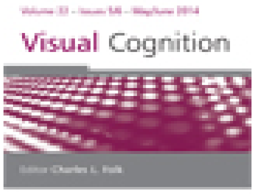

VISUAL COGNITION

ISSN: 1350-6285 (Print) 1464-0716 (Online) Journal homepage: https://www.tandfonline.com/loi/pvis20

# Categorical perception of newly learned faces

Paolo Viviani , Paola Binda & Thomas Borsato

**To cite this article:** Paolo Viviani , Paola Binda & Thomas Borsato (2007) Categorical perception of newly learned faces, VISUAL COGNITION, 15:4, 420-467, DOI: 10.1080/13506280600761134

**To link to this article:** https://doi.org/10.1080/13506280600761134

Published online: 13 Apr 2007.

VISUAL COGNITION, 2007, 15 (4), 420–467

Psychology Press
Taylor & Francis Group

# Categorical perception of newly learned faces

Paolo Viviani

*Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland, and Laboratory of Action, Perception and Cognition, Faculty of Psychology, UHSR University, Milan, Italy*

Paola Binda and Thomas Borsato

*Laboratory of Action, Perception and Cognition, Faculty of Psychology, UHSR University, Milan, Italy*

Five experiments investigated identification and discrimination of faces. Stimuli were blends of two faces generated with a morphing algorithm. Two same-gender and two different-gender pairs of faces were tested. Experiment 1 (identification) estimated the point of indifference along the morphing sequence, and the associated differential threshold. Experiment 2 (discrimination, ABX) demonstrated that novel faces are perceived categorically. Identity was a more important factor than gender in generating the perceptual categories. Experiment 3 and 4 (identification) demonstrated that categories are generated progressively in the course of the experiment and depend on the range of morphs tested in any one condition. Confidence ratings (Experiment 5) showed that the multidimensional space where faces are represented can be collapsed onto a single dimension. Response probabilities and response times for Experiments 1–4 were predicted simultaneously by a counting model postulating that quanta of discriminal information are sampled independently from the stimuli.

The phenomenon known as Categorical Perception (CP) refers collectively to any instance where the following conditions are met:

1. A class of physical stimuli is defined in some metric space with an adequate number of dimensions. For any two stimuli A and B, there is a

Please address all correspondence to Paolo Viviani, UHSR University, Faculty of Psychology, 58, via Olgettina 20132 Milan, Italy. E-mail: viviani.paolo@hsr.it

This study was partly supported by research funds to PV from the University of Geneva and from the UHSR University of Milan. Jerôme Guinot implemented the Loki algorithm. We are grateful to Bruno Rossion and to two anonymous reviewers for their constructive criticisms of a previous version of this paper. We are also grateful to our friends Christelle, Francesca, Maria, Henry, and Philippe for letting us morph away their identity.

http://www.psypress.com/viscog DOI: 10.1080/13506280600761134

connecting path such that each point C along the path also corresponds to a stimulus within the class. Moreover, for any such point, one can define an *objective* distance measure *d*(C,A) and *d*(C,B) from the path endpoints (A and B).

2. It is possible to estimate experimentally the *perceived* distance δ(C,A) and δ(C,B) of stimulus C from both A and B. As C moves along the path from A to B, δ(C,A) is a monotonously increasing function of *d*(C,A): δ(C,A) = *F*(*d*(C,A)).
3. There exist pairs A and B such that the derivative of *F* with respect to *d*(C,A) has a unique maximum somewhere along the path from A to B.

If CP occurs, the transformation that maps the stimuli into the perceptual space preserves the topological, but not the metrical properties of the objective space. Specifically, the transformation is such that, up to a point along the path from A to B, C is perceived as being closer to A than it really is. After that point, C is perceived as being closer to B than it really is. Equivalently, the phenomenon can be characterized as the result of competing "attractor fields" centred on A and B (Tanaka, Giles, Kremen, & Simon, 1998). CP is a graded rather than an all-or-none phenomenon, with the strength of the effect depending on the relative value of the maximum of the derivative of *F*. In the ideal case of "pure" CP, *F* is a step function, and its derivative is zero everywhere but in a single point, where it is infinite.

In essence, the function *F* is a classical psychophysical function relating perceived to objective value of the distance from one endpoint. Thus, it would be conceivable to address experimentally the study of CP with the standard techniques for estimating psychophysical functions (Guilford, 1954). In almost all cases, however, it has been found more expedient to take an indirect route, by measuring the effects of the warping of the representational space on identification and discrimination performances (cf. Harnad, 1987). With few exceptions—e.g., the multidimensional scaling approach adopted by Bimler and Kirkland (2001)—the experimental strategy for detecting CP has remained stable for almost 50 years. It consists in coupling two experiments. The first experiment adopts a typical *identification* task. First, the observers familiarize themselves with the endpoint stimuli A and B. Then, upon being shown an intermediate stimulus C, they must decide whether C is more similar to A or B. The results are collected as psychometric functions relating the probability *p*(B) of answering B to the objective distance *d*(C,A). By estimating the point along the path from A to B where *p*(B) = .5, one infers that, at that point, C is perceived as being equally distant from A and B (point of indifference). Contrary to what has often been claimed (see General Discussion), nothing can be inferred about CP directly from the psychometric function. However, *provided that CP occurs*, the point of indifference can be interpreted as the point where the path from A to B crosses the boundary between two regions of the representational space centred on A and B respectively. If the warping of the space is quite strong, any stimulus within either region will be difficult to discriminate from the respective centres, and the regions can be interpreted as fuzzy categories. The identification task provides the background for the second experiment, which involves a discrimination task, either in the ABX or in the "same–different" version. If CP occurs, pairs of stimuli that lay on opposite sides along the path with respect to the point of indifference should be easier to discriminate than pairs that lay on the same side, even though the distance between the stimuli is the same in both cases.

Historically, the first demonstration of CP was provided in the case of the sounds of language (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, Harris, Kinney, & Lane, 1957). The demonstration involved syllables such as /ba/ and /pa/, which are differentiated by a single physical parameter, the voice onset time. Later, CP was also found for hue (Bornstein, 1987; Bornstein & Korda, 1984; Raskin, Maital, & Bornstein, 1983) and musical intervals (Burns & Ward, 1978), which again vary along one-dimensional continua (wavelength and time). Because, at least the first two instances involve dedicated perceptual mechanisms, it has been suggested that CP reflects the evolutionary hardwired response to the need of focusing on differences that are relevant, by neglecting those that are not. In the case of language, later evidence (Kuhl, 1981; Sawusch & Gagnon, 1995) proved damaging to the hypothesis of an innate connection between CP and the phonetic system of natural languages. Actually, Massaro (1998) went as far as suggesting that CP is not the only, perhaps not even the correct way of interpreting the results in the case of speech. At the same time, research demonstrated that category boundaries emerge also between novel stimuli like musical intervals (Burns & Ward, 1978), textures (Pevtzow & Harnad, 1997), and complex unfamiliar shapes (Livingstone, Andrews, & Harnad, 1998).

Recently, a growing number of studies have addressed the issue of whether the phenomenon of CP applies also to faces. Faces are particularly interesting for several reasons. First, they are very complex stimuli varying along many relevant dimensions. Although this point has not drawn much attention, the high dimensionality of face space implies that there are infinite different paths leading from one face to another, and no obvious criterion for identifying the "shortest", more direct one. This may be important because virtually all studies on CP for faces generate intermediate stimuli with morphing techniques that select a unique path with *ad hoc* algorithms. While this path may be the shortest one in objective (pixel) space, there is no guarantee that it is also the shortest one when features rather then pixels are construed as relevant dimensions. Likewise, there is little to substantiate the implicit assumption that equal morphing steps in pixel space correspond to equal steps in feature space. Actually, Busey (1998) has cast serious doubts on both assumptions. By applying a multidimensional scaling technique to similarity judgements, this study represented a large array of face pairs as points in a multidimensional space. In the same space were also represented the morphed faces that, according to one popular algorithm (Morph™ by Gryphon Software), were halfway between each parent pair. In almost all cases, these morphs were quite distant from the hyperline connecting their parents. In future research this aspect of the generation of the stimuli should be more controlled, and described in grater detail.

The second, more important reason why faces are interesting is the paramount role that they play in social intercourse. Whereas the uniqueness of faces in terms of social and biological relevance is obvious, the extent to which such uniqueness is best characterized in terms of innateness, localization, or domain specificity (or any combination thereof) is still debated (for a review, see Liu & Chaudhuri, 2003). The tasks that the perceptual system is called to deal with differ in their degree of specificity. The most general problem is to decide whether a visual stimulus is indeed a face or something else (such as, for example, a Halloween's pumpkin). It has been argued that the fusiform gyrus in the medial occipitotemporal cortex is involved in this first step (Kanwisher, McDermott, & Chun, 1997; Kanwisher, Stanley, & Harris, 1999). However, the domain-specific role of this area has been debated (Kanwisher, 2000; Kanwisher & Moscovitch, 2000), one competing domain-general view being that its basic role is instead to individuate members of a homogenous class (not necessarily faces), who have the same components in the same spatial arrangement (Gauthier, Curran, Curby, & Collins, 2003; Gauthier & Logothetis, 2000; Gauthier et al., 2000).

The next step is to identify the face as belonging to a specific individual. Interestingly, face inversion, which does not interfere with face recognition, has a dramatically detrimental effect on face identification (Diamond & Carey, 1986; Yin, 1969), suggesting a crucial role of configurational factors. In fact, there is evidence that successful identification requires a processing stage that can be selectively disrupted. Some prosopagnosic patients who are unable to identify previously familiar faces preserve both the ability to perceive them as faces, as well as the ability to identify nonface objects. The converse syndrome has also been documented. Patient CK studied by Moscovitch, Berhmann, and Winocur (1997) is severely impaired at reading and object recognition, but can still identify faces.

Unless they have to, people rarely put up a "poker face". More often than not, the most specific task that we are confronted with when viewing a face is to identify its expression, which is a generally reliable indicator of the underlying mood or emotion. There is a vast and growing literature on the perception of facial expressions (for a review, see Adolphs, 2002). Here, we only need to stress that emotional facial expressions are just one component of a more general system of phasic, physiological changes that are innate and largely independent of cultural factors. Thus, unlike facial identity, they need not to be learned. Not surprisingly, there is clinical evidence that the ability to recognize an emotion may be impaired, whereas the ability to discriminate between faces is spared (Adolphs, Tranel, Damasio, & Damasio, 1994). There is also neuroanatomical evidence of a dissociation between areas coding face identity and areas coding for emotional expressions (Winston, Henson, Fine-Goulden, & Dolan, 2004).

CP for faces has been investigated at all levels of specificity outlined above. Evidence for CP has emerged in the case of facial expressions of emotions (Bimler & Kirkland, 2001; Calder, Young, Perret, Etcoff, & Rowland, 1996; de Gelder, Teunisse, & Benson, 1997; Etcoff & Magee, 1992; Pollak & Kistler, 2002; Suzuki, Shibui, & Shigemasu, 2004; Young et al., 1997), facial expressions related to language (Campbell, Woll, Benson, & Wallace, 1999), and in the presence of superordinate categories such as race (Levin, 1996; Levin & Angelone, 2002; Levin & Beale, 2000) and gender (Campanella, Chrysochoos, & Bruyer, 2001; see, however, Bülthoff & Newell, 2000). Whether CP exists also for within-category (i.e., same gender/same race) face identity is more controversial. The earliest study (Beale & Keil, 1995) reported CP for pairs of faces of very well-known people (e.g., President Kennedy and President Clinton), but not for faces unfamiliar to the observers before the experiment. Later, the same conclusion was reached by Angeli, Davidoff, and Valentine (2001) and Campanella et al. (2001, Exp. 3). Instead, both Campanella, Hanoteau, Seron, Joassin, and Bruyer (2003, Exp. 1) and Levin and Beale (2000) concluded that CP is present also for unfamiliar faces. Levin and Angelone (2002) reported an interaction between race and identity, with race enhancing a weaker CP effect already present in same-race pairs. McKone, Martini, and Nakayama (2001) did find CP even in the presence of noise, but the observers had a very large amount of practice (up to 3000 trials) with the face stimuli. Likewise, Stevenage (1998) after extensive training with pictures of twin sisters, obtained both a compression effect (same-twin pairs judged more similar after than before training) and an expansion effect (different-twin pairs judged more dissimilar after than before training). Finally, two brain-imaging studies (Rossion, Schiltz, Robaye, Pirenne, & Crommelinck, 2001; Rotshtein, Henson, Treves, Driver, & Dolan, 2005) also claimed to have detected CP for familiar and unfamiliar faces. However, we shall argue in the General Discussion that their claims are not adequately substantiated.

This mixed pattern of results suggests that the distinction between familiar and unfamiliar faces may not be clear-cut. Even if the face pairs are completely novel to the observers before the experiment, some amount of familiarization takes place during the experiment itself. If so, the results of McKone et al. (2001) and Stevenage (1998), both obtained with extensive training, would demonstrate that CP occurs not only with between-category stimuli, but can be induced also when the stimuli do not have a clearly identified, memorized identity.

The primary goal of the five experiments reported here is to test the hypothesis that CP can indeed arise for novel faces, and to document its development in the course of the experiment. Second, we want to assess the relative weight of facial features and gender in determining the sharpness of category boundaries. Third, we want to demonstrate that the boundary between categories is flexible and depends on the range of images shown in the course of the experiment. Finally, we want to test the assumption that, whatever the dimensionality of the perceptual space, the path connecting two faces can be collapsed onto a single one-dimensional axis. With the exception of multidimensional scaling, all experimental techniques for measuring the identification performance assume the validity of this assumption (Guilford, 1954). Yet, to our knowledge, the assumption has never been put to direct test in the case of faces. Experiment 1 (identification) and Experiment 2 (discrimination) establish the presence of CP for novel faces both within and between gender. Experiments 3 and 4 investigate the flexibility of the category boundaries. Experiment 5 deals with the dimensionality issue. Formal models of both the identification and discrimination performance are formulated and validated in order to demonstrate the coherence of the results across experiments.

## GENERAL METHODS

The experiments were conducted simultaneously at the UHSR University in Milan and at the University of Geneva with the same experimental set up. Participants seated in a quiet, isolated room kept in dim light in front of a computer screen (21 inch; resolution: 1024 × 768 pixels; refresh rate: 75 Hz), at a distance of about 60 cm (at this distance 1 cm on the screen corresponds to about 1 degree of visual angle). Responses were entered through the keyboard (see below). The computer also recorded response times (RTs) with a 1 ms accuracy. The experiments were self-paced, each trial starting 500 ms after recording the previous response. The total duration of a session varied as a function of the experimental schedule (see below). Participants could interrupt the session for a short rest by pressing, after a stimulus had been presented, the “pause” key instead of the response key. That trial was inserted again in a random position within the remaining sequence of trials. Sessions began with a verbal description of the task, and by a warm-up phase of at least 20 trials. All experimental protocols were approved by the Ethical Commission of the UHSR University and of the University of Geneva. Participants gave their informed consent.

