Cognition 121 (2011) 313–323

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Variability in photos of the same face

Rob Jenkins [a,*], David White [b], Xandra Van Montfort [a], A. Mike Burton [c]

[a] School of Psychology, University of Glasgow, 58 Hillhead Street, Glasgow G12 8QQ, United Kingdom
[b] Department of Psychology, University of New South Wales, Australia
[c] Department of Psychology, University of Aberdeen, United Kingdom

ARTICLE INFO

Article history:
Received 6 July 2010
Revised 27 July 2011
Accepted 2 August 2011
Available online 3 September 2011

Keywords:
Face perception
Identity
Photography
Face recognition
Attractiveness

ABSTRACT

Psychological studies of face recognition have typically ignored within-person variation in appearance, instead emphasising differences *between* individuals. Studies typically assume that a photograph adequately captures a person's appearance, and for that reason most studies use just one, or a small number of photos per person. Here we show that photographs are not consistent indicators of facial appearance because they are blind to within-person variability. Crucially, this within-person variability is often very large compared to the differences between people. To investigate variability in photos of the same face, we collected images from the internet to sample a realistic range for each individual. In Experiments 1 and 2, unfamiliar viewers perceived images of the same person as being different individuals, while familiar viewers perfectly identified the same photos. In Experiment 3, multiple photographs of any individual formed a continuum of good to bad likeness, which was highly sensitive to familiarity. Finally, in Experiment 4, we found that within-person variability exceeded between-person variability in attractiveness. These observations are critical to our understanding of face processing, because they suggest that a key component of face processing has been ignored. As well as its theoretical significance, this scale of variability has important practical implications. For example, our findings suggest that face photographs are unsuitable as proof of identity.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Theories of face recognition are based almost entirely on studies of photo recognition. In such studies, a person's face may be represented by a single photograph that is repeated throughout the experiment (e.g. Dyer, Neumeyer, & Chittka, 2006; Golarai et al., 2007; Gupta & Srinivasan, 2008; Mehl & Buchner, 2008; Tsukiura & Cabeza, 2011), or by a matched pair or set of photos that differ only in one respect, such as facial expression or viewpoint (e.g. D'Argembeau et al., 2003; Turati, Bulf, & Simion, 2008; Winston, Henson, Fine-Goulden, & Dolan, 2004). Here we argue that equating photographs with faces perpetuates a serious misconstrual of the face recognition problem, leading to spurious findings and theorising that misses the core issue. By the same token, recasting the problem illuminates a clear remedial path. In the discussion we outline a promising approach to this.

The problem of face recognition is often presented as a problem of telling people apart. Given that all human faces share the same basic template (two eyes above a nose above a mouth), how are we able to distinguish among many thousands of individuals? This question is often addressed in the context of within-category discrimination (e.g. Bukach, Gauthier, & Tarr, 2006; McKone, Kanwisher, & Duchaine, 2007). Since this perspective emphasises sensitivity to differences between individuals, it encourages the traditional focus on *between-person* variability. Experimentally, this often reduces to *between-photo* variability, where each person is represented by a single photo. This substitution of photos for faces implies that a photograph adequately captures a person's appearance, such that exposure to the snapshot is interchangeable with exposure to the face. The

* Corresponding author. Tel./fax: +44 141 330 4663.
E-mail address: rob.jenkins@glasgow.ac.uk (R. Jenkins).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.08.001

purpose of the present study is to challenge this idea. We show that a photograph is not a reliable indicator of facial appearance because it is blind to within-person variability. Crucially, this within-person variability is large compared with between-person variability. This is a transformative observation, not only for cognitive theories of face recognition, but also for face recognition in applied settings.

Face photographs sample three interacting layers of variation: The face itself undergoes non-rigid deformations – on the millisecond scale during muscular movement, and on the decade scale over ageing. Surface reflectance properties of the face are also affected by many factors, including cardiovascular activity in the short term, and general health in the longer term. Superimposed upon these face changes are lighting and other atmospheric changes, which vary with the ambient environment. Finally, image parameters such as resolution and depth of contrast depend on the characteristics of the camera. The interplay between these variables guarantees that no two photos of any face are the same. In practice, different photos of an individual vary greatly (see Fig. 1).

