	—
Experiment 3, Recognition				
d'	1.15	0.96	1.13	1.56
Criterion	0.01	0.34	0.00	0.42

no main effect of observer gender [$F(1,16) < 1$], but did reveal a main effect of face gender [$F(1,16) = 4.82, p < .05$], with female faces classified more slowly than male faces. The pattern of data is consistent with the presence of an interaction between face gender and observer gender, but this conclusion was not supported statistically [$F(1,16) < 1$].

The accuracy of gender classification for male and female observers on male and female faces was high in all cases (overall average = 95.9%). There was no indication of a speed–accuracy tradeoff. Furthermore, an ANOVA on errors revealed no main effects or interactions.

EXPERIMENT 2
Masculinity/Femininity and Attractiveness Ratings

Method

Observers. Eighteen observers (10 males and 8 females) from the UTD undergraduate population were recruited in exchange for a core psychology course research credit. These observers had not participated in Experiment 1.

Procedure. Observers viewed the faces one at a time on a computer screen and rated each face for attractiveness using the three-button computer mouse (0 = *unattractive*, 1 = *somewhat attractive*, and 2 = *very attractive*).[8] After the response, the face remained on the screen and observers then rated the male faces for masculinity and the female faces for femininity (0 = *not very feminine*, 1 = *somewhat feminine*, and 2 = *very feminine*). For the male faces, "femininity" was replaced in the computer prompt by "masculinity."[9] Male and female faces in this experiment were blocked and the order of these blocks was counterbalanced across observers.

Results

Mean attractiveness and masculinity/femininity ratings were computed individually for male and female observers. The group means appear in Table 1. The attractiveness rating data were analyzed using a two-factor ANOVA with the gender of observer as a between-subjects factor and gender of face as a within-subjects factor. The analysis for attractiveness revealed no main effect of observer gender [$F(1,16) < 1$], no main effect of face gender [$F(1,16) = 2.17, p > .05$], and a significant interaction between face gender and observer gender [$F(1,16) = 7.13, p < .05$]. As can be seen in Table 1, this interaction was due primarily to female observers rating female faces as more attractive than male faces. Male observers rated male and female faces to be about equally attractive.

Because only male faces were rated for masculinity and only female faces were rated for femininity, we analyzed the male and female face data in separate one-factor ANOVAs with observer gender as the independent variable. No differences were found in either case as a function of observer gender [$F(1,16) < 1$ in both cases; see Table 1 for means].

EXPERIMENT 3
Recognition

Method

Observers. Thirty-five observers (18 males and 17 females) from the UTD undergraduate population were recruited in exchange for a core psychology course research credit. These observers had not participated in Experiment 1 or 2.

Procedure. The final experiment was a standard *old/new* recognition memory study. Observers were instructed to pay close attention to the faces presented because they would be asked to remember the faces in the second part of the experiment. During the learning part of the study, 76 (38 male and 38 female) faces were presented one at a time on the computer screen for 3 sec. Observers took a short break and then viewed 152 faces (76 male and 76 female) one at a time. Half of these faces had been seen by the observer in the learning part of the study and half were new faces. Each face remained on the screen until the observer responded "old" or "new," using labeled mouse buttons. The order of faces presented in both the learning and testing phases was randomized individually for each observer. Due to the fact that we wished to use these data for computing the recognizability of male and female faces for male and female observers, elaborate counterbalancing schemes were implemented so that each face appeared equally often as *old* and *new* across observers, and also so that each face was seen equally often as *old* and *new* by equal numbers of male and female observers. Because of the slight imbalance in the number of male and female observers tested, this was not achieved precisely. Nonetheless, the d's and C values were calculated from hit and false alarm rates that were based on very close to equal numbers of *old* and *new* presentations.

Results and Discussion

A d' and criterion[10] were computed individually for each male and female observer recognizing male and female faces. The group means appear in Table 1. These data were analyzed using a two-factor ANOVA with the gender of observer as a between-subjects factor and gender of face as a within-subjects factor. The d' analysis revealed no main effect of face gender [$F(1,33) = 1.33, p > .05$], a main effect of observer gender [$F(1,33) = 4.82, p < .05$], with female observers more accurate than male observers, and an interaction between face gender and observer gender [$F(1,33) = 8.72, p < .01$], with female observers more accurate with female faces. These data are consistent with those of most previous work considering the effects of face and observer gender on face recognition accuracy (see Shepherd, 1981, for a thorough review). Shepherd (1981) noted that although face and observer gender effects have not been found consistently in the literature, when main effects have been found, they have tended to indicate an advantage for female observers. When interactions have been found, they have tended to indicate that female observers are particularly good at recognizing female faces.

The criterion analysis showed no main effect of observer gender [$F(1,33) < 1$], a main effect of face gender [$F(1,33) = 26.20, p < .01$], with observers using a stricter criterion for female faces than for male faces, and no interaction between face gender and observer gender [$F(1,33) < 1$].

Next, we computed a d' and C value for each face by compiling data across the different observers. Although the recognizability of individual faces has been measured frequently in the literature using methods based on signal detection theory (SDT; see, e.g., Hancock et al., 1996; Light et al., 1979; O'Toole et al., 1994), these studies did not make the assumptions of this model explicit as they apply to individual stimuli, rather than to individual observers. For completeness, we do so in the appendix. In the present section, we detail only the procedure used to compute the SDT measures on faces and refer readers interested in the SDT model assumptions to the appendix.

Hit rates and false alarm rates for individual faces were computed as follows. When a face was learned by an observer and was later recognized by that observer as "old," a hit was recorded for that face. The hit rate for the face was the proportion of times the face was recognized as *old* across all of the observers who had learned the face. When a face was not learned by an observer and the observer incorrectly recognized the face as *old*, a false alarm was recorded for that face. The false alarm rate for the face was the proportion of times the face was recognized as *old* across all of the observers who had *not* learned the face. A d' and criterion were computed in the standard way for each face.

COMBINED ANALYSIS ON FACES

Male and Female Faces

Using data from all three experiments, we assigned each face a value on the following five variables: (1) RT (i.e., mean latency to categorize the face by gender), (2) mean attractiveness, (3) mean femininity if female or mean masculinity if male, (4) d', and (5) criterion. In the first three cases (RT, attractiveness, and masculinity/femininity), means were computed across all observers in the appropriate experiment. In the latter two cases (d' and criterion), numbers of hits and false alarms were compiled across observers in Experiment 3 and a single d' and criterion were calculated for each face. Note that for the facial attributes of attractiveness, masculinity, and femininity, high variable values indicated high levels of the attribute in question (e.g., high numbers indicated highly attractive faces).