## Stimuli

Stimuli were frontal views of human faces generated by the LOKI morphing algorithm that we developed and implemented for the purpose of the experiments. As in other commercially available software, the algorithm generates a sequence of intermediate, equispaced images between two pictures (templates). First, one identifies the same set of salient facial features in both faces (landmarks). Then, a unique topological correspondence between templates is established by calculating the Delaunay triangulations (Okabe, Boots, Sugihara, & Chiu, 2000) of the two sets of landmarks (Figure 1). The morphing transformation acts on each triangle following a user-defined function. Different functions can be defined for any subset of landmarks. The number of steps in the transformation is limited only by the available computer memory. In this application, we applied the same linear transformation to all landmarks. By definition, the distance between morphing steps in the multidimensional pixel space is constant. It should be stressed that LOKI, unless other morphing algorithms—such as Morph$^{TM}$—makes sure that the sequence of morphs remains close to the hyperline in pixel space that connects the templates.

Five templates were used. They were high-resolution digital colour photographs of three females ([$F_1$, $F_2$, $F_3$]) and two males ([$M_1$, $M_2$]) against a black background (Figure 2, upper row). Models wore black bonnets and turtlenecks concealing hair and dressing details. We processed the raw pictures with Photoshop CS to equalize overall luminance, contrast, and chromatic spectrum. Four pairings of the templates were selected for the experiments: [$F_1$, $M_1$]; [$F_1$, $M_2$]; [$M_1$, $M_2$]; [$F_2$, $F_3$]. Figure 2 (lower panels) shows an equispaced sample of 12 pictures (including the templates) from the morphing sequence [$F_1$, $M_1$]. In all experiments stimuli were actually drawn from the same sequence of 61 pictures. However, for each experiment, we selected from this base sequence a different subset of adjacent stimuli (see later). In all cases, participants were kept unaware of the relative location of the subset within the sequence. During the experiments, participants were never exposed to the 12 initial and to the 12 final morphing steps.

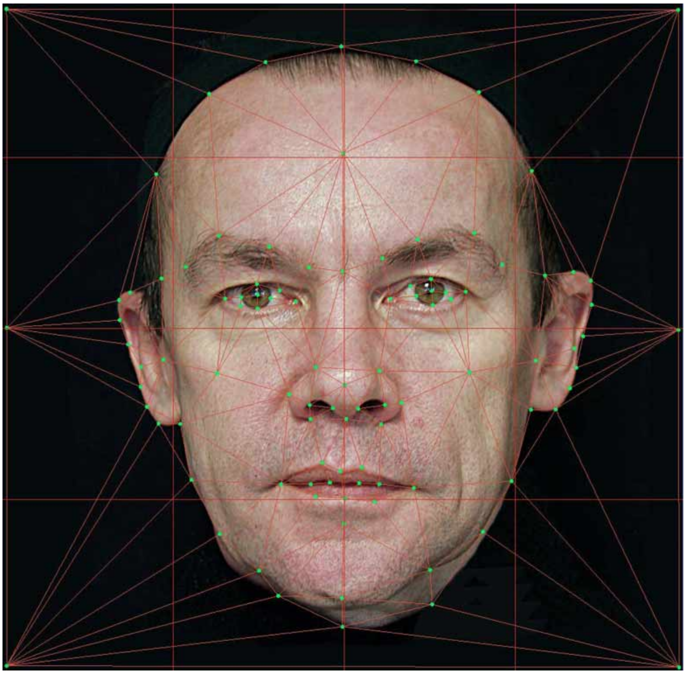

**Figure 1.** Example of the meshes used for implementing the morphing algorithm. Two sets of points are used to identify the corresponding features of the two faces. Each set of points is connected by a Delaunay triangulation, which is known to provide the optimal solution to the problem of interpolating a two-dimensional landscape f(x,y) to a finite set of samples. Because Delaunay triangulations are uniquely defined, the tessellations of the two images are topologically identical. [Colour versions of all figures are available online].

## EXPERIMENT 1

The goal of this experiment was to investigate face identification, both within and across gender, by estimating the point of indifference (median) and the differential limen (JND) of the psychometric function. The results of this experiment lay the ground for Experiment 2, which addresses more directly the issue of categorical perception.

### Methods

*Participants.* Twenty young individuals participated to the experiment (age range: 20–28 years). All participants had normal or corrected-to-normal vision. Half of the participants were students of the UHSR University of Milan. The other half were students of the University of Geneva. In both cases participants received one course credit.

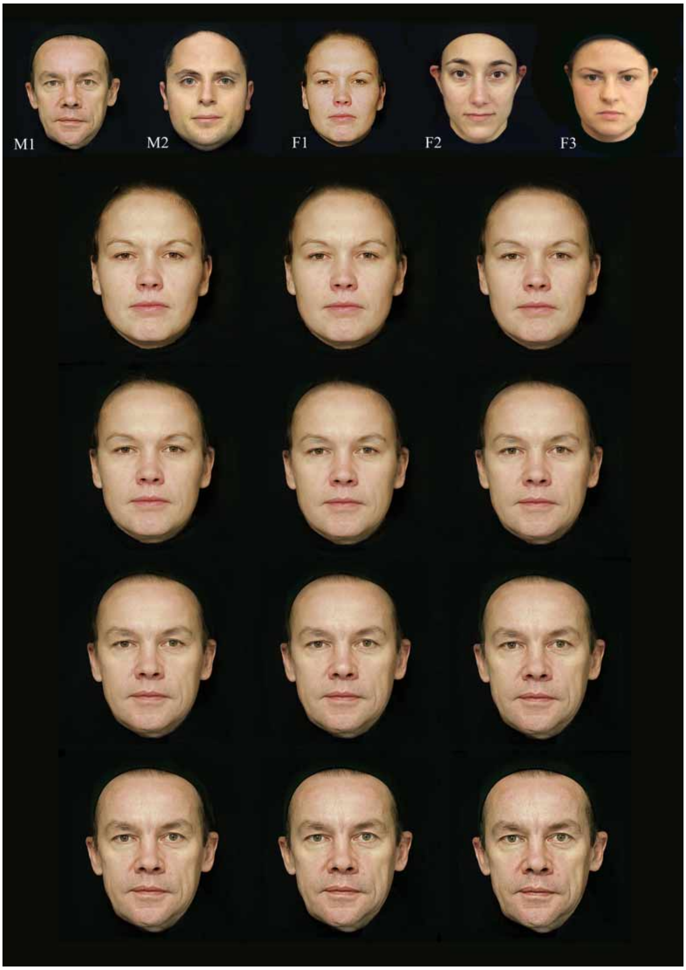

**Figure 2.** Stimuli. Upper row: The five faces used as templates. Lower panel: Twelve equispaced frames in the morphing sequence from $F_1$ to $M_1$.

*Stimuli.* The stimuli were 17 contiguous pictures, from rank order 20 to 36 within the 61-element morphing sequences [$F_1$, $M_1$], [$F_1$, $M_2$], [$M_1$, $M_2$], and [$F_2$, $F_3$]. Each stimulus was presented 50 times. The order of presentation was randomized with the constraint that successive stimuli had to be at least three steps away in the morphing sequence. In a session $17 \times 50 = 850$ stimuli were presented.

*Task and procedure.* We adopted a classical identification task. For each morphing sequence, sessions began by introducing the corresponding pair of

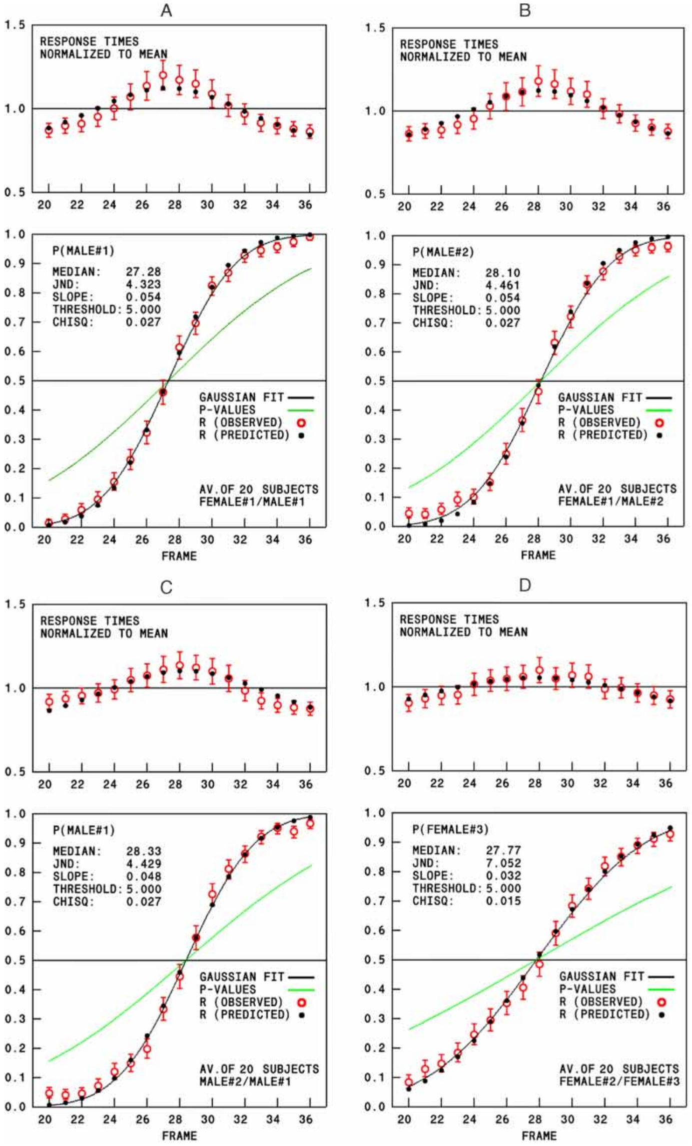

**Figure 3.** Experiment 1: Identification. Panels A to D summarize the results for the indicated pair of templates. Lower plot: psychometric function (empty circles) relating the frame rank order in the morphing sequence to the probability of identifying the stimulus with Template B. Bars are the 95% confidence intervals computed with the exact binomial theory. The continuous black line through the data points is an empirical Gaussian fitting used for estimating the median and the JND of the distribution. The predictions of the quantal model (solid dots) correspond to the indicated $p$-value function (light grey line). Upper plot: Response times normalized to the population mean (empty circles). Bars are the 95% confidence intervals of the means. Solid points: Predictions of the quantal model.

templates identified as Face A and Face B. Participants explored both faces for as long as necessary to memorize fully their peculiar features. To this end, they used the spacebar to switch back and forth between templates. On average, each face was displayed 10 times for 10 s. In each trial one stimulus was displayed, and participants had to decide (forced choice) which template the stimulus was more similar to. Responses were entered with both hands, by pressing any key between F1 and F4 for Face A, and any key between F9 and F12 for Face B. The stimulus remained visible for 3 s, and we encouraged participants to respond as soon as possible. However, the answer could be given either during the presentation, or after the stimulus had disappeared. Response times were measured from the onset of the stimuli. On average, a session lasted between 30 and 40 min. All participants were tested four times over a period of time ranging from 1 to 3 weeks. The order of the sessions—one for each pair of templates—was counterbalanced across participants.

## Results

Figure 3 summarizes the results of the experiment. The lower plot in each panel reports the psychometric function relating the rank order of the stimuli within the morphing sequence to the relative frequency $F_B$ of identifying the stimulus as Template B.

Data points (empty circles) are the observed response frequencies over all participants. Bars are the 95% confidence intervals of the observed frequencies computed with the exact binomial theory (Sachs, 1984, pp. 333–337). For all but one pair ([$F_2$, $F_3$]), the range of stimuli was sufficiently large to include stimuli that were almost always identified correctly. Psychometric functions are characterized qualitatively by the median ($F_B = .5$) and the JND, both computed from the best-fitting Gaussian interpolation of the data points (continuous black lines). Comparing these estimators across pairs of templates shows that: (1) In all cases, the point of indifference is roughly in the middle of the range of variation of the stimuli; and (2) the discriminating power depends on the pair of templates, being lower (higher JND) for the pair [$F_2$, $F_3$] than for all other pairs. Note that specific features of the faces, rather than gender seem to be responsible for this difference.

The upper plots of Figure 3 summarize the response time data. Data points are mean RTs over all participants, normalized to the general mean for all stimuli. Bars are 95% confidence intervals of the means (normal approximation). The mean RTs (over ranks) were: [$F_1$, $M_1$]: 961(±49) ms; [$F_1$, $M_2$]: 981(±59) ms; [$M_1$, $M_2$]: 935(±49) ms; [$F_2$, $F_3$]: 1021(±58) ms. Statistical analysis (ANOVA, 4[face pair] × 17[rank], repeated measure;

Greenhouse-Geisser correction) detected no significant difference among mean RTs, $F(3, 57) = 0.992$, $p = .384$. Instead, there was a highly significant effect of the stimulus rank, $F(16, 304) = 40.216$, $p < .0001$, as well as a significant interaction, $F(48, 912) = 2.729$, $p = .011$. For all pairs of templates RTs were symmetrically bell-shaped with a maximum close to the point of indifference: Quadratic regression term, $F(1, 19) = 65.887$, $p < .0001$; order 4, $F(1, 19) = 40.102$, $p < .0001$; order 6, $F(1, 19) = 17.249$, $p = .001$.

As noted in the introduction, categorical perception in identification task is best defined in terms of psychophysical functions. Specifically, let us assume that answers to stimulus $S_k$ are based on a comparison between the strengths of Templates A and B in the mixture defining that stimulus. By definition, these strengths are inversely proportional to the distances of $S_k$ for A and B. Then, any evidence of a steep slope near the point of indifference in the psychophysical function between the objective and perceived distance of $S_k$ from one template, would constitute evidence for categorical perception. Therefore, in order to use the psychometric functions for testing the hypothesis, it is necessary to make explicit assumptions on the underlying psychophysical function. The next section introduces a formal model of the mechanism that turns stimulus strength into a psychophysical variable, and this variable into a response.

*A counting model for the identification task.* In our task a stimulus $S_k$ is the image with rank order $k$ in the M-step morphing sequence from Template A to Template B. The model (Figure 4) assumes that the perceiver samples independently the image at a regular rate $R_s$, each sampling yielding only one quantum of discriminal information. There are two types of quanta: $Q_A$ and $Q_B$ favouring A and B, respectively. The perceiver counts separately A- and B-type quanta, and identifies $S_k$ with A or B as soon as the total number of either type reaches a constant criterion threshold $T$. The probability $p = p(k)$ of acquiring a quantum $Q_B$ (heretofore, *p-value function*) increases monotonically as the stimulus in the morphing sequence approaches Template B (the probability of acquiring a quantum $Q_A$ is $q = 1 - p$). In essence, the $p$-value function is a psychophysical function mapping the objective distance of $S_k$ from B into a perceived distance. In Appendix 1 we derive the exact expression for the probabilities of identifying stimulus $S_k$ with either templates ($P_A$ and $P_B$), and for the associated response times RT.