The photographs in Fig. 1 were not chosen to be especially variable. Indeed four of them are from current photo-identification documents. Notice that even this relatively modest range of variability is rarely admitted to the laboratory. The experimental convention is to minimise image variability, treating it as 'noise' that merely obscures the problem of interest. This creates a fundamental disjoint between the situation that we would like to understand and the situation that is studied in the lab. Within-person variability pervades face recognition in the real world, because no face casts the same image twice. The only exception to this is repetition of photographs, yet a great deal of experimental work is based solely on this artificial and anomalous case. Conversely, within-person variability has been almost entirely overlooked, and has never been examined in its own right.

Fig. 1. Current passport photos (Left), staff card photos (Middle), and personal photos (Right) for authors RJ (top) and AMB (bottom). Consider image similarity by rows and by columns.

It is worth considering some possible reasons why within-person variability has been so comprehensively ignored. Certainly, there is the pragmatic reason that it is much easier to present photographs in experiments than to present faces (and also somewhat easier to present one photograph of each face than to present more than one photograph of each face). However, previous face recognition research suggests a more psychologically interesting reason: perhaps within-person variability has never been directly addressed because we are simply unaware of its scale. Familiar face recognition is surprisingly robust, in the sense that we can recognise familiar faces over an enormously wide range of viewing conditions (e.g. Bruce, 1982; Burton, Wilson, Cowan, & Bruce, 1999). In cognitive terms, this corresponds to a many-to-one mapping of diverse input images onto a more abstractive representation of the individual's face (e.g. a Face Recognition Unit in Bruce & Young's 1986 framework). It is possible that this funnel-like connectivity attenuates sensitivity to variation in input, leading to underestimation of within-person variability in familiar faces. We return to this issue in the discussion.

In contrast to familiar face recognition, unfamiliar face recognition is surprisingly fragile. It can be disrupted by even superficial changes in the input image (Bruce, 1982; Burton et al., 1999; Megreya & Burton, 2006, 2008). Perhaps less intuitively, this too may lead within-person variability to be underestimated. Outside of the psychology experiments, we seldom receive feedback on recognition errors. So if we encounter an unfamiliar person on one day, and then fail to recognise the same person on a later day, we can simply assume that the second sighting was of a different person. This is a reasonable interpretation in the absence of feedback, but it is an error arising from a narrow view of within-person variability. The data presented below highlight the very large discrepancy between the expected range of this variability and the actual range.

Interestingly, a number of recent studies have begun to uncover large variability in the face recognition *ability* of observers. Duchaine and Nakayama (2006), and Russell, Duchaine, and Nakayama (2009) have described groups of individuals at opposite ends of this spectrum. 'Developmental prosopagnosics' (Duchaine & Nakayama, 2006) have profound difficulty with face recognition, despite having otherwise intact visual abilities and no history of brain damage. In contrast, 'Super-recognizers' perform exceptionally well on a range of face recognition tasks (Russell et al., 2009). Megreya and Burton (2006) have reported large and stable individual differences for a number of face processing tasks, and recently Burton, White, and McNeill (2010) developed the Glasgow Face Matching Test (GFMT) as an instrument for assessing subjects' ability to match unfamiliar faces. All of these studies point to substantial variability among *perceivers*. However, no theory yet addresses variability in the *person perceived*. We hope to persuade readers that within-person variability must be built into our theorising if the problem of face recognition is to be properly understood.

We begin in Studies 1 and 2 by using a photo sorting task to compare actual within-person variability with the expectations of naïve observers. In Study 3 we address the everyday notion of 'good likeness' and 'bad likeness'

photographs by examining the distribution of likeness ratings both within individuals and between individuals. Finally, in Study 4 we turn to within-person variability in facial attractiveness. The overall message from these studies is that photographs are not stable representations of facial appearance. This is true for forensically important judgements of identity. It is also true for socially important judgements of attractiveness.