Table 2
Correlations Among Human Rating and Recognition Performance Measures: Upper Triangle Contains Correlations for Female Faces, Lower Triangle for Male Faces

	Reaction Time	Attractiveness	Femininity/ Masculinity	d'	Criterion
Reaction time	1.00	−.37†	−.55†	−.07	.02
Attractiveness	−.24*	1.00	.88†	.08	.13
Femininity/ masculinity	−.62†	.23*	1.00	.07	.05
d'	.09	−.30†	.10	1.00	.46†
Criterion	.10	−.17	−.15	.28†	1.00

$^{*}p < .05$. $^{\dagger}p < .001$.

The faces were then separated by gender and a varimax-rotated PCA was applied separately to the correlation matrices of the raw rating and performance data for male and female faces. For completeness, we present these raw correlation matrices in Table 2. Correlations for the female faces appear in the upper triangle of the matrix, whereas correlations for the male faces appear in the lower triangle of the matrix. Note that an interpretation of PCA data requires a decision about the number of axes/factors to retain. We based this choice on the structure of the resultant data and presented as many axes as we were able to interpret easily. Additionally, because there is no significance test to indicate the size of the loading to be considered important, it is useful to choose a loading value for all analyses that will be considered as a threshold for interpreting the variable loadings. For comparison with past work (O'Toole et al., 1994), we will restrict our conclusions to loadings greater than or equal to .30, which are marked with an asterisk in the tables.

Table 3
Human Rating and Recognition Performance for Male and Female Faces for the First Two Rotated Factors

	Female Faces		Male Faces	
	Classification	Recognition	Classification	Recognition
Reaction time	−.71*	.04	−.85*	.13
Attractiveness	.89*	.11	.32*	−.61
Masculinity/ femininity	.96*	.04	.90*	.00
d'	.05	.84*	−.15	.83*
Criterion	.01	.86*	.10	.65*
Proportion of variance accounted for by axis	.45	.28	.37	.26

The results of the PCA analysis appear in Table 3. We retained the first two axes, which explained 73% of the variance for female faces and 63% of the variance for male faces. Several points are worth noting. First, in terms of the performance measures, RT and d' appear independently on the first and second axes, respectively, for both male and female faces. For convenience and brevity, we will henceforth refer to these two axes as the "classification" and "recognition" axes, respectively. The independence of d' and RT is at odds with our simple conceptualization of human classification and recognition performance being dependent on a face's distance from a local subcategory prototype. We consider this question in more detail in the discussion.

Second, beginning with female faces, it is clear from the first axis that the attractiveness and femininity judgments were strongly related to RT to classify the faces as female. Faces rated as highly feminine and highly attractive were classified as female more quickly than faces judged to be less feminine and less attractive. The similarly sized loadings for attractiveness and femininity on this axis suggest that observers used the attractiveness and femininity ratings in very similar ways. The similar usage of these ratings sits uncomfortably with our tentative conceptualizations of attractiveness as "average" and femininity as a "caricature" of female (i.e., its contrast from male faces).

The classification axis for male faces is more complicated. In general, faces judged to be masculine and attractive were classified as male more quickly than faces judged less masculine and unattractive. However, the difference in the size of the loading for attractiveness and masculinity suggests that for male faces "attractive" did not equal "masculine." Masculinity proved more closely tied to the speed required to classify the face as male than did attractiveness. As expected from Light et al. (1981), the attractiveness rating loads in the opposite direction from d' on the recognizability axis for male faces, indicating that attractive faces were not well recognized. In contrast to the female face data, the dissociation of the attractiveness and masculinity ratings is reasonably consistent with the notion that at least part of the attractiveness rating for male faces captures "averageness," and that masculinity captures something caricatured about male faces.

The recognition axis for both the male and female faces is dominated by the relationship between d' and criterion. Additionally, the attractiveness rating did not load on the recognizability axis for female faces, as it had for male faces. This result replicates Shepherd and Ellis's (1973) finding that attractiveness and recognizability for female faces are unrelated. The present data, combined with the results of Light et al. (1981) and Shepherd and Ellis, suggest that the negative relationship between attractiveness and recognizability holds only for male faces.

Finally, d' and criterion were not independent for either the male or the female faces. This result seemed inconsistent with earlier data collected on these faces (O'Toole et al., 1994). This earlier study indicated that d' and criterion were independent for same-race Caucasian faces but not for other-race Japanese faces. The major difference between the recognition experiment carried out in that study and the one performed here was the balance of male and female observers. In the O'Toole et al. (1994) study, approximately 80% of the observers were female, whereas here, the proportion was close to 50%.

Male and Female Observers and Male and Female Faces

To examine the possibility that the difference in the balance of male and female observers between O'Toole et al. (1994) and the present study was responsible for the nonindependence between d' and criterion found here, and also to extend the present results to look at differences between male and female observers, we repeated the PCA analysis, separating the data further by the gender of observer. Thus, we performed four PCA analyses: (1) female observers with female faces, (2) female observers with male faces, (3) male observers with female faces, and (4) male observers with male faces. Again, in all four of these analyses, the first axis was interpretable as a classification axis. Since these data were similar to those found previously with male and female observers combined, we will not consider the first axis further. The second axis, however, although again identifiable as a recognition axis, differed as a function of observer and face gender. This axis appears in Table 4 for each of the four analyses. First, in all cases, it is clear that d' and criterion were not independent. The pattern of nonindependence differed as a function of observer and face gender. For male observers of both male and female faces, well-recognized faces tended to be recognized with strict (more conservative) criteria. This was also true for female observers with female faces. For female observers viewing male faces, however, the relationship between d' and criterion was in the opposite direction. Well-recognized faces tended to be recognized with looser (more liberal) criteria. Thus, it seems that this opposing relationship for female observers on male and female faces, combined with the preponderance of female observers in O'Toole et al. (1994), can explain why the data from that study showed independence between d' and criterion.