The experimental data (cf. Figure 3) consist of the relative frequency $F_B(k)$ of Response B, and of the mean response time (in ms) RT($k$) as a function of the position $k$ of the stimulus $S_k$ in the morphing sequence. To fit the model to the data, one must insert in the expressions for $P_B(p)$ and $N_Q(p)$ the $p$-value function $p = p(k)$, which describes how the sampling probability for $Q_B$ increases when the stimulus rank order ranges between $k = 1$ and $k = \mathrm{M}$. We adopt the following description of the psychophysical function $p$:

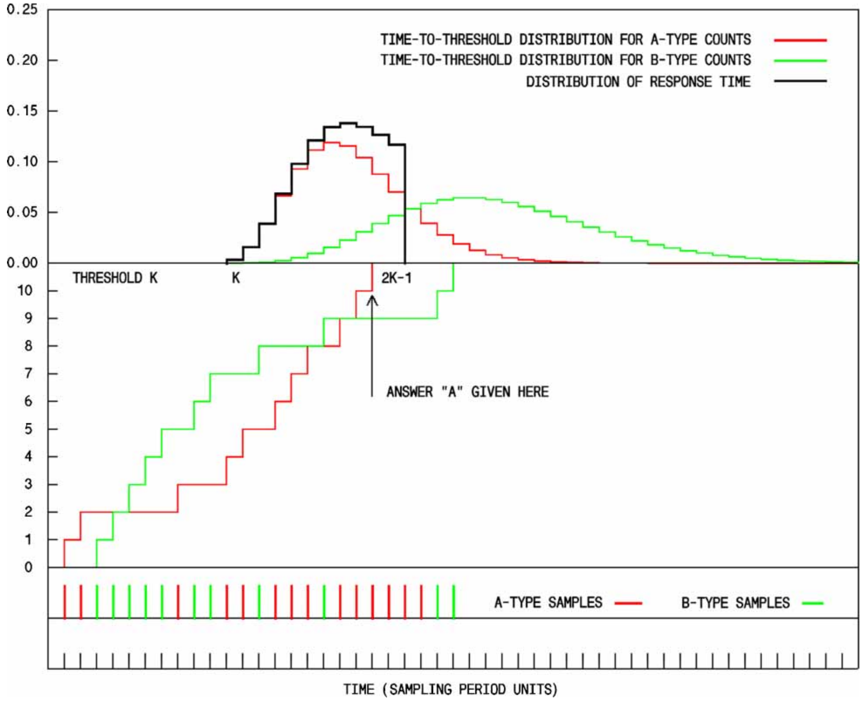

**Figure 4.** Quantal model for the identification task. Lower panel: Time evolution of the cumulative number of A- and B-type quanta (medium and light grey lines, respectively) sampled from a stimulus in the morphing sequence (abscissa: Time; ordinate: Number of quanta). In this simulated run, we set the probability of sampling an A-type quantum to $p_A = .6$, and the response threshold to $T = 11$. Upper panel: Probability density functions for the number of samples before reaching threshold for the two types of quanta (medium and light grey lines), and probability density function for the total number of samples before either cumulative sum reaches threshold (thick black line; abscissa: Number of quanta; ordinate: Probability of reaching the threshold).

$$p(k) = \sum_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(i-\mu)^2}{2\sigma^2}} [1 \leq k \leq N]$$

i.e., the discrete approximation to a cumulative Gaussian function with parameters μ and σ. The parameter μ sets the point along the morphing sequence that is perceived as being equally distant from A and B. The parameter σ is inversely proportional to the slope of $p(k)$ at $k = \mu$, where the slope is maximum. Within the interval of $k$-values for which response frequencies are available, relatively large values of σ result in an almost linear increase of $p$ as a function of $k$. Conversely, smaller values of σ result in an increasingly marked nonlinearity of the relation between $k$ and $p$. Thus, σ estimates the extent to which categorical perception distorts the perceptual continuum between the Templates A and B. The details of the fitting strategy are reported at the end of Appendix 1.

The best-fitting predictions of the model (solid dots in Figure 3) are in excellent agreement with the data for both response probabilities and response times. With only two free parameters, the model captured accurately the covariation between the identification accuracy, estimated by the slope of the psychometric function, and the increase of the response times near the point of indifference. Note also that the same value of the threshold ($T = 5$) was optimal for all four pairs of templates. As described above, the model captures the degree of categorical perception through the nonlinearity of the $p$-value functions (light grey lines). On this basis, it appears that the phenomenon is clearly present for the two pairs [$F_1$, $M_1$] and [$F_1$, $M_2$] (different gender), slightly less marked for the pair [$M_1$, $M_2$], and almost inexistent for the pair [$F_2$, $F_3$].

Inserting the best-fitting threshold value in the theoretical expression of the average number of samples accumulated before a response (see Appendix 1), one obtains $\overline{N}_Q = 6.45$. Thus, the average sampling rate $R_s = \overline{N}_Q / RT_{mean}$ for the four pairs was fairly similar [$F_1$, $M_1$]: 7.4 sample/s; [$F_1$, $M_2$]: 7.3 sample/s; [$M_1$, $M_2$]: 7.6 sample/s; [$F_2$, $F_3$]: 6.9 sample/s.

# EXPERIMENT 2

Experiment 1 provided some indirect evidence of categorical perception for three pairs of templates. The goal of this experiment was to confirm more directly the presence of the phenomenon by analysing the performance in a discrimination task.

## Methods

*Participants.* We tested the same 20 individuals who had participated to Experiment 1. Also for this experiment participants received one course credit.

*Stimuli.* We adopted the ABX version of the discrimination task. In each session the stimuli were 13 triples of pictures ($S_A$, $S_B$, and $S_X$) drawn from the same 61-element morphing sequences [$F_1$, $M_1$], [$F_1$, $M_2$], [$M_1$, $M_2$], and [$F_2$, $F_3$] already used in Experiment 1. $S_A$ and $S_B$ were always at a fixed rank distance of 8 along the sequence. The rank order of the first picture ($S_A$) ranged from 18 to 30, that of the second one ($S_B$) from 26 to 38. Picture $S_X$ was the same as either $S_A$ or $S_B$. In one trial the three pictures were presented sequentially, $S_A$ or $S_B$ for 1000 ms each, and $S_X$ for a maximum period of 3000 ms. The interstimulus interval was 300 ms. Each of the four possible combinations ($S_A$,$S_B$,$S_A$), ($S_A$,$S_B$,$S_B$), ($S_B$,$S_A$,$S_A$), and ($S_B$,$S_A$,$S_B$) was presented 10 times in a different pseudorandom order for each participant, with the constraint that the first pictures in successive trials had to be at least three positions away. The total number of trials in one session was 13 (initial position) × 40 (repetition) = 520.

*Task and procedure.* As in Experiment 1, sessions began by introducing the corresponding pair of templates identified as Face A and Face B. Participants explored again both faces with the same modality of Experiment 1. Participants were informed that the third stimulus ($S_X$) was always the same as either the first or the second. In each trial, participants had to decide (forced choice) which of the two possibilities had occurred. Response were entered with both hands, by pressing any key between F1 and F4 if $S_X$ was equal to the first picture, and any key between F9 and F12 if $S_X$ was equal to the second picture. The third stimulus disappeared as soon as the answer was entered. Participants were encouraged to respond as quickly as possible. In all but a few exceptional cases, responses were entered before the third stimulus had disappeared. Response times were measured from the onset of the third stimulus $S_X$. A new trial began 500 ms after entering the answer to the previous one. On average, a session lasted between 40 and 50 min. The four sequences [$F_1$, $M_1$], [$F_1$, $M_2$], [$M_1$, $M_2$], and [$F_2$, $F_3$] were tested in separate sessions over a period of 1–3 weeks. The order of the sessions was counterbalanced across participants. The experiment took place after completing Experiment 1.

## Results

Figure 5 summarizes the results of the experiment. The lower plot in each panel reports the probability of a mistaken identification of the third picture ($S_X$) as a function of the midpoint $k$ between the rank orders of the first two pictures ($S_A$ and $S_B$). Thus, for instance, responses for $S_A = 22$ and $S_B = 30$ are displayed as frame rank order $k = 26$. As in Experiment 1, probabilities (empty circles) were estimated by pooling the responses for all participants, and bars encompass the 95% confidence interval of the means. The average error rates for the four face pairs were: [$F_1$, $M_1$]: .211; [$F_1$, $M_2$]: .209; [$M_1$, $M_2$]: .220; [$F_2$, $F_3$]: .297. Statistical analysis (ANOVA, 4[face pair] × 13[midpoint rank], repeated measure; Greenhouse-Geisser correction) showed a significant difference among pairs, $F(3, 57) = 6.425$, $p = .002$, the error rate for [$F_2$, $F_3$] (i.e., the pair for which the psychometric function was most shallow, see Figure 5D) being higher than that for each of the other three pairs. The performance never reached chance level, some discriminating power remaining even when $S_A$ and $S_B$ were positioned at the extreme of the

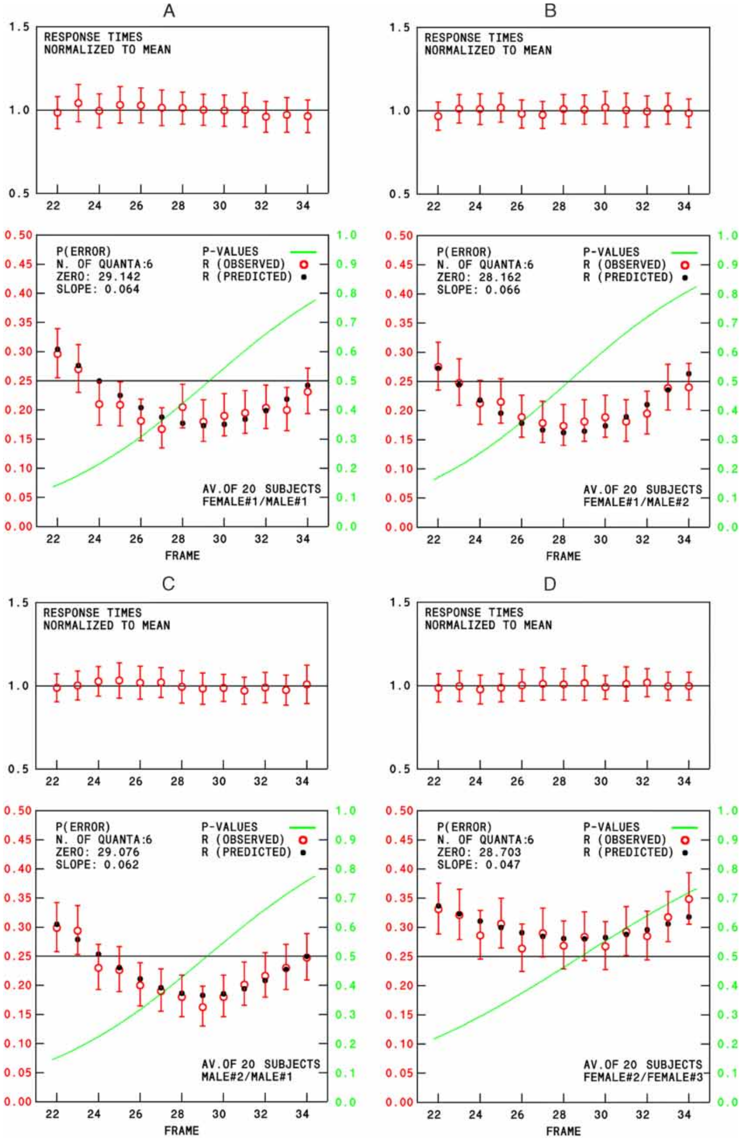

**Figure 5.** Experiment 2: ABX discrimination task. Panels A–D summarize the results for the indicated pair of templates. Lower plot: Probability of incorrect discrimination (empty circles) as a function of the midpoint of the 9-frame interval between stimuli A and B. Bars around the data points are the 95% confidence intervals computed with the exact binomial theory. The predictions of the quantal model (solid dots) correspond to the indicated $p$-value function (light grey lines) . For all pairs of templates, the error rate is minimum when the A–B interval straddles the same frame rank order for which the identification performance is at chance level (cf. Figure 4). Upper plot: Response times normalized to the population mean (empty circles). Bars are the 95% confidence intervals of the means.

tested range. More importantly, the effect of the midpoint rank was also highly significant, $F(12, 228) = 12.375$, $p < .001$. There was no interaction between factors, $F(36, 684) = 0.853$, $p = .582$. For all face pairs, the error probability dropped significantly near the middle of the tested range of stimuli: Quadratic regression term, $F(1, 19) = 91.042$, $p < .001$, the minimum being reached almost exactly at the point of indifference found in Experiment 1. Thus the results provide clear evidence of categorical perception for all pairs of faces, including the one ([$F_2$, $F_3$]) for which no such evidence emerged from the identification task. The upper plot in each panel reports the response times normalized as in Experiment 1. Mean RTs over midpoint rank were: [$F_1$, $M_1$]: 1097(±69) ms; [$F_1$, $M_2$]: 957(±54) ms; [$M_1$, $M_2$]: 983(±60) ms; [$F_2$, $F_3$]: 992(±49) ms (grand mean: 1007 ms). There was a significant difference among face pairs, $F(3, 57) = 4.555$, $p = .006$. Instead, RTs were unaffected by the position of $S_A$ and $S_B$ along the morphing sequence, $F(12, 228) = 1.338$, $p = .224$.

We have analysed the results of the identification task (Experiment 1) within the framework of a model of the relationship between the objective and perceived distance of the mixtures from the templates. We tested whether the same intuition underlying that model is able to capture also the results of the ABX task. The next section describes how the model was adapted to predict the error probabilities in Experiment 2.

*Evolution of categorical effects.* We investigated whether the higher discriminability of A,B pairs that straddle the category boundary depends on learning. The analysis was performed separately for each pair of templates, by pooling the results of all participants. For each trial rank order within an experimental session, we computed two average error rates. The first average included the responses for the four A–B intervals [18–26], [19–27], [29–37], and [30–38] whose midpoints (i.e., 22, 23, 33, and 34) were most distant from the middle ($k = 28$) of the tested range of stimuli. The second average included the responses for the three A–B intervals [23–31], [24–32], and [25–33] whose midpoints 27, 28, and 29 flanked and included the middle of the range. Figure 6 plots the difference between the first and the second average for trial rank orders ranging from 150 to 520 (the results for the first 150 trials were too scattered). The final difference between the error rates at the extreme and in the middle of the range of stimuli (categorical effect) depended on the pair of templates being tested, being maximum for the pair [$M_1$, $M_2$] and minimum for the pair [$F_2$, $F_3$]. However, in all cases the difference increased in the course of the session, suggesting that the categorical effect builds up progressively as the stimuli become increasingly familiar.