2. Experiment 1

The purpose of this experiment was to examine face matching in the context of realistic within-person variability. Our main interest was observers' tolerance to this variability when matching photographs for identity. To investigate this, we developed a new sorting task using multiple photographs of different faces. In this task, participants are simply asked to group the photographs according to identity, so that different photos of the same person are gathered together. Participants are not told how many identities to expect, and are free to group the images however they wish. The crux of the study is the provenance of the images. A common approach to acquiring experimental face stimuli is to take new photographs that meet the particular requirements of the study (e.g., Megreya & Burton, 2006). Typically these are taken under controlled conditions, specifically to *minimise* image variability. Our intention here was the opposite: We sought to represent the full range of natural variability in images by using pre-existing photographs collected from the internet. We refer to such photos as *ambient images*, to emphasise that they are drawn from the surrounding environment rather than an experimental pool.

By allowing participants to cleave the photo set into as many or as few identities as they perceived, we hoped to reveal the range of variability that they would tolerate for a single identity. We predicted that participants would find it difficult to map diverse photos onto the same face, leading them to produce solutions that contained more identities than were actually presented.

2.1. Method

2.1.1. Stimuli

Twenty images of each of two Dutch celebrities (Chantel Janzen and Bridget Maasland) were downloaded from the internet (40 images in total). These individuals are well known in the Netherlands, and photographs of them are easy to find online. Importantly however, they were not known to our UK participants. The images were collected via Google Image, using the celebrities' names as search terms. We accepted the first 20 images of each face that (i) exceeded 150 pixels in height, (ii) showed the face in roughly frontal aspect, and (iii) were free from occlusions. All photos were converted to greyscale and printed onto laminated cards measuring 38 × 50 mm. Copyright restrictions prevent us from reproducing the images here. However, readers can easily replicate our search by using the celebrities' names as Google Image search terms. Fig. 2 shows a similar range of images for two other individuals.

2.2. Participants

Twenty UK undergraduates took part in the study in exchange for a small payment.

2.3. Procedure

Participants were given a shuffled deck of 40 face photos (20 photos per face), and were asked to sort them by identity, so that photos of the same face were grouped together. There was no time restriction on this task, and participants were free to create as many or as few groups as they wished.

2.4. Results and discussion

The median number of identities in participants' solutions was 7.5 (Mode 9; Range 3–16), reflecting the number of distinct identities perceived in the set. A one-sample *t*-test confirmed that this was significantly higher than the 2 identities that were actually presented [$t(19) = 7.82$, $p < 0.001$, $d = 1.8$]. In fact, none of our participants arrived at the correct solution. Photos of the same face were often deemed too dissimilar to go together, leading participants falsely to fractionate a single identity into several identities. By contrast, misidentification errors (i.e. sorting the two different people into the same pile) were infrequent, at less than 1 error per participant on average (Mode 0; Range 0–3). This pattern indicates that the problem is primarily one of integrating dissimilar images. It is difficult to find commonalities among photos of the same face that justify grouping them together. At the same time, it is easy to find differences that justify grouping them separately.

3. Experiment 2

In view of the very poor performance in Experiment 1, we next sought to rule out the possibility that the photo sets were inherently difficult for participants to process, perhaps due to poor image quality or biased sampling. To this end, we recruited 20 Dutch participants who were familiar with both of the faces shown in the task. If the images are identifiable in principle, then participants who are familiar with the faces should have no trouble sorting them correctly. On the other hand, if the images are somehow misrepresentative, even participants who know the faces should struggle with the task.

3.1. Method

The method was the same as for Study 1, except that the participants were now 20 Dutch volunteers who were familiar with the faces on the cards.