Table 4
Recognition Axis as a Function of Face and Observer Gender

	Observers			
	Female		Male	
	Female Faces	Male Faces	Female Faces	Male Faces
Reaction time	−.17	.00	.12	.10
Attractiveness	.10	−.64*	.04	−.64*
Masculinity/femininity	.02	−.10	.02	−.01
d'	.83*	.78*	.77*	.71*
Criterion	.82*	−.49*	.81*	.68*
Proportion of variance accounted for by axis	.25	.26	.26	.26

At present, we can offer neither an explanation nor an interpretation of the nonindependence of d' and criterion in these data. We are further unsure as to why there were differences in the pattern of nonindependence as a function of gender of observer and gender of face. The result is important, however, because it suggests that correlations between facial attribute ratings and single components of d' (i.e., hit rate and false alarm rate) can be very difficult to interpret and can be misleading in some cases (see O'Toole et al., 1994, for a discussion of the problem). Because of the complexity of the issue and the fact that it is not central to the theme of the present paper, we will not consider criterion further, though we believe it (and its relationship to d' in standard face recognition studies) to be worthy of further study in its own right.

Results, Summary, and Conclusions

From the combined analysis of the first three experiments, several points are worth noting. First, independence was found between the recognizability of a face and the speed in classifying it as male or female. As noted, nonindependence might be predicted by a model that (1) considers the RT to classify a particular face by gender as a measure of the distance of that face to the subcategory center or prototype (i.e., in this case, to the average male or female face), and that (2) assumes that faces are more densely clustered (and hence, less distinguishable) close to this subcategory center. Independence of RT and d' is more consistent with a model in which caricatured rather than prototypical aspects of gender appearance underlie gender classification performance, whereas the similarity structure or density of stimuli (presumably highest around the subcategory prototype) underlies face recognizability. We consider these issues further after presenting our computational model.

Second, for female and male faces, speed of classifying faces by gender was related both to the attractiveness and femininity/masculinity of the faces. The similarity of the attractiveness and femininity loadings indicates that to a first approximation, observers tended to use the attractiveness and femininity ratings in very similar ways—a finding that is inconsistent with the conceptualization of attractiveness as "average" and femininity as our tentatively conjectured "caricature" of female.

Third, consistent with the findings of Shepherd and Ellis (1973) for female faces, attractiveness ratings and recognizability were not related. For male faces, attractiveness ratings were made of two independent components, one related to the masculinity of the face and to the speed of classifying it as male, and a second related to the recognizability of the face. This suggests that observers base attractiveness ratings of male faces on two kinds of information about the faces, one related to masculinity and RT, and the other to recognizability and criterion. The latter component is consistent with the findings of Light et al. (1981), using male faces. Combined with the findings of Shepherd and Ellis, the present data suggest that the negative relationship between recognizability and attractiveness holds only for male faces. The differences seen between the pattern of rating and performance measures for male and female faces indicate that caution should be exercised in interpreting the results of experiments using these ratings with both male and female faces.

Finally, differences in the way male and female observers processed these faces were confined primarily to the recognition axis, on which male and female observers showed different patterns of relationships between recognizability and response bias.

MODEL DESCRIPTION

The psychological data indicated differences in the pattern of interrelationships among gender-related facial ratings and recognition performance measures for male and female faces. How do stimulus structure differences between/among male and female faces relate to these differences? Answering this question requires an ability to quantify the information in faces in a way that captures the visually derived semantic information relevant for categorizing faces by gender. We coded each individual face as a vector of pixels created by concatenating the rows of the face image. We then applied a PCA to the cross-product matrix made from the set of face vectors. As noted, a face is represented in this model as a weighted combination of "features" (PCs, eigenvectors, axes, and dimensions). Because the PCA is applied to images, each eigenvector is interpretable or "displayable" as an image. Figure 1 shows the first nine eigenvectors extracted from a matrix made of male and female faces. In (re)constructing a particular face, these eigenvectors are combined linearly along with the remaining eigenvectors in the set.

In the context of representing faces, eigenvectors have two defining characteristics. First, eigenvectors can be ordered according to the amount of variance (referred to as the eigenvalue of the eigenvector) each explains in the cross-product matrix made from the set of faces. This is a measure of the importance of the eigenvector for representing *all* faces in the set and is important for understanding properties of the representation that relate to the heterogeneity of the face set. Second, different "amounts" of each eigenvector are required to reconstruct *particular* faces. These amounts measure the importance of the individual eigenvectors for representing individual faces and are important for understanding how a particular face differs from other faces in the set. We refer to these amounts for a particular face as the face's "weights" with respect to the eigenvectors.

We have concentrated on eigenvectors explaining large proportions of variance in the face set, since a number of studies have indicated that these eigenvectors contain reliable information for predicting the gender of a face (Abdi et al., 1995; O'Toole et al., 1993; O'Toole et al., 1997). This is not surprising, given that gender is one of the basic "features" on which faces can be contrasted, and hence is likely to explain a large proportion of variance in a set of faces.

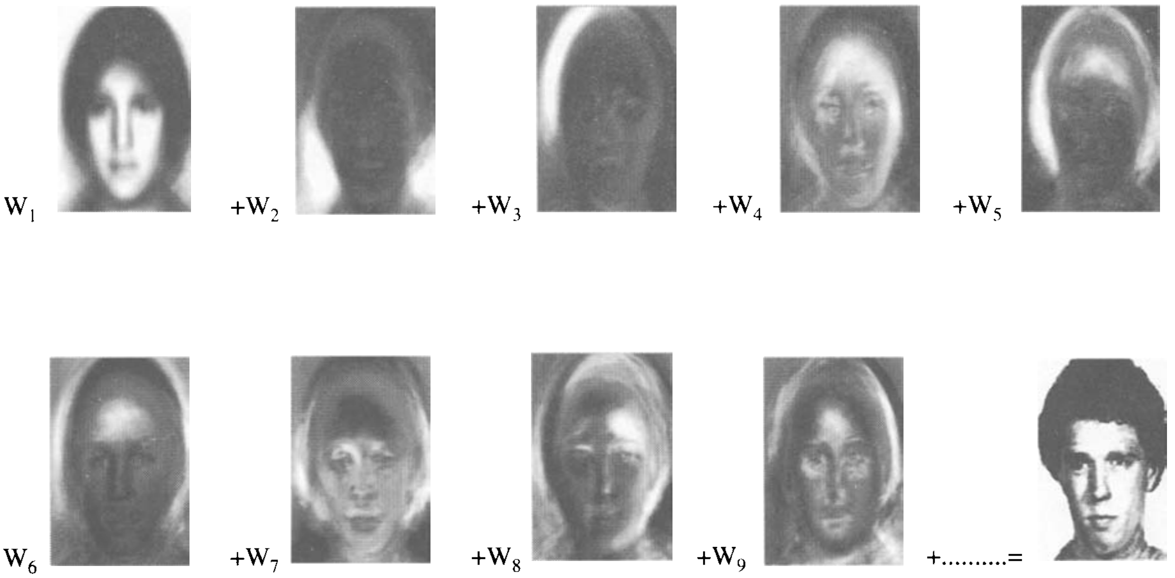

Figure 1. Schematic of the combination of eigen-images to create a face. The first nine eigen-images are displayed. For illustration purposes, the weights for these first nine eigen-images for the particular face shown are as follows: .89 × the first eigen-image $(e_1) + .30e_2 + .07e_3 - .04e_4 + 0e_5 - .09e_6 + .03e_7 + 0e_8 - .02e_9 + \ldots + w_n e_n$.