*A version of the counting model for the ABX task.* A stimulus in the ABX task is a triple of pictures $\{S_A = S_k,\ S_B = S_{k+\Delta},\ S_X\}$, extracted from the sequence $S_1 \ldots S_M$. $\Delta$ is a positive or negative integer whose absolute value is kept constant throughout the experiment (if $\Delta > 0$, $[1 \leq \Delta \leq M-k]$; if $\Delta < 0$, $[-\Delta + 1 \leq k \leq M]$). $S_X$ is always equal to either $S_A$ or $S_B$. As in the identification task, the model assumes that images are sampled independently at a constant rate $R_s$. $S_X$ was shown for 3000 ms. However, the mean RT was also very close to the duration of presentation of both $S_A$ and $S_B$ (see above). Thus, we also assume a constant presentation time of 1000 ms for all three stimuli. During the presentation interval, the observer counts separately the number $N_A$, $N_B$, and $N_X$ of B-type quanta sampled from $S_A$, $S_B$, and $S_X$, respectively (because the total number of samples is constant, it is irrelevant which type of quanta are reckoned). We postulate the simple response rule:

$$\text{If } N_X - N_A < N_X - N_B \rightarrow \text{“}S_A\text{”}.$$

$$\text{If } N_X - N_A > N_X - N_B \rightarrow \text{“}S_B\text{”}.$$

If $N_X$-$N_A = N_X$-$N_B$ → randomly “$S_A$” or “$S_B$” with equal probability.

Based on these assumptions, the quantal model predicts the probability $P_E$ of an incorrect discrimination. Appendix 2 describes the derivation of the prediction and the strategy for model fitting. The best-fitting predictions of the model (solid dots in Figure 5) are in good agreement with the error probabilities. The optimal value of the parameter $N$ was the same ($N = 6$) for all four pairs of faces. Thus, the rate at which quanta are supposed to be sampled is the same ($R_s = 6$ samples/s), independently of the pair of faces that are tested.

## EXPERIMENT 3

Together, the results of the first two experiments demonstrated a significant anisotropy in the perceptual space within which faces are represented. The results, however, do not permit one to decide whether the anisotropy, which is most evident near the middle of the morphing sequences, depends only on the weighing of the templates in each image per se, or, more generally, on the entire set of stimuli presented in a session. In Experiment 3 we address this issue by comparing the results obtained with different stimulus ranges within the morphing sequence. The experiment adopted again the identification paradigm (Experiment 1), which affords a precise estimation of the point of subjective indifference.

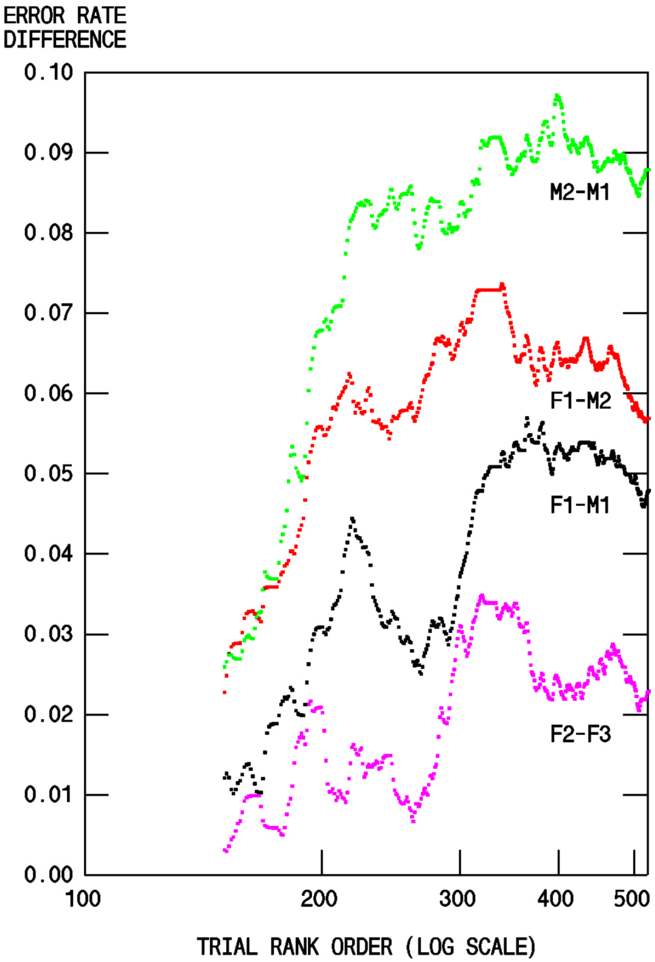

**Figure 6.** Experiment 2: ABX discrimination task. Evolution of categorical perception. For all indicated pairs of templates, data points describe the evolution along the experiment of the difference between the average error rate for the four outermost positions of the $S_A - S_B$ interval (corresponding to frames 22, 23, 33, and 34 in Figure 5), and the average error rate for the three central positions (corresponding to frames 27, 28, and 29 in Figure 5). Averages for trial rank order less than 150 were based on too few responses to be meaningful.

## Methods

*Participants.* Fourteen young individuals participated to the experiment (age range: 21–27 years). All participants had normal or corrected-to-normal vision. Half of the participants were students of the UHSR University of Milan. The other half was students of the University of

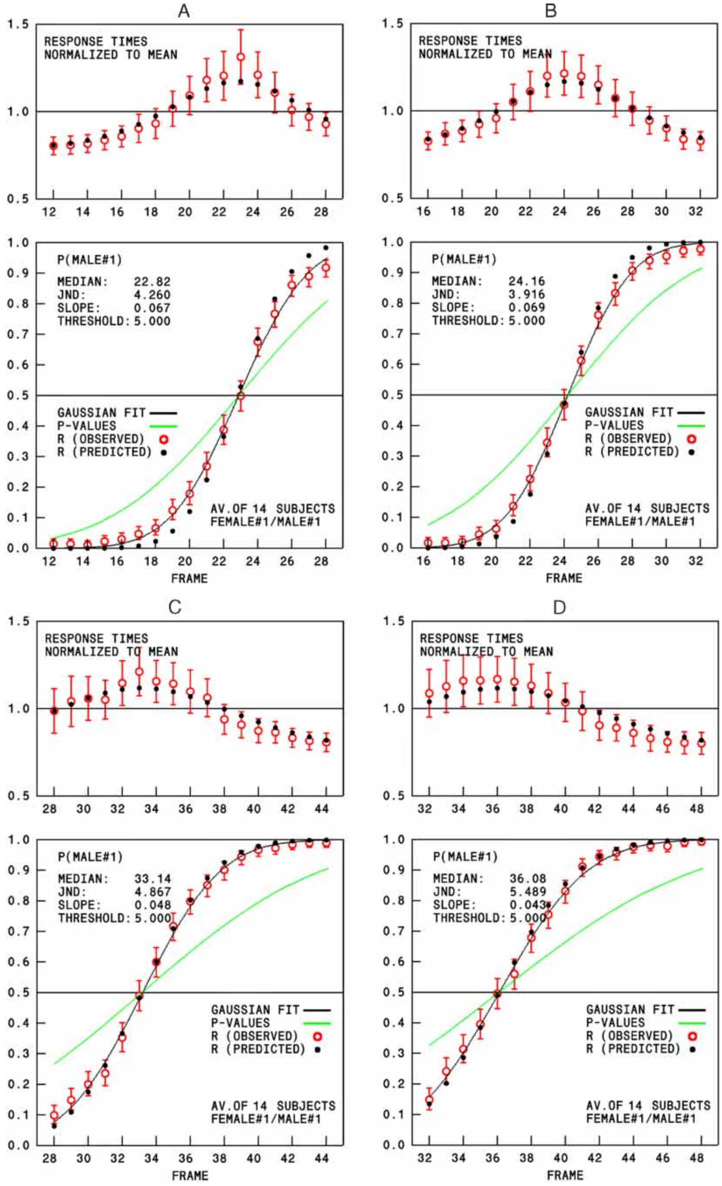

**Figure 7.** Experiment 3: Context effects in the identification task. Psychometric functions and response times for the [$F_1$, $M_1$] template pair (same format as in Figure 3). In separate sessions, the same sample of participants was tested with a different range of frames within the morphing sequence (A: [12–28], B: [16–32], C: [28–44], D: [32–48]). The frame for which the identification performance dropped at chance level (median) tended to move towards the middle of the range.

Geneva. None of them had participated to the first two experiments. In both cases participants received one course credit.

*Stimuli.* The stimuli were 17 contiguous pictures drawn from the morphing sequence for the pair of templates [$M_1$, $F_1$]. We tested four ranges of ranks (12–28, 16–32, 28–44, 32–48). The number of repetitions and the randomization procedure were the same as in Experiment 1.

*Task and procedure.* The task and the procedure were the same as in Experiment 1. The four ranges of ranks were tested in separate sessions over a period of time varying from 1 to 3 weeks. The order of the sessions was counterbalanced across participants.

## Results

Figure 7 shows, with the same conventions of Figure 4, the response probabilities and the response times for the indicated ranges of ranks. The range of variation of the stimuli had a profound effect on the psychometric functions. Table 1, which also includes data from Experiments 1 and 4, summarizes this effect. The point of indifference tended to move *pari passu* with the midpoint of the range. In fact, the whole psychometric function was shifted along the sequence. The differential threshold (JND) also varied with the range, being minimum for the range [16–32]. The mean RTs were quite similar [12–28]: 1080($\pm$47) ms; [16–32]: 1109($\pm$53) ms; [28–44]: 1016($\pm$62) ms; [32–50]: 1115($\pm$90) ms. Statistical analysis (ANOVA, 4[range] × 17[rank], repeated measure; Greenhouse-Geisser correction) detected no main effect of the range, $F(3, 39) = 0.843$, $p = .454$. Instead, the rank order had a significant effect, $F(16, 208) = 8.807, p < .0001$. As in Experiment 1, the RT values had a clear maximum, which moved along with the median of the corresponding psychometric functions. As signalled by the significant interaction between the range and rank factors, $F(48, 624) = 8.556$,

TABLE 1
Experiments 1, 3, and 4: Median and JND of the psychometric curve for the pair of templates [$F_1$, $M_1$] as a function of the midpoint of the range of tested stimuli

| *Range* | *Midpoint* | *Median* | *JND* |
|---|---|---|---|
| [12–28] (Exp. 3) | 20 | 21.82 | 4.26 |
| [16–32] (Exp. 3) | 24 | 23.16 | 3.92 |
| [20–36] (Exp. 1) | 28 | 27.28 | 4.32 |
| [24–40] (Exp. 4) | 32 | 30.93 | 4.27 |
| [28–44] (Exp. 3) | 36 | 33.14 | 4.88 |
| [32–48] (Exp. 3) | 40 | 36.08 | 5.49 |

$p < .0001$, the shape of the RT curves depended on the range. Separate ANOVAs showed that the curve for the [16–32] range was symmetrically bell-shaped: Quadratic regression term, $F(1, 13) = 43.972, p < .0001$; order 4, $F(1, 13) = 40.705$, $p < .0001$, whereas the curve was markedly asymmetrical for [12–28]: Linear term, $F(1, 13) = 23.460$, $p < .0001$; cubic term, $F(1, 13) = 16.885, p = .001$; [28–44]: Linear term, $F(1, 13) = 9.877, p = .008$; cubic term, $F(1, 13) = 4.726$, $p = .049$; and [32–48]: Linear term, $F(1, 13) = 10.810$, $p = .006$; cubic term, $F(1, 13) = 6.957, p = .020$. The quantal model developed for Experiment 1 was again fitted to these data. In all cases the model (black dots) interpolated quite satisfactorily the data points for both response probabilities and response times. By construction, the fit was obtained by adopting different $p$-value functions for each range. The parallel shift of the medians and of the RT curves strongly suggest a perceptual effect. Indeed, a mere shift of the median could also be interpreted as a biased response strategy induced by the implicit assumption that the subset of stimuli was always centred with respect to the templates. However, such a response bias would not explain why RTs should peak exactly in correspondence with the medians.

The shift of the psychometric function with the range of tested ranks set in progressively as a result of a process of adaptation. Figure 8 plots the evolution along the experiment of the median of the psychometric function for the indicated rank ranges (including data from Experiment 1 and 4). Medians were estimated by fitting a linear regression to the $z$-transform of the response probabilities (probabilities estimated by pooling the results of all participants). Reliable estimates of the medians were only possible for trial rank orders greater then 20. However, the results show clearly that the final values of the medians (Table 1) were indeed reached at the end of a progressive divergence from a common origin (approximately, around frame 29).

## EXPERIMENT 4

Experiment 3 demonstrated that judgements depend on the entire range of images presented in a session. However, the demonstration involved only one pair of templates, and was obtained with a different sample of participants from the one tested in Experiments 1 and 2. Experiment 4 is a control in which a subset of the population of participants who served in Experiment 1 was tested again using all four pairs of templates. Moreover, we used a slightly different range of images ([24–40] instead of [20–36]). This permitted us to estimate, for the same participants, how sensitive the psychometric function is to the effect of the range. In addition, by compounding the results of Experiments 1, 3, and 4, the original morphing

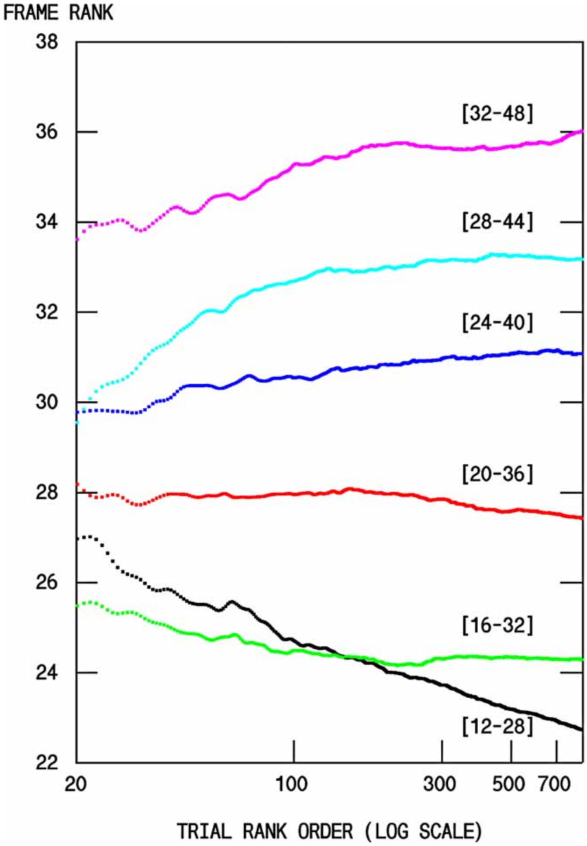

**Figure 8.** Evolution of the median of the psychometric function in the course of a trial. The figure includes data from Experiment 1 (range [20–36]), Experiment 3 (ranges [12–28], [16–32], [28–44], [32–48]), and Experiment 4 (range [24–40]). In all cases there is a clear drift of the median towards the final observed values.

sequence for one pair of templates was explored uniformly by six nonoverlapping ranges of images.

## Methods

*Participants.* Ten individuals chosen randomly among those who had already participated in Experiments 1 and 2.