3.2. Results and discussion

Dutch participants straightforwardly sorted the photos into two groups, almost all of them performing perfectly (Median 2; Mode 2; Range 2–5). An independent samples *t*-test confirmed that the Dutch participants perceived

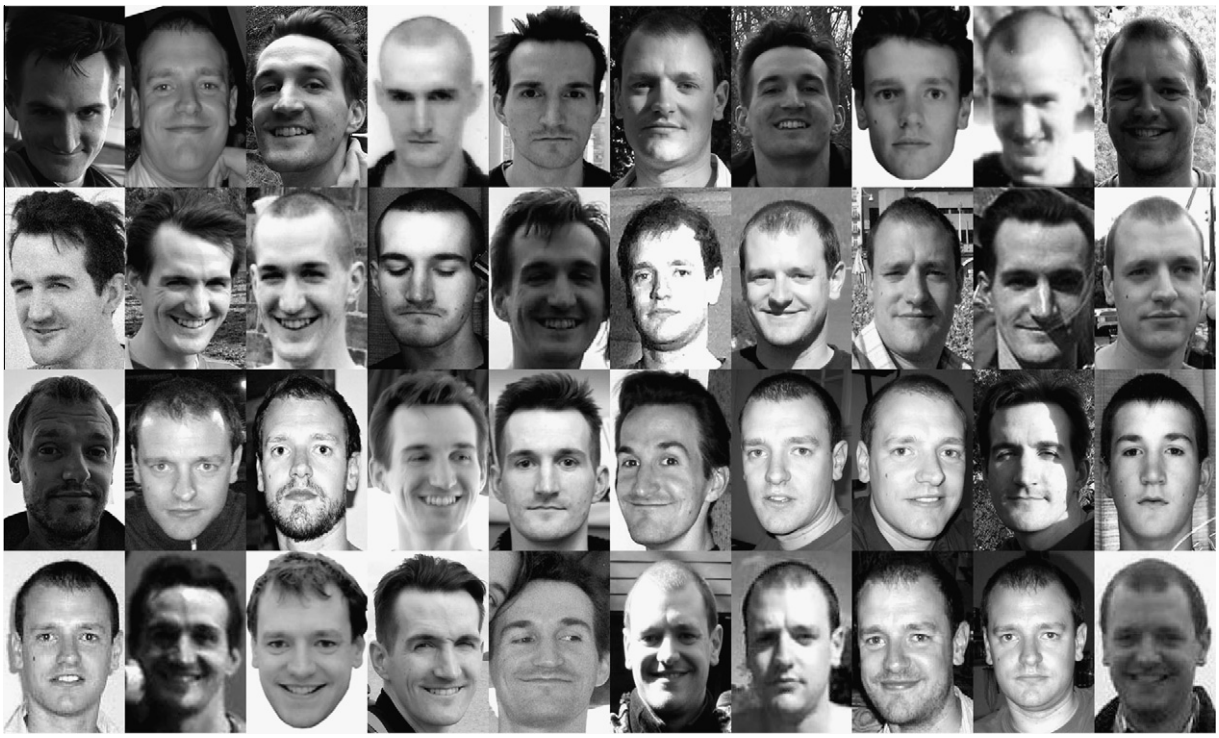

Fig. 2. Sorting face photos by identity is a difficult task, unless the faces are familiar. The solution for this set is given in Appendix I.

significantly fewer identities that the UK participants [$t(38) = 5.99$, $p < 0.001$, $d = 1.9$]. Misidentification errors were again low, at less than 1 error per participant on average (Mode 0; Range 0–3). These results confirm that the photographs in this task were all recognisable in principle. The problem for unfamiliar observers lies in separating image changes from face changes. Familiarity solves that problem.

4. Experiment 3

In the preceding experiments, ambient photos of an individual face were thought to depict different people, unless the observer was familiar with the face concerned. Given the image variability associated with each person, we next asked whether some photos capture identity better than others. To investigate this formally, we collected multiple images for a set of well-known celebrities, and asked participants to rate each photo for likeness (i.e., degree of resemblance to the depicted person). As likeness ratings allow for more graded responses than the sorting task, we anticipated some variability among these ratings, even though the faces were familiar to the raters. Our main interest was in the range of likeness ratings for each face, and its consistency across individuals.

4.1. Method

4.1.1. Stimuli

For each of 40 UK celebrities (20 males; 20 females), 12 images were downloaded from the internet (480 images in total). All the images met the criteria set out in Experiment 1. Fig. 3 shows 12 photos of Bill Clinton, which illustrate the range of within-person variability encountered.