The relationship between individual eigenvectors and the gender of a face has been established in previous work by computing a point biserial correlation between the weights of faces on particular eigenvectors and the gender of the faces[11] (O'Toole et al., 1993). Using the same set of faces that we used in the present study, O'Toole et al. found statistically reliable relationships for 12 eigenvectors with relatively large eigenvalues. The strongest relationship was found for the second eigenvector (r =.66, df = 157, p < .0001). In general, a positive weight on this eigenvector was required to reconstruct male faces, whereas a negative weight was required to reconstruct female faces. Accordingly, O'Toole et al. showed that adding the second eigenvector to the first produced a face with a male appearance, whereas subtracting the second eigenvector from the first produced a face with a female appearance.[12] This demonstration is reproduced in Figure 2. The first row of the figure illustrates, from left to right, the first three eigenvectors. Row 2 of the figure shows the result of adding the first eigenvector to the second (left face) and the result of subtracting the second eigenvector from the first (right face). This eigenvector captures hair length and face shape differences between male and female faces (see also Abdi et al., 1995, and O'Toole et al., 1997, for an analysis of the computational utility and generalizability of eigenvectors for the gender classification task).

O'Toole et al. (1993) also found the third eigenvector weight to be a reliable, though much less powerful, predictor of face gender (r = .21, df = 157, p < .006). Again, male faces generally required positive values of this eigenvector, whereas female faces generally required negative values. We then combined this eigenvector with the first eigenvector in positive and negative combinations. The results appear in row 3 of Figure 2. The face on the left is the result of adding the third eigenvector to the first, whereas the face on the right is the result of subtracting the third eigenvector from the first eigenvector. Surprisingly, the eyes and head of the "female" face are turned very slightly in comparison to the "male" face.[13] This would indicate that some female subjects did not gaze directly at the camera, but rather, just to the side—a surprising result in that the pictures of males and females in this set were taken under identical pose conditions and with identical instructions (A. Goldstein, personal communication).[14] Nevertheless, this difference proved a reliable discriminator of face gender.

We concentrated on 3 of the 12 eigenvector weights found to be predictive of face gender by O'Toole et al. (1993)—the first, second, and third eigenvector weights. Combined, these three weights explained 58.86% of the total variance in gender prediction, and 65.4% of the variance explained by the 12 weights that were statistically significant gender predictors. We have chosen to include these three eigenvector weights in the present analysis because they were the most strongly predictive of face gender and because we can offer a prima facie interpretation of the information they capture. We have mentioned our interpretation of the second and third eigenvectors. The first eigenvector also related to face gender in the study of O'Toole et al. (1993); (r = .33, df = 157, p < .0001). Due to the fact that we did not subtract the mean face prior to the extraction of eigenvectors, the highly similar nature of the face images is such

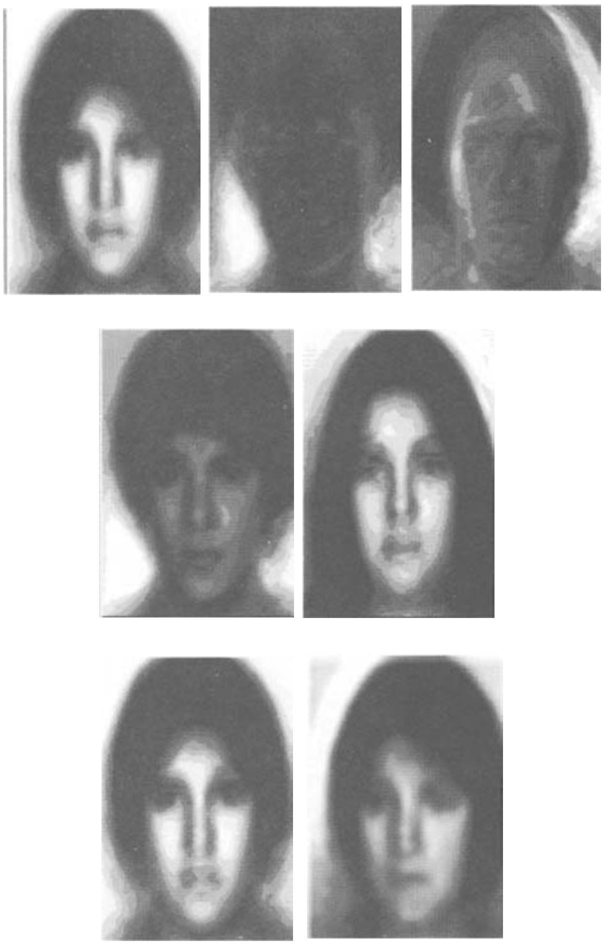

Figure 2. The first three eigen-images of a face matrix composed of equal numbers of male and female faces (row 1). The first eigenvector plus the second eigenvector appears on the left of row 2, making a face with a male appearance. The first eigenvector minus the second eigenvector appears on the right of row 2, making a face with a female appearance (row 2). The first eigenvector plus the third eigenvector appears on the left of row 3. The first eigenvector minus the third eigenvector appears on the right of row 3. The face on the right appears more feminine than the face on the left.

that this eigenvector approximates the average face (for more details, see Valentin, Abdi, & O'Toole, in press). Accordingly, unlike the positive/negative image differences we saw for the second and third eigenvectors, both male and female faces required large positive values of this eigenvector to be reconstructed, with the reconstructions of male faces requiring a significantly larger amount of this eigenvector (i.e., the average) than the female faces. The fact that the "amount" of the average face required to reconstruct a face correlated significantly with the gender of the face suggests an interesting aspect of the stimulus structure of our particular set of male and female faces. Specifically, it would seem that, on the average, male faces were "closer" to the general face average than were female faces.[15]

SIMULATION
Quantifying Gender Information in Faces

The present methods are very similar to those applied in previous studies, and so we provide only a brief overview of the computational analysis. A more detailed description appears in O'Toole et al. (1994) and a tutorial explanation can be found in Abdi (1994).