*Stimuli, task, and procedure.* The stimuli, the experimental procedure, and the task were as in Experiment 1. The only difference was the range of

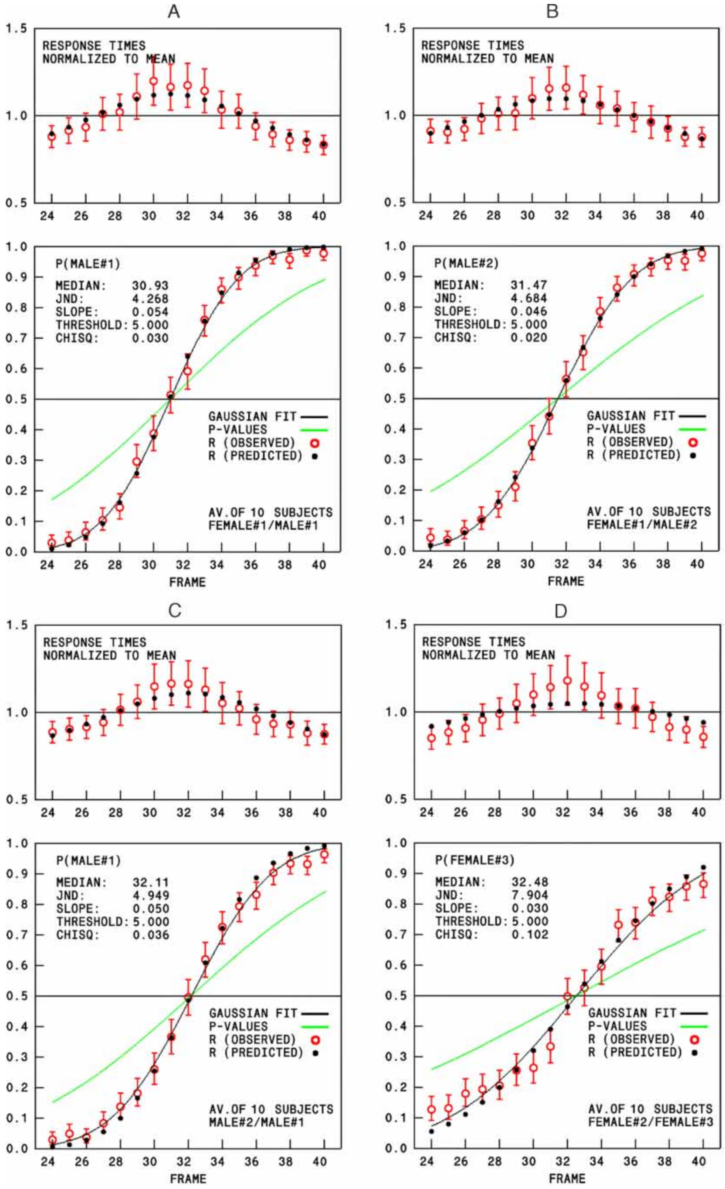

**Figure 9.** Experiment 4: Control. Psychometric functions and response times for all template pairs (same format as in Figure 3). The same sample of individuals who participated in Experiment 1 was tested again with a slightly different range of frames within the morphing sequence ([24–40]). In all cases, even this small difference resulted in a corresponding shift of the median.

variation of the stimuli which was set to [24–40]. The experiment took place after completing Experiment 2.

## Results

The results are reported in Figure 9 with the same format as Figure 1. They show that a shift of just four frames along the morphing sequence was reflected faithfully on the entire psychometric function. Specifically, the shift between medians in Experiments 1 and 4 was roughly the same for the four pairs of faces ([$F_1$, $M_1$]: 3.65; [$F_1$, $M_2$]: 3.37; [$M_1$, $M_2$]: 3.78; [$F_2$, $F_3$]: 4.64). The peak relative values of the RTs also moved across conditions mirroring the shift of the medians. The mean RTs (over ranks) were: [$F_1$, $M_1$]: 916($\pm$ 55) ms; [$F_1$, $M_2$]: 867($\pm$67) ms; [$M_1$, $M_2$]: 954($\pm$75) ms; [$F_2$, $F_3$]: 992($\pm$80) ms. Statistical analysis (ANOVA, 4[face pair] × 17[rank], repeated measure; Greenhouse-Geisser correction) detected no significant difference among mean RTs, $F(3, 27) = 1.467$, $p = .246$. There was a clear effect of the stimulus rank order, $F(16, 144) = 13.310$, $p = .001$, but no significant interaction, $F(48, 432) = 1.277$, $p = .301$. As in Experiment 1, for all pairs of templates RTs were symmetrically bell-shaped with a maximum close to the point of indifference: Quadratic term, $F(1, 9) = 18.846$, $p = .002$; order 4, $F(1, 9) = 13.497$, $p = .005$; order 6, $F(1, 9) = 10.752$, $p = .010$. Unlike the median of the distribution, the JND was almost independent of the range of variation of the stimuli (average across pair of faces: 5.06 for Experiment 1, 4.86 for Experiment 4). On the one hand, the experiment generalized to all pairs of templates the effect of the range demonstrated by Experiment 3. On the other hand, the results fit in nicely with the trend that emerged when Experiments 1, 3, and 4 are considered together (Table 1).

# EXPERIMENT 5

The analysis and the modelling of the first four experiments are based on two assumptions: (1) The objective figural information sampled from the stimuli is summarized by a prothetic discriminal variable, and (2) responses are related in a principled way to the distribution of this variable. In other words, we have assumed that, as the stimulus moves along the morphing sequence, the corresponding percept moves along a path in representational space connecting the two templates. As mentioned in the introduction, this path may differ considerably from the hyperline connecting the templates (Busey, 1998). The deviation may be due partly to the morphing algorithm, and partly to the very nature of the representational space. Whatever the cause, a strong deviation implies that, for the purpose of psychophysical analysis, perceptual space cannot be collapsed onto a single dimension. Therefore, analyses based this assumption may be severely biased. In fact, strong deviations have documented even with stimuli much simpler than faces, the best example being pitch perception. From the physical point of view, pitch is a simple one-dimensional variable. Yet, as Révész (1913) pointed out long ago, many people, particularly trained musicians, perceive pitches as *tones*, i.e., pitches within the context of the tonal scale. Unlike pitches, tones are two-dimensional objects defined jointly by the *height* of the sound, and its position within the octave (the *chroma*). Analysis (cf. Deutsch, 1982) has shown the tones are perceived in a cyclic manner, in which the cycle repeats at the octave. In other words, a sequence of pitches is perceived as moving in perceptual space along a helix. Thus, a note on the piano is perceived closer to its homologue in the upper octave than a note within its own octave. Experiment 5 was designed to verify whether the proximity structure of face space is distorted in a similar way. We reasoned that a confidence rating paradigm would expose the distortion, because judgements take into account the entire range of represented stimuli. Indeed, if a single stochastic variable underlies the responses, reliability estimates should be congruent with the categorical judgements. By contrast, the distortion would have eluded the simple forced-choice paradigm adopted in Experiment 1.

## Methods

*Participants.* The 14 individuals who participated in Experiment 3 were tested also in this experiment.

*Stimuli.* The stimuli were 12 pictures (rank order: [24–35]) from the four morphing sequences ([$F_1$, $M_1$]; [$F_1$, $M_2$]; [$M_1$, $M_2$]; [$F_2$, $F_3$]) used in the previous experiments. Stimuli were presented for a maximum of 4000 ms. The successive stimulus was presented 500 ms after entering a response. Each stimulus was presented 50 times in a pseudorandom order with the constraint that successive stimuli had to be at least three positions apart within the morphing sequence. In each experimental session $12 \times 50 = 600$ stimuli were presented.

*Task and procedure.* The task was to indicate which template the stimulus was more similar to (forced choice). However, unlike Experiment 1, participants complemented the indication of the selected template with a subjective estimate of the likelihood that the response was correct (confidence rating). By adopting a 6-point (ordinal) scale, there were 12 possible responses from A6 ("Sure A") to B6 ("Sure B"), each corresponding to one of the keys F1–F12 on the keyboard. Responses had to be entered as soon as possible. In all but a few exceptional cases responses were given before the end of the presentation of the stimulus, causing the disappearance of the stimulus. The familiarization phase was much more extensive than that in the other experiments, and consisted of 200 trials. The experiment involved four sessions (one for each pair of templates), which were administered in counterbalanced order, at least one week apart. The experiment took place after Experiment 3.

## Results

The results are reported in Figure 10. The left plots in each panel report the frequency distributions of the confidence ratings for all indicated stimulus ranks (see figure caption). It was assumed that to each stimulus is associated a Gaussian discriminal variable with a constant variance and an average that depends on the stimulus. The frequency of each possible response is then predicted by the probability with which the discriminal variable falls within the response categories marked on the common axis (abscissa) by equispaced boundaries. The empirical distributions were fitted simultaneously by assuming a principled relation (actually, a psychophysical function) between the stimulus rank order $k$ and the average of the distribution:

$$\mu(k) = \frac{A}{2} - A\left(1 - \frac{1}{1 + \exp(-\alpha(k - k_0))}\right)$$

where A is range of variation of the dependent variable, $k_0$ is the value for which the function is zero, and $\alpha$ sets the slope of the function at $k_0$. The category scale was fixed by setting to 1 the (constant) width of the response categories.

The upper right plots show the variation of the distribution average yielding the best approximation to the empirical distributions. The best-fitting values of the parameters ($A$: Amplitude; $k_0$: zero; $\alpha$: $4 \times \text{slope}/A$), estimated with a standard Simplex algorithm, are reported inset. The lower right plots compare the empirical psychometric function (binary responses) derived for the confidence responses shown in the left panels (circles), with the theoretical curve corresponding to best-fitting $\mu(k)$ function (black line).

For all pairs of templates participants were able to provide reliable confidence ratings, the frequency distribution of the responses being a well-behaved function of the stimulus rank order within the morphing sequence. The distributions were predicted accurately by postulating a prothetic discriminal variable with a Gaussian density function. The average $\mu(k)$ of the discriminal variable was a markedly nonlinear function of the stimulus

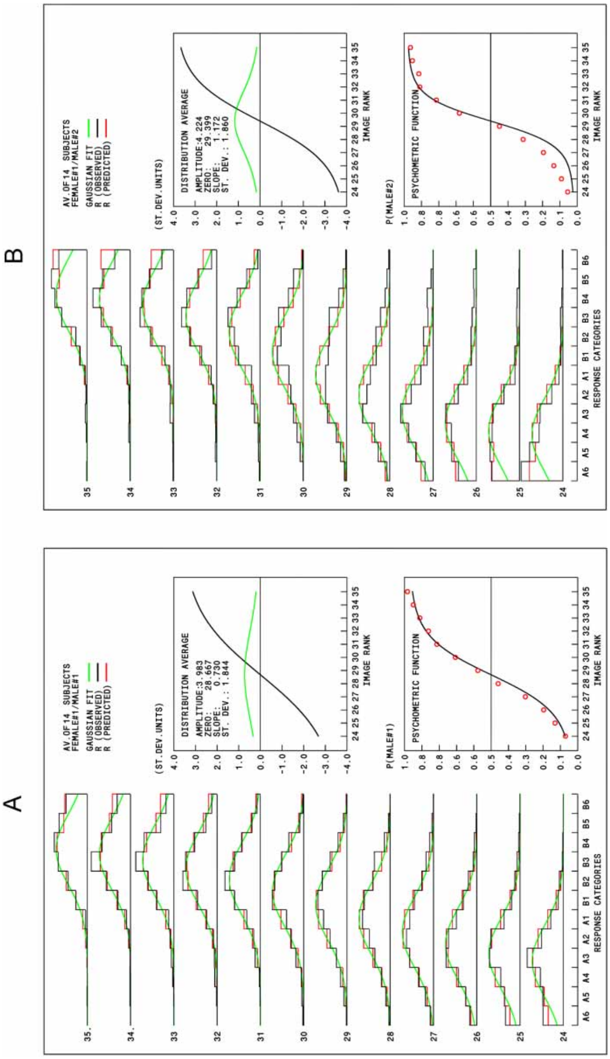
A
AV.OF14 SUBJECTS
FEMALE#1/MALE#1
GAUSSIAN FIT
R (OBSERVED)
R (PREDICTED)
(ST.DEV.UNITS)
DISTRIBUTION AVERAGE
AMPLITUDE:3.983
ZERO: 28.667
SLOPE: 0.730
ST. DEV.: 1.844
P(MALE#1)
PSYCHOMETRIC FUNCTION
IMAGE RANK
RESPONSE CATEGORIES
A6 A5 A4 A3 A2 A1 B1 B2 B3 B4 B5 B6
B
AV.OF14 SUBJECTS
FEMALE#1/MALE#2
GAUSSIAN FIT
R (OBSERVED)
R (PREDICTED)
(ST.DEV.UNITS)
DISTRIBUTION AVERAGE
AMPLITUDE:4.224
ZERO: 29.399
SLOPE: 1.172
ST. DEV.: 1.860
P(MALE#2)
PSYCHOMETRIC FUNCTION
IMAGE RANK
RESPONSE CATEGORIES
A6 A5 A4 A3 A2 A1 B1 B2 B3 B4 B5 B6

Figure 10

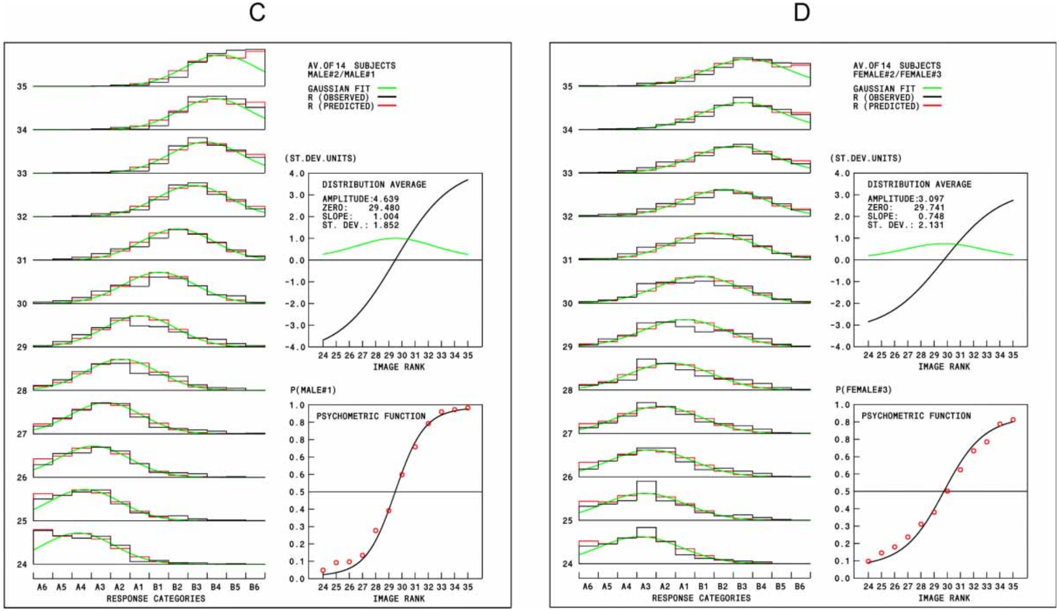

**Figure 10.** Experiment 5: Confidence ratings. A: [F$_1$, M$_1$]; B: [F$_1$, M$_2$]; C: [M$_1$, M$_2$]; D: [F$_2$, F$_3$]. Left panels: Frequency distributions of confidence rating responses for the indicated stimulus rank order (black histograms), Gaussian fit to the distributions (continuous lines), and predicted distributions (light grey histograms). Upper right panels: Theoretical evolution of the average of the Gaussian distribution as a function of the stimulus rank order (black line), and its derivative (grey line). This function was used to interpolate the frequency distributions. Lower right panels: Psychometric function derived from the confidence ratings (circles) and associated Gaussian fit (continuous line).

rank order. Finally, the psychometric functions derived from the confidence rating were quite similar to those estimated in Experiment 4 with an almost identical range of variation of the stimuli. In particular, in both experiments the identification of the pair [$F_2$, $F_3$] turned out to be more difficult than that of all other pairs. In conclusion, the experiment confirmed that the identification of ambiguous faces is based on a single discriminal variable, which summarizes all the figural cues sampled from the stimuli. The nonlinearity of the psychophysical function $\mu(k)$ also confirmed the suggestion (Experiments 1 and 2) of a perceptual anisotropy in face identification.