4.2. Participants

Twenty UK undergraduates took part in the study in exchange for a small payment.

4.3. Procedure

The 480 face photographs were blocked according to identity, so that for each celebrity, all 12 photos were presented in a random sequence. Each participant received a different block order. The celebrity's name was displayed on screen throughout the block to avoid any ambiguity concerning identity. For each photo, participants were asked to provide a likeness rating using a 7-point Likert scale, where 1 indicated an extremely poor likeness, and 7 indicated an extremely good likeness. If participants were not familiar with a particular celebrity, they proceeded to the next block. No time limit was imposed for the task, and each image stayed on the screen until a response was made.

4.4. Results and discussion

Participants were familiar with 91% of the celebrities on average. For each photograph, we calculated a mean likeness score by averaging ratings across participants. We also calculated an overall likeness score for each celebrity

Fig. 3. Ambient photos of Bill Clinton. Some look more like Bill Clinton than others.

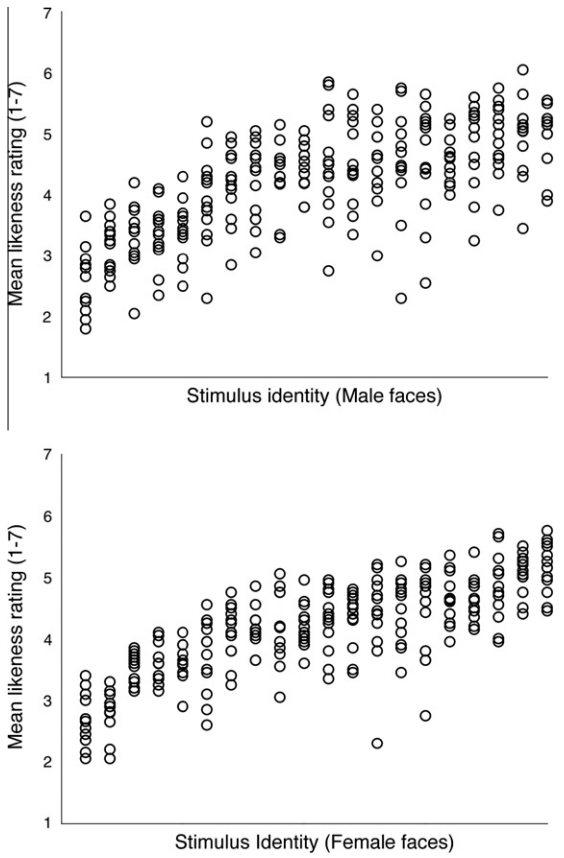

Fig. 4. Mean likeness ratings in Experiment 3, plotted separately for male faces (top panel) and female faces (bottom panel). Each column represents a single identity, and each point represents a single photo. Identities are ranked along the *x*-axis by overall likeness. See main text for details.

by pooling over photo ratings within identity. Fig. 4 shows the photo means, separately for male and female celebrities, with identities ranked on the *x*-axis in order of overall likeness.

The data contain two interesting patterns. First, there is substantial *within*-person variability. Some photographs encapsulate a person's appearance better than others, and this is true for every individual we had rated. In some ways this is curious finding. Given that a photograph captures the actual distribution of light, one might expect all photos to look like the person they depict. Instead, the variability seen here implies a continuum of resemblance, even among photographs that were good enough to be published.

The second pattern concerns the substantial *between*-person variability in likeness ratings. This may seem puzzling at first, as it seems to imply that while some people look like themselves in photographs, others do not. We suggest that the between-person differences reflect different degrees of familiarity (see Clutterbuck & Johnston, 2002, 2004, 2005). Support for this interpretation comes from a very strong correlation between overall likeness ratings for the different identities, and the proportion of participants who were familiar with those identities ($r = 0.95$, $p < 0.001$; see Fig. 5).