Method

Stimuli. The same 152 faces used in Experiments 1–3 were used for the simulation as well. For the model, these faces were 151 pixels in width and 225 pixels in length, and were digitized to 16 gray levels. These faces constitute 152 of the 159 faces used by O'Toole et al. (1993).

Procedure. For model-predicted gender information, we extracted the weights for each face with respect to the first, second, and third eigenvectors. These three weights for a given face served as a model-derived measure of the gender-related information in that face with respect to the information captured by these eigenvectors.

For model-predicted recognizability information, the recognizability measure should capture the extent to which individual faces are distinctive or unusual with respect to other faces in the set—in this case, distinctive with respect to the basic categorical structure defined by gender. In short, our measure tries to answer the question, "How distinctive is a face once we have partialed out some of this basic categorical structure information?" Or, "How different is it from the gender subcategory average?" This was determined in two steps. First, because we know that the first three eigenvectors or eigen-images relate to face gender, we can reconstruct faces eliminating these eigen-images. Second, we computed the cosine or normalized correlation between each reconstructed face vector and its original face vector (see O'Toole et al., 1994). This measures the similarity between the two vectors and provides an indication of how much information is "left over" after eliminating most of the useful gender information in the faces. Relatively high similarity of these reconstructions to the originals (i.e., the cosine is high) indicates that there is a relatively large amount of information leftover for distinguishing a particular face from the gender subcategory to which it belongs. Relatively low similarity of these reconstructions to the originals indicates that there is a relatively little information leftover for distinguishing the face from the category prototype.

Canonical correlation analysis. Before examining the structure among the variables, we assessed the strength and reliability of the relationship between model and human measures. Canonical correlation can be used to assess the statistical reliability of the linear relationship between two *sets* of variables. In this analysis, a linear combination within each set of variables was computed so as to maximize the correlation between the two sets of variables (Kshirsagar, 1972). We carried out separate analyses for male and female faces[16] so that the "gender" of the face alone could not be responsible for any correlation found. The model measures were the weights on the first three eigenvectors and the cosine, computed as indicated previously. The human measures were attractiveness, masculinity/femininity, RT, and d'.[17] This yielded a canonical correlation of .49 (maximum likelihood ratio test, $p < .03$) for the female faces and .52 (maximum likelihood ratio test, $p < .001$) for the male faces. The results of this analysis indicate that there is a statistically reliable relationship between the model and human measures.

Divided further by the gender of observer, the canonical correlations were not significant, possibly due to the loss of power incurred in dividing the number of cases by 2. Given the lack of significance and the minor differences due to observer gender in the psychological data, we do not present these analyses.

Table 5a
Human and Model Data for Female Faces for the First Three Rotated Factors

	Classification	Recognition	Axis 3
Reaction time	−.68*	−.14	.34*
Attractiveness	.89*	−.10	.19
Femininity	.96*	−.04	.00
d′	.18	.65*	.20
EV 1 weight	.05	−.57*	.76*
EV 2 weight	−.04	.08	.88*
EV 3 weight	−.31*	.48*	.09
Cosine	−.01	.80*	−.27
Proportion of variance accounted for by axis	.30	.27	.13

Table 5b
Human and Model Data for Male Faces for the First Three Rotated Factors

	Classification	Recognition	Axis 3
Reaction time	−.86*	−.01	−.09
Attractiveness	.54*	−.36*	−.21
Masculinity	.80*	−.20	.10
d′	.14	.71*	.14
EV 1 weight	−.21	−.79*	.07
EV 2 weight	.15	.10	.88*
EV 3 weight	.35*	.16	−.76*
Cosine	−.03	.90*	−.13
Proportion of variance accounted for by axis	.24	.28	.16

Varimax PCA analysis combining human and model data. Next we examined the structure of the relationship among model and human measures by applying a varimax-rotated PCA to the combined model and human data for the male and female faces. We simply supplemented the human rating and performance measures for each face with the four additional model-derived measures for each face. The results of the varimax-rotated PCA analysis appear in Table 5. Note that for all three of the eigenvector weights, high numbers indicate values toward the male end of the scale, low numbers indicate values toward the female end (see Figure 2), and high cosines indicate high-quality reconstructions (reconstructions similar to the original faces).

In both the male and female analyses, we retained three axes, accounting for a total of 70% of the variance for female faces and 68% of the variance for male faces. The classification and recognition axes seen in the psychological data are again identifiable as the first two axes. We discuss the third axis individually for the male and female faces. Because there is no appropriate common label for this axis, in Tables 5a and 5b we simply label it "Axis 3."

Several points are worth noting. First, for both male and female faces, the strongest overlap between model and human measures occurred on the recognition axis, which shows roughly equally sized loadings for model and human measures (see Tables 5a and 5b). The common structural element of this recognition axis for male and female faces is a loading of *d′* and cosine in the same direction, opposing an inverse loading of the first eigenvector coefficient.[18] The direction of the cosine-*d′* relationship loading indicates that faces with higher quality model representations were better recognized by human observers. That is, these faces were more readily distinguished from their subcategory prototype and were, therefore, presumably more distant from it in the face space. This replicates a similar finding by O'Toole et al. (1994) when identity-specific information in the face representations was preserved (i.e., when faces were reconstructed with eigenvectors with relatively smaller eigenvalues).[19]

The first eigenvector coefficient, or "amount of the average face required to reconstruct the face," also loaded in opposition to *d′*. Faces more similar to this common face base were less recognizable than were faces less similar to this base. This is consistent with common interpretations of prototype theory for faces.

The pattern of results for male and female faces diverged on the recognition axis in two ways. For male faces, attractiveness loads on this axis in the direction expected—that is, with the first eigenvector coefficient, and against *d′* and cosine. This is consistent with the suggestion that one component of attractiveness in male faces makes for a less recognizable, and in part, more "average" face. For female faces, the third eigenvector coefficient also loaded in the same direction as *d′* and cosine, so that female faces with more masculine values of this weight—that is, more frontal-looking faces—were more recognizable. It might be possible that the latter component captures something related to the attractiveness–recognizability relationship found for male faces, with these female faces being treated like unattractive (more discriminable) male faces, rather than female faces.