# GENERAL DISCUSSION

## Overview

The study addressed a number of issues concerning the perception of face identity that are not fully adjudged yet. The identification experiment (Experiment 1) provided reliable estimates of the median (point of indifference) and of JND of the psychometric functions describing the choice behaviour between templates (Figure 4). In a noncommittal interpretation of the data, medians delimit the basins of attractions of the two templates, and indicate their relative strength. However, if indeed individual faces are perceived categorically, the medians also identify the point along the morphing continuum where lays the boundary between perceptual categories. If so, they provide the background for the analysis of the discrimination data. Quite independently of the validity of the CP hypothesis, the JND estimate the difficulty of the identification. Thus, so to speak, they suggest how distant the two templates are in representational space. The discrimination experiment (Experiment 2) addressed directly the CP hypothesis. All four comparisons (Figure 5) demonstrated unambiguously that pairs of stimuli that straddled the median (as measured by Experiment 1) were discriminated more easily than pairs that were either to the left or to the right of the median. By general consensus, this is considered the hallmark of CP. The advantage of straddling over nonstraddling pairs built up progressively in the course of the experimental session (Figure 6). Experiments 3 and 4 demonstrated that the point of indifference in the identification task depends heavily on the position within the morphing sequence of the range of stimuli (Figures 7 and 9). Also in this case, the effect of the context increased as the session progressed (Figure 8), suggesting an important role of perceptual learning. Finally, the analysis of confidence ratings for identification (Experiment 5), demonstrated that, whatever the true dimensionality of the representational space for faces, the perceptual anisotropy responsible for the discrimination results may be represented by a one-dimensional metrics. In the following, we consider first some methodological issues and the motivation for modelling the results. Then, we discuss the import of the results in the context of current views on category learning.

## Comparing identification and discrimination performance

Ever since the seminal paper by Liberman et al. (1957), it has been claimed that a cross-validation of the results of identification and discrimination tasks is crucially important for supporting the CP hypothesis. Indeed, the claim has been voiced even by those who believe that CP is a myth: "[I]n standard CP research, it is *necessary* to show how discrimination is *directly* predicted by identification performance" (Massaro, 1998, p. 2277, our emphasis). The issue is how to produce such a direct demonstration.

For the purpose of this discussion, we qualify the question by asking how a psychometric function, which summarizes the identification performance, can be processed to predict response accuracy in the ABX task. The solution adopted by several authors (e.g., Calder et al., 1996; Levin & Angelone, 2002; Liberman et al., 1957; Newell & Bülthoff, 2002; Young et al., 1997) consists of adding the scaled difference in identification rate for the two stimuli presented in ABX task to the mean discriminability for the pairs at the end of the continuum. Aside from the fact, already pointed out by McKone et al. (2001), that the scaling factor is chosen ad hoc, it is doubtful that any manipulation whatsoever of the identification data is able to provide a sensible prediction of the discrimination data.

The reason for scepticism is best appreciated within the classic Thurstonian framework for interpreting psychometric functions (Guilford, 1954, pp. 154–156). However, the argument remains valid for any other sensible framework. In the simplest version of the classical approach, perceptual processes transform the physical stimulus into a discriminal stochastic variable $Z$ with a unimodal probability density function (pdf) $f_Z(\mu_Z, \sigma_Z^2, z)$. The distribution mean $\mu_Z$ depends on the intensity $x$ of the physical stimulus through a psychophysical function $\mu_Z = \mu_Z(x)$, whereas the variance $\sigma_Z^2$ is constant (Thurstone Type-V case). In the absence of biases, identification responses are dictated by a deterministic rule involving a threshold $T$: If $z < T$, answer A; if $z > T$, answer B. Then, the probability of observing an answer B (psychometric function) is:

$$P_B(x) = \int_T^{\infty} f_Z(\mu_Z(x), \sigma_Z^2, z)dz$$

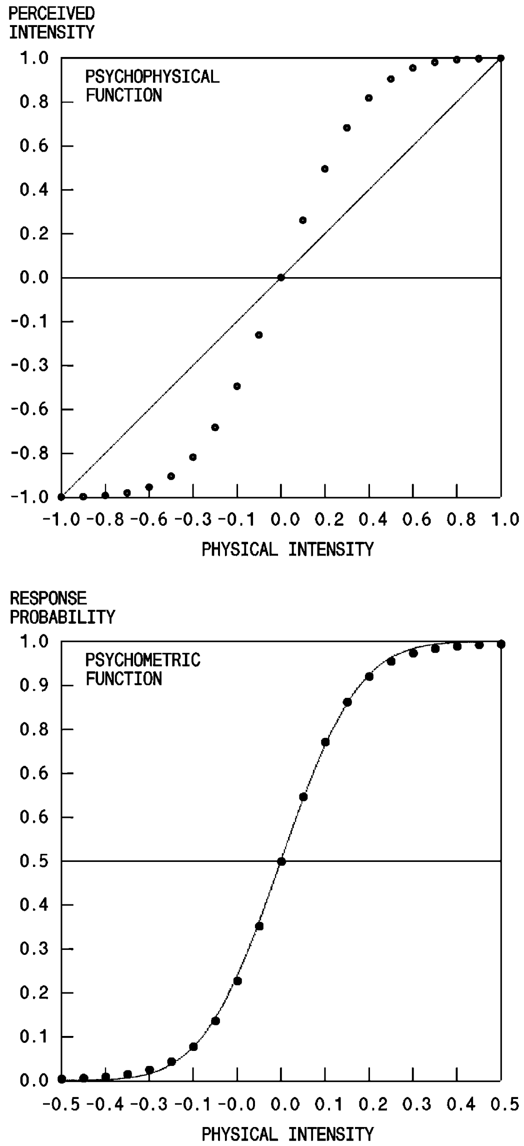

**Figure 11.** Simulated data (Thurston model). Upper plot: Two hypothetical psychophysical functions. One (filled dots) is strongly nonlinear, and suggests a categorical boundary in the middle of the range of values of the physical stimulus. The other one (continuous line) is linear. Lower plot: By choosing appropriately the variance of the discriminal variable that translates perceived intensity into response probabilities, the two psychophysical functions in the upper plot predict the same psychometric function.

The shape of the psychometric function depends jointly on the variance of the pdf $f_Z(\mu_Z,\sigma_Z^2, z)$, which models the variability intrinsic to the perceptual and decision processes, and on the psychophysical function $\mu_Z = \mu_Z(x)$, which models whatever systematic distortion is associated to the transformation between the physical stimulus and the discriminal variable. Specifically, in the presence of CP, one would expect $\mu_Z(x)$ to rise more steeply around the boundary between categories than both before and after the boundary. If so, in order to secure evidence in favour of CP, one should manipulate the psychometric function to demonstrate that $\mu_Z(x)$ is significantly nonlinear.

The following simulation shows that this cannot be done reliably. For the sake of simplicity, we assume that the physical intensity of the stimulus ranges between $-1$ for Template A to $+1$ for Template B, that the discriminal variable has a Gaussian distribution with a constant variance $\sigma_Z^2$, and that the threshold $T=0$. The upper plot in Figure 11 contrasts two psychophysical functions. The first one (filled dots) is assumed to be the integral Gaussian: $\mu_Z(x) = \mathrm{erf}(\sqrt{2}\ x/0.6)$, which is strongly nonlinear, and would certainly constitute clear evidence of CP. The second function (continuous line) is instead strictly linear: $\mu_Z(x) = x$ (no CP). The lower plot in Figure 11 compares the psychometric function predicted by the nonlinear psychophysical function when one assumes $\sigma_Z^2 = 0.35$ (filled dots) with the psychometric function predicted by the linear psychophysical function when one assumes instead $\sigma_Z^2 = 0.14$ (continuous line). The two predictions are virtually indistinguishable. Clearly, unless we had an independent estimate of $\sigma_Z^2$, no experimental data would be able to discriminate between the consequences of these two very different assumptions concerning the psychophysical function. By contrast, the results of an ABX task cannot be ambiguous: Only a strongly nonlinear psychophysical function can yield a significant difference between the identification rate for pairs of stimuli that straddle the category boundary and pairs of stimuli that do not. For this reason no empirical processing of the psychometric function can produce a direct prediction of the results of the ABX task. For the same reason, the psychometric function per se provides no direct evidence for or against CP. Therefore claims such as "Steep identification slopes provide evidence for categorical perception" (Pollak & Kistler, 2002, p. 9074), and similar statements by several other authors (Campanella et al., 2001, p. 247; de Gelder et al., 1997, p. 2; Etcoff & Magee, 1992, p. 233; Newell & Bülthoff, 2002, p. 122; Rossion et al., 2001, p. 1021; Rotshtein et al., 2005, p. 108; Suzuki et al., 2004, p. 1304; Young et al., 1997, p. 287), are not adequately substantiated.

## Why modelling the results?

One reason for developing quantal models for the identification and discrimination tasks was to circumvent the difficulty described in the previous section. Although different in several respects, the two models share the common intuition that responses are based on a comparison between the numbers of quanta of discriminal information sampled while the stimuli are displayed. Moreover, both models assume that the relative strength of the two templates within each morph maps—through the same $p$-functions—into the probability of sampling a quantum of either type. In essence, through the sampling mechanism and the associated $p$-functions, we formulated a principled hypothesis about the shape of the psychometric function. The excellent fit to the experimental data afforded reliable best-fit estimates of the parameters of the $p$-functions. The fact that the $p$-functions for the two tasks varied *pari passu* across pairs of templates proves at least the mutual consistency of the performances. Thus, although this cannot be construed as direct evidence, the fact that the $p$-functions were significantly nonlinear for three of the four pairs tested in the identification experiment does in fact corroborate the more conclusive demonstration of CP from the discrimination experiment.

The second reason for developing a quantal model was to provide a unified account of both response frequencies and response latencies (RT). It is natural to expect that, as the difficulty of a perceptual task increases, the deterioration of the performance is accompanied by an increase of the time to complete the task. This general statement has to be qualified depending on the nature of the task. In all identification tasks (Experiments 1, 3, and 4) RTs were indeed longest for the most ambiguous stimuli, near the point where the psychometric function crosses the value 0.5. This, however, simply suggests a self-terminating process in which answers are given as soon as the accumulated discriminal evidence reaches a certain decision threshold (Laming, 1968; Luce, 1986). Although the presence of such an increase in RT dictated the formulation the model, RT data by themselves speak neither in favour nor against CP. In fact, a linear $p$-function (i.e., no CP) would still predict an increase of RT near the point of indifference. Relatively few studies of CP for faces have measured RTs in identification tasks (Campanella et al., 2001; de Gelder et al., 1997; Levin, 1996; Young et al., 1997). All reported a significant scalloping of the RT curves, but none of them provided a unified framework for relating latencies data with identification rate. Instead, the fact that one simple hypothesis on the accumulation of discriminal evidence predicted accurately both response frequencies and response latencies justifies the claim that identifying a morph to one template becomes increasingly difficult as the morph approaches the point of indifference.

Two discrimination experiments (Campanella et al., 2003; Newell & Bülthoff, 2002) reported that RTs for face pairs straddling the category boundary were faster than RTs for within-category pairs. Instead, response latencies in the ABX task (Experiment 2) were not significantly affected by the position of the $S_A$–$S_B$ pair along the morphing sequence (Figure 5). Because discriminating between-category pairs was easier than discriminating within-category pairs, this seems to be inconsistent with the intuition that difficult perceptual tasks ought to require more time than easy tasks. Yet, the result is in fact coherent with the general logic of the model. In the version of the model adopted for the identification task, answers are dictated by the *absolute* number of A- and B-type quanta sampled from the stimuli. Therefore, for any intermediate morph, accuracy increases with viewing time. Conversely, the time to reach a fixed decision threshold increases when the sampling probabilities for A- and B-type quanta become equal. By contrast, the version of the model adopted for the discrimination task assumes that answers are dictated by the *relative* size of the differences $N_X$–$N_A$ and $N_X$–$N_B$ between the number of quanta sampled from $S_X$, $S_A$, and $S_B$. Thus, because the sampling rate was assumed to be constant, and because both $S_A$ and $S_B$ remained on for 1 s, increasing the viewing time of $S_X$ beyond 1 s would not have improved accuracy. In fact, although the third ($S_X$) stimulus remained on for a maximum of 3 s, the average RT across face pairs was also close to 1 s. The point to be stressed is that, by assuming also for $S_X$ a sampling period equal to the RT, the model did predict very accurately the variations of the error rate.

## Gender and identity

It is possible that CP effects for faces are enhanced whenever the endpoint stimuli belong to different gender. The fact that same identity implies same sex but not vice versa introduces an inevitable confound in any attempt to disentangle the affects of gender and identity in face perception. Yet, these two attributes of a face differ in several ways: (1) Gender is a binary attribute; (2) identifying the gender of a person is a vital skill equally developed across individuals; and (3) the hierarchy of feature saliency for gender identification is rather stable across identities (Brown & Perret, 1993; Roberts & Bruce, 1988). By contrast, (1) there are millions of clearly different faces of either gender; (2) identifying a face requires the retrieval of a memory trace, which some people do more efficiently than others; and (3) there is no obvious hierarchy of discriminal features for face identification.