Presumably, celebrities who were known to all participants have received more media exposure than celebrities who were only known to some. This in turn should lead to differential levels of familiarity, even among people who know the faces. The strong correlation between familiarity and likeness implies that as a face is learned, tolerance to image variability increases, in the sense that more images are judged to be acceptable representations of the face. This accords with the findings of Experiments 1 and 2 above. It also converges with evidence from face matching tasks (Clutterbuck & Johnston, 2002, 2004, 2005).

We next conducted separate analyses for male and female faces to establish whether image variability made a significant contribution to overall variability in each case. These analyses involve a statistical comparison of two

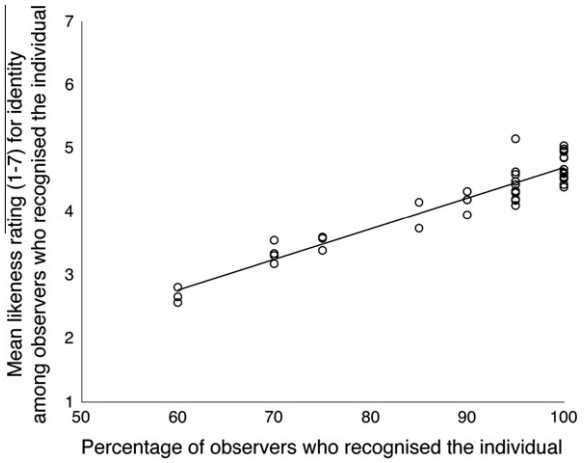

Fig. 5. Correlation between level of fame and rated likeness, using data from Experiment 3. Not all of the faces in the experiment were known to all of the observers. The *x*-axis shows the proportion of observers who were familiar with each face (i.e. the level of fame of the face). The *y*-axis shows the mean likeness ratings for each face, from observers who were familiar with them (i.e. the overall likeness rating). The correlation between level of fame and overall likeness rating is extremely reliable, suggesting that high exposure leads to high likeness ratings.

different correlations. We refer to the first correlation as the *Rank-Identity correlation*. To compute this Rank-Identity correlation, we first calculated an overall likeness rating for each identity by averaging together the mean likeness ratings for each photo of that person. We then ranked these overall likeness ratings by arranging them in ascending order. This resulted in two numbers for each identity – an overall likeness rating (ranging from 1 to 7), and a rank (an integer between 1 and 20). The Rank-Identity correlation is the correlation between these two sets of numbers. The second correlation is the *Rank-Image correlation*. This is similar to the Rank-Identity correlation, but analyses likeness data at the image level, rather than at the identity level. Here each image has a mean likeness rating (ranging from 1 to 7), and a rank (which, for each image of a person, is the rank of that person from 1 to 20 as calculated above). The Rank-Image correlation is the correlation between these two sets of numbers. We then compared the Rank-Identity and Rank-Image correlations using Fisher's *z* test to establish whether or not they were reliably different. Table 1 summarises the results of this analysis. The significant difference between the two correlations indicates that there is variability in the likeness ratings which is not accounted for by changes in identity. This confirms that different photos capture an individual's appearance to varying degrees.

We next compared the photographs against UK passport regulations (Identity & Passport Service, 2005), to test whether compliance with these regulations predicted high likeness ratings. We coded as *Acceptable* all photographs in which the subject was facing forward, looking straight at the camera, with a neutral expression and the mouth closed, showing the full head, free from shadows, without any covering. (Note that for these ambient images we had no control over lighting or background.) Photographs that violated one or more of these guidelines were coded as *Unacceptable*. We also classified the same images by emotional expression. Table 2 shows mean likeness ratings for these categories.

Table 1
Variability analysis in Experiment 3. Correlation coefficients (*r*), Fisher's *z*, and *p*-values, are shown separately for male and female faces. See text for details of this analysis.