For the classification axes (see Tables 5a and b), the pattern of human data is again characterized by the opposition of RT and the combination of attractiveness and masculinity/femininity. Note that in contrast to the recognition axis, where model and human measures had roughly an equal foothold, this classification axis was primarily dominated by the human measures. Surprisingly, the model measure that loaded most strongly on this axis, and the only model measure to load above our criterion, was the weight on the third eigenvector. Despite its modest size, this loading appears for both male and female faces. For female faces, femininity and attractiveness loaded in a direction opposing the third eigenvector weight and RT. Thus, feminine and attractive faces, which were classified as female relatively quickly, tended to have smaller (more negative) values of the third eigenvector and hence tended to be slightly turned from the camera. For male faces, masculinity and attractiveness loaded in the same direction as the third eigenvector weight. In other words, masculine and attractive faces tended to have larger (more positive) values of the third eigenvector and hence tended to gaze directly at the camera.

Finally, the third axis retained in this analysis (see Tables 5a and b), though dominated by model measures, is interesting for female faces in that it contains a second orthogonal component of RT, related only to model measures. RT appears on this axis in the same direction as the first and second eigenvector weights. This indicates that faces requiring larger amounts of the average face to be reconstructed (i.e., female faces with more male values of this eigenvector weight) were classified as female more slowly than were faces requiring less of the general average. Additionally, female faces with more male values of the second eigenvector weight were classified more slowly as female. In general, these are female faces with short hair and more male-shaped faces, as defined by this eigenvector (see Figure 2). This second component of RT was detected in this analysis, but not in the analysis of the purely psychological data, due to the presence of model measures relevant to the information on which it was based. This information was apparently not captured in the facial characteristic ratings or recognition performance measures.

For male faces, the third axis was completely dominated by model measures and shows only that the second and third eigenvector weights loaded in opposition for male faces.[20] Since no human measure loaded in this axis at a level meeting our criterion, this simply indicates an axis dissociating typically masculine values on the second and third eigenvectors.

SUMMARY AND DISCUSSION

In summarizing the specifics of these results for female faces, "attractive" was very nearly synonymous with "feminine" and was related to the time required to call the face "female." This finding indicates that our tentatively advanced conceptualizations of "attractiveness as average" and "femininity as a caricature" cannot both be correct. We would argue that "caricature," rather than "average," may be a better descriptor of the information captured by these ratings. Supporting this conclusion, the model data indicated that the distinguishability of the face when some of the basic gender information was eliminated (distinctiveness with respect to the prototype female face) was not related either to the attractiveness/femininity rating or to the RT to classify the face as female. By contrast, this model-based distinctiveness information was quite strongly related to the recognizability of the face. The premise here is that a caricature is built by opposition to a contrastive category. For example, a "caricatured female" emphasizes/exaggerates the features that most distinguish it from male faces. Recognition memory performance, on the other hand, would be more concerned with the local category structure (i.e., female), since it is presumably most related to the number of similar distracting items for an individual face.

In summarizing the specifics of these results for male faces, "attractive" was not synonymous with "masculine." Rather, attractiveness was a 2-D entity, one dimension of which mirrored the unidimensional attractiveness rating seen for female faces and related to masculinity and the time required to classify the face as male. The second dimension of attractiveness related to the model-derived measure of the distinguishability of the face from the male prototype and, importantly, to the recognizability of the face for human observers. These results indicate that although masculinity may be seen as an "attractive" property of male faces, it is possible that extreme masculinity in a face may render it a bit too "strong" looking. It is possible, therefore, that the masculine component of attractiveness may need to be tempered or toned down somewhat for a male face to be judged attractive. By contrast, it is somewhat hard to imagine extreme femininity rendering a female face unattractive.

In relating these findings to past work, our psychological data clear up the disagreement in the literature concerning the relationship between attractiveness and recognizability, replicating the findings of both Light et al. (1979) and Shepherd and Ellis (1973). The critical factor explaining the difference in results between these studies is face gender, which dissociates two subcomponents of attractiveness for male, but not for female, faces. This dissociation is important for interpreting results that draw on rating and performance measures gathered on both male and/or female faces.

Additionally, with reference to past work, the fact that we did not find a relationship between attractiveness and "average" for the female faces is not necessarily inconsistent with the claim that an averaged or composite female face is more attractive than most single noncomposite faces. In contrast to past work (Langlois & Roggman, 1990; Langlois et al., 1994), the attractiveness ratings used in the present study were collected on "unprocessed" faces (i.e., single noncomposite faces), rather than on composite or averaged faces. The process of averaging faces can selectively obliterate relatively low-contrast, high-spatial-frequency (i.e., finely detailed) information that is specific to only one or a few of the faces in the set to be averaged (cf. Langlois et al., 1994). This could include small skin irregularities such as blemishes, which may render a face less attractive, as well as dimples or long eye lashes, which may render a face more attractive (see also Perrett et al., 1994, for more discussion of this issue). Primarily, with respect to the model proposed here, this kind of information is likely to be contained in eigenvectors with relatively small eigenvalues (see O'Toole et al., 1993). The presence of this low-contrast, high spatial frequency information, by its very definition, is likely to have a negligible effect on the arithmetically computed distance of a face to the average face. This information may be, nonetheless, clearly detectable for human observers in a single or unaveraged face and may have very important consequences for perceived attractiveness. In other words, although *averaged* faces may be generally judged to be more attractive than single faces, single faces that are close to the average may contain low-contrast, but detectable, features that are very important for the human judgment of attractiveness.

One last speculative point we wish to make refers to the nature of stimulus information contributing to masculinity and femininity judgments. This concerns the small but consistent loading of the third eigenvector coefficient on the classification axis for both male and female faces. Our interpretation of the information provided by this model measure is that it conveys information about a facial mannerism. Thus, it seems possible that a face can be made to appear (at any given instant) more feminine or more masculine via some very simple facial mannerisms. For example, looking straight ahead and seeking direct eye contact may lend any face a more masculine appearance, whereas averting the eyes and gazing downward may lend a face a more feminine appearance. These are simple, though subtle, changes in the orientation of faces, which (1) were useful in explaining variance in the face set (i.e., were captured by the third eigenvector), (2) were useful in predicting the gender of a face in purely computational terms, and (3) related to the human measures captured by the classification axis.