Using unfamiliar faces, Campanella et al. (2001) reported that discrimination performance with pairs of stimuli straddling the category boundary was well above chance when there was a gender difference, but essentially random otherwise. Our results did not confirm this effect of a superordinate category. Experiments 1, 2, 4, and 5 tested both pairs of templates that crossed gender ([$F_1$–$M_1$], [$F_1$–$M_2$]) and pairs that did not ([$M_1$–$M_2$], [$F_2$–$F_3$]). One might reason that if gender per se is perceived categorically, then the effect of gender-switching should add to whatever effect is produced by identity. Thus, a sharper category boundary should exist for pairs [$F_1$, $M_1$] and [$F_1$, $M_2$] than for pairs [$M_1$, $M_2$] and [$F_2$, $F_3$]. The JND associated to the psychometric functions from Experiments 1 (Figure 4) and 4 (Figure 9) did not support this line of reasoning. Although the two female faces were harder to identify ($F_2$, $F_3$: JND = 7.05) than either pair with different gender ($F_1$, $M_1$: JND = 4.32; $F_1$, $M_2$: JND = 4.46), the two male faces were not ($M_1$, $M_2$: JND = 4.43). Moreover, if a gender additive effect existed, in the discrimination task (Experiment 2), the difference between the error rates for pairs of stimuli that did and did not straddle the category boundary should be higher for different-gender than for same-gender faces. The results (Figure 6) also failed to confirm this prediction. Although the lowest difference between error rates was indeed observed for [$F_2$, $F_3$], the highest difference was found for the other same-gender pair [$M_1$, $M_2$], with different-gender pairs yielding intermediate values. Finally, the relation between the rank order position of the stimuli within the morphing sequence and the distribution of the confidence ratings in Experiment 5 was again shallower for [$F_2$, $F_3$] (Figure 10D) than for [$F_1$, $M_2$] (Figure 10B), but quite comparable to that for [$F_1$, $M_1$]. In conclusion, at least for the face pairs tested in our experiments, the gender category did not seem to enhance significantly the CP effect. To detect a more subtle interaction between identity and gender, it may be necessary to test a larger sample of between- and within-gender pairs, by controlling their similarity.

As for identity, the results across experimental conditions suggest that idiosyncratic differences among face pairs determined a stable pattern of distances among faces in the multidimensional space were they are represented. More importantly, these differences, as they were noticed and learned in the course of the experiment from single views of unfamiliar individuals, were sufficiently marked to warp the representational space in a way that is compatible with the CP hypothesis. Because the notion of familiarity is not clear-cut, our evidence does not necessarily conflict with the logical argument that memory traces are necessary for perceptual categories to emerge. By expanding this point, the next section will also address the question of why neither Beale and Keil (1995) nor Campanella et al. (2001) had found evidence of CP with unfamiliar faces.

## Learning perceptual categories

Informally, the familiarity of a face looks like a simple attribute ranging from that of our closest relatives, to the vague feeling that we may express as "Didn't I meet that guy before?" However, the term "familiarity" actually conflates diverse dimensions that must be distinguished when discussing the issue of categorical perception. Along one dimension, familiarity varies as a function of the number of times we have seen the *image* of a face. Thus, if dollar bills were the only source of information, the face of Alexander Hamilton would be more familiar than that of Ulysses Grant just because there are more $10 bills around that $50 dollar bills. However, George Washington is in a different league, not just because there are so many $1 bills, but also because we have all seen many *diverse views* of that face, whereas relatively few people have seen pictures of either Hamilton or Grant other than those on the bills. If we were to meet them, we would recognize Washington more easily than Hamilton or Grant: because it is defined by more templates, the identity category "Washington" is far richer than the other two. Yet another dimension is movement. Faces that we have seen moving, either because we know them personally, or because we have seen on TV, have a different kind of familiarity from those we have only seen pictures of. This is particularly relevant for the issue at hand because, as pointed out by Beale and Keil (1995), only through movement can one assess the range of deformation of a face that is compatible with the invariant bone structure of the head. Learning about this range may have an impact on the way we adjudge the identity of a morphed image. If so, the finding that categorical perception occurs with famous faces, but not with unfamiliar ones (Beale & Keil, 1995) must be qualified insofar as the familiarity of faces such as President Kennedy's cannot be gauged on the same scale as a face of someone the observers have never seen before the experiment, and was presented in just one fixed view.

Because identification and discrimination studies usually involve hundreds of presentations, it should be clear that the term "unfamiliar" refers only to the status of the stimuli before the experiments. Whether or not the original (unblended) faces are included in the stimulus, the sheer number of repetition of identical or similar face images is likely to generate, through perceptual learning, a certain degree of familiarity.

Although, our experiments deal only with the kind of token-specific familiarity exemplified by the portrait of President Grant on the $50 bill, Levin and Beale (2000) seem to be right in claiming that the on-line learning process that goes on during the experiment is both necessary and sufficient to explain the progressive emergence of categorical behaviour. If so, three questions arise. The first question is why learning does not happen systematically. Again, we agree with suggestion by Levin and Beale that mixing discrimination trials from different continua, as both Beale and Keil (1995) and Campanella et al. (2001) did, might have been detrimental for establishing a robust memory trace. The second question concerns the time course of the learning process. Levin and Beale suggest that new perceptual categories emerge quickly, within the first half of the session. Our results from Experiment 2 (Figure 6) suggest instead a slower learning process that levels off much later in the session. Possibly, the single most important reason for this discrepancy is the fact that, after the familiarization phase, we never showed the unblended faces again. In fact, the endpoints of the tested interval of stimuli (rank order 18 and 38) were rather far away from the templates.

The third and most important question is what is actually learned (Goldstone, 1998). Before each session, participants familiarized thoroughly with the faces (templates) from which we derived the stimuli. Even though only one view was shown of each person, his/her identity was well established. Thus, although they were never shown again during the testing phase, one might assume that the pair of templates used in a given session was assumed to be representative tokens of a corresponding pair of perceptual categories. This, in turn, leads to a specific hypothesis on what is modified in the course of the experiment. Goldstone (1994) showed that training participants on simple categorization rules modified their perceptual discrimination ability. Recently, Notman, Sowden, and Özgen (2005) have suggested that the modifications induced by category learning take place as early on along the visual pathway as the primary visual cortex. Along similar lines, Livingstone et al. (1998) argued that the relationship between perceptual categories and similarities in the representational space may not point in the most obvious direction, with items falling in one category *because* they are similar to each other, but actually in the reverse direction: items that we have learned to put in the same category become more similar to each other than to any other item in other categories. If stable perceptual categories had been established during the familiarization phase, as in the identical twin faces experiment by Stevenage (1998), the evolution of the error rate in Experiment 2, and of the medians in Experiments 3 and 4, would describe the progressive impact of the categories on the metrics of the representational space. This hypothesis, however, predicts a convergence towards a stable pattern of similarity among the blends of templates, quite irrespective of the range of blends covered in any one session. Instead, Experiment 3 showed (Figure 8) that this range affected profoundly the median of the psychometric functions (i.e., our best estimate of the category boundaries). Actually, the entire psychometric function and the associated RT depended on the range of frames shown in any one session (Figure 7), which implies a corresponding dramatic effect on the pattern of similarity among blends. For instance, when frame 28 was the rightmost item in the range (Figure 7A), it had a probability of about .9 of being identified with $M_1$. The probability dropped to about .1 when the same frame was the leftmost item (Figure 7C). In conclusion, it seems that, no matter how well the templates were memorized, they did not have a major role in the learning process.

As an alternative to notion that the templates drove the learning process, the results of Experiments 3 and 4 suggest instead that learning was driven by the specific distribution of facial features presented in each session. To elaborate this point, let us consider again the session in which the stimuli were selected in the range of ranks [12–28] of the morphing sequence. Objectively, all stimuli were closer to Template A than to Template B. Yet stimuli with rank greater than 23 were more likely to be identified with B than with A (Figure 7A). By extrapolating to faces the "perceptual magnet" idea developed for speech signals (Iverson & Kuhl, 1995), we suggest that, through repeated exposure to the distribution of morphs, the endpoints of the range (stimuli 12 and 28) progressively acquired the status of prototypes and attracted the neighbouring stimuli. Of course, the endpoints were not identified as such. Thus, their peculiar status emerged implicitly, through some kind of abstraction process involving the neighbouring stimuli, which were similar, but not identical to each other. In this view, which is in keeping with a recent study on the adaptation to facial categories (Webster, Kaping, Mizokami, & Duhamel, 2004), stimuli closer to either endpoint tended to become equivalent (compression), while those near the middle of the range acquired distinctiveness (expansion). Note, however, that the effect of the range endpoints was compounded with a (weaker) effect of the true templates, because the point of indifference was always shifted towards the midpoint of the complete morphing sequence.

There has been considerable debate on the relative weight of compression and expansion effects. It has often been suggested (Goldstone, 1994; Livingstone et al., 1998; Stevenage, 1998) that baselines established with appropriate control groups should help to decide whether the learned-induced warping of the similarity space is associated with compression, expansion, or both. In practice, this strategy has produced rather inconclusive results. Our data do not permit us to address this issue. We note, however, that if the distance between the centres of gravity of the categories is not affected by learning—which may be hard to find out—compression and expansion must necessarily balance out. Therefore, perhaps, the only sensible question might be whether compression is the primary effect, and expansion is the inevitable consequence, or the other way around. Clearly, the hypothesis that we have set out above favours the former possibility.

## The dimensionality of the face space

As argued by Valentine (1991), assuming that faces are encoded as points in multidimensional space provides a congenial framework for investigating face identification and discrimination. Of course, no one really knows the dimensionality of this space. Actually, the number of significant dimensions is likely to depend on the context in which faces are perceived. For example, one multidimensional scaling study suggests that six dimensions capture most of the salient differences of neutral faces (Busey, 1998), while a study of facial expressions reckons only three dimensions (Bimler & Kirkland, 2001). The only fairly obvious fact is that the psychological face space has far fewer dimensions than the pixels face space, and only a loose connection with it.

The question that we want to address here is the minimum number of dimensions that one has to reckon when studying identification and discrimination. The question has methodological relevance whenever morphing algorithms are used to generate the stimuli. As noted in the introduction, there is no way of knowing a priori how the path that the algorithm traces from one face to another in pixel space maps into a corresponding path in psychological face space. Yet, virtually all techniques for describing identification and discrimination performance postulate an internal one-dimensional discriminal variable that is a monotonic function of the physical parameter that ranks the morphing sequence. In other words, the (generally implicit) assumption is made that, for the purpose of measuring identification and discrimination accuracy, the minimum number of relevant dimensions is one. In our quantal model, this assumption was made explicit by chaining two processes. First, the relative strength of the templates in the morph is transformed into sampling probabilities ($p$-functions). Then, the discriminal variable is set up by Bernouillian process that counts the number of quanta sampled from the morph. However, quite independently from this specific model, we felt it was necessary to verify that indeed a one-dimensional discriminal variable captures all the perceptual evidence that decisions are based on. Experiment 5 provided the required validation. By assuming that observers could set up a stable set of confidence ratings, we were able to describe the discriminal variable as a one-dimensional stochastic variable (Figure 10). More importantly, the way in which the mean of the associated pdf varied along the morphing sequence turned out to be quite similar to the $p$-functions predicted by the quantal model. The fact that a purely descriptive analysis of Experiment 5 yielded results that are fully compatible with those of Experiments 1–4, suggests that: (1) The path connecting the templates generated by the morphing algorithm was well-behaved; (2) the position of the morphs along the path can be parameterized by a prothetic variable; and (3) the premises upon which the quantal model is based are not obviously false.

## Concluding remarks

The experiments confirmed the presence of the phenomenon that is generally considered to be diagnostic for CP, namely the increased discriminability of pairs of morphs that straddle the point of subjective indifference identified by the identification function. Thus, it seems safe to conclude that even faces that were not familiar before the experiments are ultimately perceived categorically. Livingstone et al. (1998) argue that the relative merit of the two views—*similarities generate categories* versus *categories generate similarities*—is an empirical issue that can be adjudged experimentally. This amounts to suggest that the warping of the perceptual space on the one side, and the concept of category on the other are logically distinct concepts. However, the fact that the categorical effect builds up progressively as a result of perceptual learning processes taking place during the experiment casts some doubts on the validity of such a distinction, at least in the case of novel faces. To the extent that perceptual categories are defined only on terms of their effects on the metrics of the psychological space, as measured though a choice behaviour, the two concepts may actually collapse into a single one. This should not be taken to imply that the term "perceptual category" is in all cases just short-hand for a form of perceptual learning. Perceptual categories were first introduced to describe the perception of natural kinds—such as phonemes and colours—that is stable and involves little or no learning. Only later, the demonstration that the same diagnostic features observed in those cases are detected also with attributes such as the identity of unknown persons, which need to be learned, motivated the introduction of the notion of "induced category". What we are suggesting here is that, perhaps, confusion would be avoided if the term "categorical perception" were reserved to those instances where there is evidence that the clustering of the percepts in psychological space originates from within the perceptual system. For example, this may be the case of the perception of emotional expressions, which are known to be universal and genetically determined. Along the same line of thinking, one may prefer the term *CP-effects* in all those instances where the warping of the psychological space is the result of extensive training.

# REFERENCES

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, *1*(1), 21–61.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expression following bilateral damage of the human amygdala. *Nature*, *372*, 669–672.

Angeli, A., Davidoff, J., & Valentine, T. (2001). Investigating the factors producing categorical perception of face identity. In M. Corley (Ed.), *Proceedings of the XIIth ESCOP and XVIIIth*

*BPS Cognitive Society conference* (p. 35). Edinburgh, UK: ESCOP/British Psychological Society: Academia Press.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*, 217–239.

Bimler, D., & Kirkland, J. (2001). Categorical perception of facial expressions of emotions: Evidence from multidimensional scaling. *Cognition and Emotion*, *15*(5), 633–658.

Bornstein, M. H. (1987). Perceptual categories in vision and audition. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 535–565). New York: Cambridge University Press.

Bornstein, M. H., & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological Research*, *46*, 207–222.

Brown, E., & Perret, D. I. (1993). What give a face its gender? *Perception*, *22*, 829–840.

Burns, E. M., & Ward, W. D. (1978). Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *Journal of the Acoustical Society of America*, *63*, 456–468.

Busey, T. A. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science*, *9*(6), 476–483.

Bülthoff, I., & Newell, F. N. (2000). There is no categorical effect for the discrimination of face gender using 3D-morphs of laser scans of heads. *Investigative Ophthalmology and Visual Science*, *41*(4), S225.

Calder, A. J., Young, A. W., Perret, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, *3*, 81–117.

Campanella, S., Chrysochoos, A., & Bruyer, R. (2001). Categorical perception of facial gender information: Behavioural evidence and the face–space metaphor. *Visual Cognition*, *8*(2), 237–262.

Campanella, S., Hanoteau, C., Seron, X., Joassin, F., & Bruyer, R. (2003). Categorical perception of unfamiliar facial identities, the face–space metaphor, and the morphing technique. *Visual Cognition*, *10*(2), 129–156.

Campbell, R., Woll, B., Benson, P. J., & Wallace, S. B. (1999). Categorical perception of face actions: Their role in sign language and in communicative facial displays. *Quarterly Journal of Experimental Psychology*, *52A*(1), 67–95.

De Gelder, B., Teunisse, J.-P., & Benson, P. J. (1997). Categorical perception of facial expressions: Categories and their internal structure. *Cognition and Emotion*, *11*, 1–23.