Examining the interrelationships among commonly assessed facial rating and performance measures can give insight into the potentially multidimensional components of these measures. The consistency of this relationship across groups of faces that vary in base configural properties such as gender, race, or age may be an important element in developing and refining face processing theories to fit the heterogeneous nature of the faces we encounter in the course of our social experience.

REFERENCES

Abdi, H. (1994). A neural network primer. *Journal of Biological Systems*, **2**, 247-281.

Abdi, H., Valentin, D., Edelman, B., & O'Toole, A. J. (1995). More about the difference between men and women: Evidence from linear neural networks and the principal component approach. *Perception*, **24**, 539-562.

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

Brown, E., & Perrett, D. I. (1993). What gives a face its gender? *Perception*, **22**, 829-840.

Bruce, V., Burton, A. M., Dench, N., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., & Linney, A. (1993). Sex discrimination: How do we tell the difference between male and female faces? *Perception*, **22**, 131-152.

Bruce, V., Ellis, H., Gibling, F., & Young, A. (1987). Parallel processing of the gender and familiarity of faces. *Canadian Journal of Psychology*, **41**, 510-520.

Bruce, V., & Langton, S. (1994). The use of pigmentation and shading information in recognising the sex and identities of faces. *Perception*, **23**, 803-822.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology*, **77**, 305-327.

Burton, A. M., Bruce, V., & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception*, **22**, 153-176.

Chronicle, E. P., Chan, M., Hawkings, C., Mason, K., Smethurst, K., Stallybrass, K., Westrope, K., & Wright, K. (1995). You can tell by the nose—Judging sex from an isolated facial feature. *Perception*, **24**, 969-973.

Cottrell, G. W., & Fleming, M. K. (1990). Face recognition using unsupervised feature extraction. In *Proceedings of the International Conference on Neural Networks* (pp. 322-325). Dordrecht: Kluwer.

Enlow, D. (1982). *Handbook of facial growth*. Philadelphia: W. H. Saunders.

Goldstein, A. G., & Chance, J. E. (1980). Memory for faces and schema theory. *Journal of Psychology*, **105**, 47-59.

Golomb, B. A., Lawrence, D. T., & Sejnowski, T. J. (1991). SEXnet: A neural network identifies sex from human faces. In R. P. Lippmann, J. Moody, & D. S. Touretsky (Eds.), *Advances in neural information processing systems 3* (pp. 572-577). San Mateo, CA: Morgan Kaufmann.

Gray, M. S., Lawrence, D. T., Golomb, B. A., & Sejnowski, T. J. (1995). A perceptron reveals the face of gender. *Neural Computation*, **7**, 1160-1164.

Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition*, **24**, 26-40.

Kshirsagar, A. M. (1972). *Multivariate analysis*. New York: Marcel Dekker.

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

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

Langlois, J. H., Roggman, L. A., Mussleman, L., & Acton, S. (1991). A picture is worth a thousand words: A reply to "On the difficulty of averaging faces." *Psychological Science*, **2**, 354-357.

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

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Memory & Learning*, **5**, 212-228.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: A two-process theory of infant face recognition. *Psychological Review*, **98**, 164-181.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A*, **10**, 405-410.

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1995). A perceptual learning theory of the information in faces. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition* (pp. 159-182). London: Routledge.

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, **22**, 208-224.

O'Toole, A. J., Peterson, J., & Deffenbacher, K. A. (1996). An other-race effect for categorizing faces by sex. *Perception*, **25**, 669-676.

O'Toole, A. J., Vetter, T., Troje, N. F., & Bülthoff, H. H. (1997). Sex classification is better with three-dimensional structure than with image intensity information. *Perception*, **26**, 75-84.

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

Pittenger, J. B. (1991). On the difficulty of averaging faces. *Psychological Science*, **2**, 351-353.

Rhodes, G. (1988). Looking at faces: First-order and second-order features as determinants of facial appearance. *Perception*, **17**, 43-63.

Roberts, T., & Bruce, V. (1988). Feature saliency in judging the sex and familiarity of faces. *Perception*, **17**, 475-481.

Rowland, D. A., & Perrett, D. I. (1995). Manipulating facial appearance through shape and color. *IEEE Transactions on Computer Graphics & Applications*, **15**, 70-76.

Shepherd, J. W. (1981). Social factors in face recognition. In G. M. Davies, H. D. Ellis, & J. W. Shepherd (Eds.), *Perceiving and remembering faces* (pp. 55-79). London: Academic Press.

Shepherd, J. W., & Ellis, H. D. (1973). The effect of attractiveness on recognition memory for faces. *American Journal of Psychology*, **86**, 627-633.

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, **3**, 519-524.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, **3**, 71-86.

Valentin, D., Abdi, H., & O'Toole, A. J. (in press). Principal component and neural network analysis of face images: Explorations into the nature of information available for classifying faces by gender. *Journal of Mathematical Psychology*.

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

Valentine, T., & Bruce, V. (1986). Recognizing familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology*, **40**, 300-305.

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

Yamaguchi, M. K., Hirukawa, T., & Kanazawa, S. (1995). Judgment of gender through facial parts. *Perception*, **24**, 563-575.

NOTES

1. However, age may have less of a "categorical" and more of a continuous structure than race and gender.

2. The 3-D features were derived from full and profile views of the faces.

3. These have been referred to as "eigen-pictures" by Sirovich and Kirby (1987) and "eigen-faces" by Turk and Pentland (1991).

4. Although masculinity and femininity are considered orthogonal dimensions in the personality literature, in the face perception litera-

ture, femininity/masculinity have assumed a single scale (see, e.g., Bruce, Ellis, Gibling, & Young, 1987; Burton et al., 1993).

5. The number of observers may seem relatively small in Experiments 1 and 2, but our primary analysis treats "cases" as faces, rather than as observers. Additionally, of the eight possible main effects/interactions tested in these two experiments, in all but one case, the observer-based analyses of variance yielded F values that were either less than 1 or proved statistically significant, indicating that it would be unlikely that more observers would have changed results substantially.

6. For comparison purposes, it should be noted that these faces constituted 152 of the 159 Caucasian faces used in O'Toole et al. (1993) and O'Toole et al. (1994).

7. These mean reaction times are substantially longer than those reported in other studies of gender classification times (e.g., Bruce, Ellis, Gibling, & Young, 1987, who found gender classification times of slightly over 600 msec). Perhaps the major difference between this task and similar ones was the inclusion of a certainty rating task intervening between speeded classification trials in our study. This led us to wonder if this intervening task could have "broken the stride" of the observers in the reaction time task. We thus tested an additional 5 observers in the classification task, eliminating the rating task, and found that the mean reaction time dropped by 211 msec. Although still longer than in other studies, the factor analysis will show that reaction times related to the other psychological measures of gender-related attributes in interpretable ways.

8. In theory, a 5- or 7-point rating scale would have been better, but a 3-point scale was sufficiently sensitive in this study to capture stong and meaningful variations among the face measures (see Table 2). The correlations between measures define lower bounds on the reliability of the measures (Nunnally, 1978), thus allaying concerns about the consistency with which observers rated faces.

9. We note that this method allows for possible order effects of always rating attractiveness before masculinity/femininity.

10. We used C, a measure of the displacement of the criterion in z-score units, computed as $-0.5(z_{\mathrm{H}} + z_{\mathrm{FA}})$. With this measure, smaller values imply looser criteria (Snodgrass & Corwin, 1988).

11. Face gender was defined as 0 for female and 1 for male.

12. See also O'Toole et al. (1997) for a replication of this finding with 3-D data from laser scans of human heads.

13. When we showed these faces to people informally, all agreed that the face on the right appears female, but not all agreed that the face on the left appeared male. In any case, all seemed to agree that the right-hand face appears more feminine than the left-hand face.

14. It is worth noting that this is a very subtle cue. We have used this set of faces in many experiments and have never noticed differences in the gaze direction of the male and females in the photographs, though our PCA model detected it. Although this may be considered a "problem" for the standardization of the photographs, it would apply perhaps to other face sets that have not been similarly analyzed for systematic "nonfeature" differences between male and female faces.

15. The weight on the first eigenvector is simply the dot product between the vector of pixel values specifying a face and the first eigenvector and hence is a direct measure of the physical similarity between the two.

16. A joint analysis done over all faces was not necessary since we knew already that the three model gender measures were correlated with the sex of the face, which is embodied in the masculinity and femininity judgments. For completeness, however, when the male and female faces were combined, the canonical correlation between model and human measures was .74 (maximum likelihood ratio test, $p < .0001$).

17. As noted, we omitted the criterion from further consideration. For completeness, we carried out canonical correlations including criterion, but the inclusion of criterion did not change the size of the correlation substantially.

18. The opposition of the two model measures, cosine against the first eigenvector coefficient, is in part artifactual because (1) the first eigenvector is highly related to the mean, so all faces will have strong positive values on it; (2) the larger this weight for a given face, the larger the variation in the face explained by the first eigenvector, and the less explained by the eigenvectors contributing to the cosine measure.

19. In O'Toole et al. (1994), a much larger range of the eigenvectors was eliminated in this identity-specific condition. It would appear, therefore, that eliminating only the first three eigenvectors in the present analysis was sufficient to replicate this finding.

20. For specialists of PCA, who may be disturbed by a cross-loading of two eigenvector weights from orthogonal eigenvectors, recall that the PCA on faces was carried out for male and female faces combined, whereas the varimax-rotated PCA on the face measures, where we see this cross-loading, was done individually for male and female faces.

APPENDIX
Signal Detection Model Comparison for Observers and Faces

Signal detection model for observers. When computing the measures d' and C for a particular observer in a particular condition of a recognition experiment, data from many different faces are combined. Each hit that contributes to the hit rate and each false alarm that contributes to the false alarm rate comes from a different face. All the different faces that an observer learned in the learning phase of a recognition experiment contribute to the *old* distribution, and all the faces that the observer did not learn, but that are used as test faces, contribute to the *new* distribution. In general, the dimension on which these faces are distributed is thought to be an indication of the level of familiarity that *a particular observer* experiences when looking at faces. To be able to recognize faces at a level above chance, the observer must experience generally higher levels of familiarity when viewing faces he/she has seen before than when viewing faces that he/she has not seen before.

The d' measures the overlap of evoked familiarity feelings for an observer when *old* versus *new* faces are being viewed. For observers with good recognition skills, there will be relatively little overlap between the *old* and *new* distributions, and for observers with poorer recognition skills, there will be relatively more overlap between the distributions. The differences in d' yielded by different observers under identical experimental conditions are thought to reflect the *characteristics of the individual observers* such as their visual and perceptual abilities, memory capacity, motivation, and experience with the task.

The criterion measures the observer's level of conservativeness during the experiment for responding that he/she has seen faces previously. In other words, how familiar must a face in the experiment seem for the observer to be comfortable responding "known." Criterion is generally thought to reflect *both the characteristics of the individual observers and the characteristics of different situations.* The former include inherent aspects of the observer's personality such as the liberalness/conservativeness of guessing strategy, and the latter include aspects of the experimental situation, including task demands and context, such as the proportion of faces that are *actually* old versus new in recognition test.

Signal detection model for faces. When computing the measures d' and C for a particular face in a particular condition of a recognition experiment, data from many different observers are combined. Each hit that contributes to the hit rate and each false alarm that contributes to the false alarm rate comes from a different observer. All of the observers in the experiment who learned a particular face in the learning phase of a recognition experiment contribute to the *old* distribution for that face, and all of the observers who did not learn this face but see it in the recognition test contribute to the *new* distribution for the face. The dimension on which these observers are distributed is an indication of the level of familiarity people experience when looking *at the face in question*. For the face to be recognizable at a level above chance, observers who have seen the face before should generally experience higher levels of familiarity than observers who have not seen the face before. More formally, the distribution composed of the observers who have seen the face before (*old*) should not overlap completely with the distribution composed of the observers who have not seen the face before (*new*).

The *d′* is a measure of the overlap of familiarity levels experienced by the observers who have seen the face before and the distribution of observers who have not seen the face before—that is, its *recognizability*. For highly recognizable faces, there will be relatively little overlap between the *old* and the *new* distributions, whereas for less recognizable faces, there will be more overlap between the distributions. Differences in the *d′*s yielded by different faces under identical experimental conditions are thought to reflect the *characteristics of the individual faces*, including whether or not they have moles, buck teeth, and so on.

The criterion measures the tendency of the face to evoke *old* versus *new* responses from observers in a particular experiment. Criterion reflects *both the characteristics of the individual faces and the characteristics of the experimental situation*. An example combining both factors might be as follows. Male faces with long hair may evoke lots of *old* responses in a task in which they constitute 80% of the faces used, but may evoke many fewer *old* responses when they represent a small minority of the faces.

(Manuscript received July 17, 1996;
revision accepted for publication November 19, 1996.)