Deutsch, D. (1982). Structural representations of musical pitch. In D. Deutsch (Ed.), *The psychology of music* (pp. 343–390). New York: Academic Press.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*, 107–117.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*, 227–240.

Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience*, *6*, 428–432.

Gauthier, I., & Logothetis, N. K. (2000). Is face recognition not so unique after all? *Cognitive Neuropsychology*, *17*(1/2/3), 125–142.

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience*, *12*(3), 495–504.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*, 178–200.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Harnad, S. (1987). *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, *97*, 553–562.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, *3*(8), 759–763.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Kanwisher, N., & Moscovitch, M. (2000). The cognitive neuroscience of face processing: An introduction. *Cognitive Neuropsychology*, *17*(1/2/3), 1–11.

Kanwisher, N., Stanley, D., & Harris, A. (1999). The fusiform face area is selective for faces not animals. *NeuroReport*, *10*, 183–187.

Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *Journal of the Acoustical Society of America*, *70*, 340–349.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. London/New York: Academic Press.

Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1364–1382.

Levin, D. T., & Angelone, B. I. (2002). Categorical perception of race. *Perception*, *31*, 567–578.

Levin, D. T., & Beale, J. M. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception and Psychophysics*, *62*(2), 386–401.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Liberman, A. M., Harris, K. S., Kinney, J., & Lane, H. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368.

Liu, C. H., & Chaudhuri, A. (2003). What determines whether faces are special? *Visual Cognition*, *10*(4), 385–408.

Livingstone, K. R., Andrews, J. K., & Harnad, W. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 732–753.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford, UK/New York: Oxford University Press.

Massaro, D. W. (1998). Categorical perception: Important phenomenon or lasting myth? In R. H. Mannell & J. Robert-Ribes (Eds.), *Proceedings of the 5th International Congress of Spoken Language Processing* (pp. 2275–2279). Sydney, Australia: Australian Speech Science and Technology Association Inc.

McKone, E., Martini, P., & Nakayama, K. (2001). Categorical perception of face identity in noise isolates configural processing. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(3), 573–599.

Moscovitch, M., Berhmann, M., & Winocur, G. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9*, 555–604.

Newell, F. N., & Bülthoff, H. H. (2002). Categorical perception of familiar objects. *Cognition*, *85*, 113–143.

Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: A psychophysical approach. *Cognition*, *95*, B1–B14.

Okabe, A., Boots, B., Sugihara, K., & Chiu, S. N. (2000). *Spatial tassellation: Concepts and application of Voronoi diagrams*. Chichester, UK: John Wiley.

Pevtzow, R., & Harnad, S. (1997). Warping similarity space in category learning by human subjects: The role of task difficulty. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Paine (Eds.), *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization. Department of Artificial Intelligence, Edinburgh University* (pp. 189–195). Edinburgh, UK: Edinburgh University Press.

Pollak, S. D., & Kistler, D. J. (2002). Early experience is associated with the development of categorical representations for facial expressions of emotion. *Proceedings of the National Academy of Science*, *99*(13), 9072–9076.

Raskin, L. A., Maital, S., & Bornstein, M. H. (1983). Perceptual categorization of color: A life-span study. *Psychological Research*, *45*, 135–145.

Révész, G. (1913). *Zur Grundlegung der Tonpsychologie [On the foundations of the psychology of tones]*. Leipzig, Germany: Feit.

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

Rossion, B., Schiltz, C., Robaye, L., Pirenne, D., & Crommelinck, M. (2001). How does the brain discriminate familiar and unfamiliar faces? A PET study of face categorical perception. *Journal of Cognitive Neuroscience*, *13*(7), 1019–1034.

Rotshtein, P., Henson, R. N. A., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, *8*(1), 107–113.

Sachs, L. (1984). *Applied statistics*. New York/Berlin: Springer-Verlag.

Sawusch, J. R., & Gagnon, D. A. (1995). Auditory coding, cues, and coherence in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 635–652.

Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. *British Journal of Psychology*, *89*, 39–57.

Suzuki, A., Shibui, S., & Shigemasu, K. (2004). Temporal characteristics of categorical perception of emotional facial expressions. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-sixth annual conference of the Cognitive Science Society* (pp. 1303–1308).

Tanaka, J., Giles, M., Kremen, S., & Simon, V. (1998). Mapping attractor fields in face space: The atypicality bias in face recognition. *Cognition*, *68*, 199–220. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*, 557–561.

Winston, J. S., Henson, R. N. A., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, *92*, 1830–1839.

Yin, R. K. (1969). Looking at upside down faces. *Journal of Experimental Psychology*, *81*, 141–145.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perret, D. I. (1997). Facial expressions megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, *63*, 271–313.

*Manuscript received September 2005*
*Manuscript accepted March 2006*
*First published online June 2006*

# APPENDIX 1

We derive the formal predictions of the quantal model for the identification experiment. The following notation is adopted. A, B: Templates. $S_k$: Stimulus with rank order $k$ within the morphing sequence from A to B. $p = p(k)$: Probability of acquiring a B-type quantum from $S_k$. The sampling of the stimulus is a Bernoulli process, and the probability that $j$ samples include exactly $(j-T)$ A-type quanta and $T$ B-type quanta is:

$$P(T,j,p) = C_{jT} p^T (1-p)^{j-T}$$

By Bayes' rule, the probability that exactly $j$ samples must be acquired to obtain $T$ quanta of either type is:

$$P_A(j,p) = \frac{T}{j} C_{jT} p^{j-T} (1-p)^T \quad P_B(j,p) = \frac{T}{j} C_{jT} p^T (1-p)^{j-T}$$

The number of samples acquired before either the Type-A, or the Type-B count reaches threshold ranges between $T$ and $2T-1$. Thus, the probability of obtaining an identification response A or B is:

$$P_A(p) = 1 - P_B(p)$$

$$P_B(p) = \sum_{j=T}^{2T-1} P_B(j,p) = \sum_{j=T}^{2T-1} \frac{T}{j} C_{jT} p^T (1-p)^{j-T}$$

$$= 1 - \frac{1}{2} C_{2T,T} p^T (1-p)^T H([1,2T],[T+1],1-p)$$

where H is the generalized hypergeometric function.

The number of quanta sampled before reaching a decision is a random variable with a discrete distribution:

$$D_Q(j,p) = P_A(j,p) + P_B(j,p) =$$

$$\frac{T}{j} C_{jT} (p^T (1-p)^{j-T} + p^{j-T} (1-p)^T) \quad [T \le j \le T-1]$$

$$D_Q(j,p) = 0 \ [0 \le j < T] \quad [2T-1 < j]$$

Thus, the average number of quanta sampled from $S_k$ at response time is:

$$N_Q(p) = \sum_{j=T}^{2T-1} j D_Q(j,p) = \sum_{j=T}^{2T-1} T C_{jT} (p^T (1-p)^{j-T} + p^{j-T} (1-p)^T)$$

The general average over all $p$-values is:

$$\overline{N}_Q = \int_0^1 N_q(p)dp = 2T(\psi(2T-1) - \psi(T+1)$$

where $\psi$ is the Digamma function. The average response time to stimulus $S_k$ is RT($k$) = $N_Q(p) \times R_s$. RT($k$) is minimum (RT$_{\mathrm{min}}$ = $T$) for $p$ = 0 and $p$ = 1, and maximum for $p$ = .5 where it has the value:

$$RT_{\max} = 2T\left(\frac{\mathrm{C}_{\mathrm{2T,T}}}{2^{\mathrm{2T}}}\mathrm{H}([1,\mathrm{2T}+1],[\mathrm{T}+1],\frac{1}{2}) - 2\right)$$

The rate at which quanta are sampled can be estimated by dividing the mean response time RT$_{\mathrm{mean}}$ for all stimuli by$\overline{N}_Q$. $\overline{N}_Q$ is an increasing function of $T$ with an almost constant slope that converges rapidly to 2log(2) as $T$ increases. Thus, the mean response time is approximately 2log(2) × $R_s$.

The discrimination power can be estimated by the slope of the function $P = P_B(p)$ at $p$ = .5:

$$\left|\frac{\partial P_B(p)}{\partial p}\right|_{p=0.5} = \frac{TC_{2T,T}}{2^{2T}(T+1)}H\left([2,2T+1],[T+2],\frac{1}{2}\right)$$

which is inversely related to the JND. The slope is an increasing function of $T$. Thus, the model predicts that discrimination improves with the threshold.

## Fitting the model to the data

To fit the model to the data, we adopted the following strategy. Individual response times were normalized by subtracting their mean RTs and dividing by the population mean RT$_{\mathrm{pop}}$. Response frequencies were computed by pooling for each $k$-value the number of B responses of all participants. The parameter $\mu$ was estimated directly from the psychometric function, by interpolating the value of $k$ for which the response frequency is .5 (continuous black lines in Figure 3). Thus, the $p$-value function actually has only the free parameter $\sigma$, which was estimated along with the threshold $T$ by minimizing the quantity:

$$\chi^2 = \sum_{k=1}^{N}\left[(P_B(p(k),T) - F_B(k))^2 + \left(\frac{N_Q(p(k),T) - \overline{N}_Q}{\overline{N}_Q} - \frac{RT(k)-1}{M_{RT}}\right)^2\right]$$

with a constrained Simplex minimization routine ($T$ constrained to integer values).

# APPENDIX 2

We derive the formal predictions of the quantal model for the discrimination experiment. Let $p_A$ and $p_B$ be the probabilities of sampling a B-type quantum $Q_A$ from $S_A$ and $S_B$, respectively. Let $T_p$ be the common presentation interval of the three images, $R_s$ the sampling rate, and $N = R_s T_p$ the (constant) total number of quanta sampled within the interval $T_p$. The number of B-type quanta sampled within $T_p$ has the Bernoullian distributions $\mathrm{B}(N,i,p_A)$ for $S_A$, $\mathrm{B}(N,i,p_B)$ for $S_B$, and $\mathrm{B}(N,i,p_A)$ or $\mathrm{B}(N,i, p_B)$ for $S_X$. The distribution of the differences between counts is a convolution that depends on $S_X$:

$$\text{Distribution of } N_X - N_A\!: P_{AA}(j) = \mathrm{B}(N, -j, p_A)^{*}\mathrm{B}(N, i, p_A) \quad \text{if } S_X = S_A$$
$$\text{Distribution of } N_X - N_A\!: P_{AB}(j) = \mathrm{B}(N, -j, p_A)^{*}\mathrm{B}(N, i, p_B) \quad \text{if } S_X = S_B$$
$$\text{Distribution of } N_X - N_B\!: P_{BA}(j) = \mathrm{B}(N, -j, p_B)^{*}\mathrm{B}(N, i, p_A) \quad \text{if } S_X = S_A$$
$$\text{Distribution of } N_X - N_B\!: P_{BB}(j) = \mathrm{B}(N, -j, p_B)^{*}\mathrm{B}(N, i, p_B) \quad \text{if } S_X = S_B$$

The four distributions have the following expressions:

$$P_{AA}(j) = \sum_{i=j}^{N} C_{N,i-j} C_{N,i} p_A^{2i-j} (1 - p_A)^{2N-2i+j}$$

$$P_{AB}(j) = \sum_{i=j}^{N} C_{N,i-j} C_{N,i} p_A^{2i-j} (1 - p_A)^{2N-2i+j} p_B^{i} (1 - p_B)^{N-i}$$

$$P_{BA}(j) = \sum_{i=j}^{N} C_{N,i-j} C_{N,i} p_B^{2i-j} (1 - p_B)^{2N-2i+j} p_A^{i} (1 - p_A)^{N-i}$$

$$P_{BB}(j) = \sum_{i=j}^{N} C_{N,i-j} C_{N,i} p_B^{2i-j} (1 - p_B)^{2N-2i+j}$$

$$[-N \le j \le N]$$

Because images are sampled independently, the joint distribution of $N_X - N_A$ and $N_X - N_B$ is $P_A(j,k) = P_{AA}(j) \times P_{BA}(k)$ and $P_B(j,k) = P_{BB}(j) \times P_{AB}(k)$, $[-N \le j, k \le N]$ for $S_X = S_A$ and $S_X = S_B$, respectively. Thus, the conditional answer probabilities are:

$$P(''S_X = S_A''|S_X = S_A) = \sum_{j=-(k\text{-}1)}^{k-1} \sum_{k=1}^{N} P_A(j,k) +$$
$$\sum_{j=(k+1)}^{\text{-}(k+1)} \sum_{k=-N}^{-1} P_A(j,k) + \frac{1}{2}\left(\sum_{k=-N}^{N} P_A(k,k) + \sum_{k=-N}^{N} P_A(-k,k) - P_A(0,0)\right)$$

$$P(''S_X = S_A''|S_X = S_B) = \sum_{k=-(j\text{-}1)}^{j-1} \sum_{j=1}^{N} P_B(j,k) +$$

$$\sum_{k=(j+1)}^{\text{-}(j+1)} \sum_{j=-N}^{-1} P_B(j,k) + \frac{1}{2}\left(\sum_{k=-N}^{N} P_B(k,k) + \sum_{k=-N}^{N} P_B(-k,k) - P_B(0,0)\right)$$

$$P(''S_X = S_B''|S_X = S_A) = 1 - P(''S_X = S_A''|X = S_A)$$
$$P(''S_X = S_B''|S_X = S_B) = 1 - P(''S_X = S_A''|X = S_B)$$

Finally, the probability $P_E$ of a wrong answer is:

$$P_E = P(''S_X = S_B''|S_X = S_A) \times P(S_X = S_A)$$
$$+ P(''S_X = S_A''|S_X = S_B) \times P(S_X = S_B)$$

If $P(S_X = S_A) = P(S_X = S_B) = .5$, the model is doubly symmetric with respect to the sampling probabilities. There, $P_E(p_1 = \alpha,\ p_2 = \beta) = P_E(p_1 = \beta,\ p_2 = \alpha) = P_E(p_1 = 1-\beta,\ p_2 = 1-\alpha)$. All distributions and probabilities defined above may be expressed in terms of generalized Legendre P functions.

## Fitting the model to the data

As in the identification model, we assume that the $p$-value function between the position of the images along the morphing sequence and the sampling probabilities is described by the discrete approximation to a cumulative Gaussian function with the same parameters:

$$p_A = p_A(k) = \sum_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(i-\mu)^2}{2\sigma^2}} \qquad p_B = p_B(k) = \sum_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(i+\Delta-\mu)^2}{2\sigma^2}}$$

For a given distance $\Delta$ between $S_A$ and $S_B$, the model has three parameters, namely the total number $N$ of quanta sampled while the images are shown, and the quantities and s that characterize the $p$-value functions $p_A$ and $p_B$. The best-fitting values of the parameters were computed by minimizing the quantity:

$$\chi^2 = \sum_{k=1}^{N} (P_E(p_A(k), P_B(k), N) - F_E(k))^2$$

with a constrained Simplex minimization routine ($N$ constrained to integer values, $\Delta$ set to 8).