Statistic	Male faces	Female faces
Rank-Identity *r*	0.936	0.949
Rank-Image *r*	0.683	0.746
Fisher's *z*	3.47	3.25
p	<0.01	<0.01

Acceptable images received significantly lower likeness ratings than *Unacceptable* images [$t(476) = 3.21$, $p < 0.001$, $d = .3$], indicating that passport compliant photographs captured identity especially poorly. The breakdown by facial expression suggests that the likeness cost for passport compliance can be explained in relation to open-mouth smiles. Likeness ratings were significantly higher in the *Open-mouth smile* category than in the *Neutral* category [$t(415) = 5.03$, $p < 0.001$, $d = 0.5$]. To test whether any other differences, besides the smile, could account for the passport photo cost, we also split the *Neutral* photos into *Acceptable* and *Unacceptable* subcategories using the criteria described above. Likeness ratings for these subcategories were not significantly different [$t(121) = 1.59$, n.s], suggesting that other factors make a relatively small contribution to the passport cost, compared with facial expression. This finding is consistent with previous studies showing that famous faces are easier to identify when smiling (e.g. Endo, Endo, Kirita, & Maruyama, 1992; Gallegos & Tranel, 2005; Kottoor, 1989; Sansone & Tiberghien, 1994).

5. Experiment 4

The preceding experiments demonstrate that photographs are not stable representations of facial identity. However, identity is just one of many signals that we read from the face. In this final experiment, we examined within-person variability for another socially significant signal – facial attractiveness. Previous studies of facial attractiveness have typically focused on biological variation between individuals (e.g., Perrett et al., 1998; Thornhill & Gangestad, 1999; Roberts et al., 2004; Rhodes, 2006), or the effects of isolated variables (e.g., gaze direction) on the attractiveness of an individual (e.g., Ewing, Rhodes, & Pellicano,

Table 2
Likeness ratings for the passport compliance and facial expression comparisons in Experiment 3.

Category	*N*	Likeness	SE
Acceptable	93	3.92	0.10
Unacceptable	387	4.23	0.04
Neutral	123	3.87	0.09
Open-mouth smile	292	4.34	0.05
Closed-mouth smile	35	3.89	0.14
Frown	6	3.63	0.23
Other	24	4.18	0.15

2010; Kampe, Frith, Dolan, & Frith, 2001). In such studies, identity is typically held constant across conditions in order to equate every facial variable except that which is under examination. Within-person comparisons have thus arisen incidentally, as a by-product of stimulus control, but not as a matter for study in their own right. In the present study we took a very different approach. Instead of measuring the effects of predefined variables on attractiveness ratings, we sampled the natural variation among ambient photographs. To ensure that knowledge of the individuals' characters did not influence observers' impressions, we presented only unfamiliar faces in this study. Participants made attractiveness judgements for multiple photographs of each face. We expected clear separation between faces, such that some individuals would be rated as more attractive than others. Of greater interest was the range of attractiveness ratings among photos of the same face, and its relation to variability across individuals.

5.1. Method

5.1.1. Stimuli

For each of 20 Dutch celebrities (10 males; 10 females), 20 images were downloaded from the internet (400 images in total). All of these images met the inclusion criteria set out in Experiment 1. None of the identities were known to our participants. Fig. 6 shows example photos of two other faces which illustrate the range of variability encountered.

Fig. 6. Attractiveness judgements can be reversed by photo choice. Both of the photos on the left show one person, and both of the photos on the right show another person. In the top row, most observers prefer the face on the left. In the bottom row, most observers prefer the face on the right.

5.2. Participants

Forty UK undergraduates (20 male; 20 female) took part in the study in exchange for a small payment.

5.3. Procedure

The 400 unfamiliar face photographs were separated into male and female blocks, and block order was counterbalanced across participants. Within each block, the 200 photos were presented in a random order. For each photo, participants made a Yes/No attractiveness judgement via keypress. Participants were informed that faces could appear more than once, and that attractiveness should be assessed on an image-by-image basis. No time limit was imposed for this task. Each image stayed on screen until a response was made.

5.4. Results and discussion

For each image, we calculated an attractiveness score out of 20 by aggregating 'Yes' responses across participants. We also calculated an overall attractiveness score for each person by averaging these image scores within identity. Fig. 7 shows the image attractiveness scores and the identity attractiveness scores, separately for male and female participants, and for male and female faces. The most striking finding is that, for any pair of faces, it was possible to choose photographs that reversed underlying person-level preferences.

Identity attractiveness scores were submitted to a 2×2 mixed ANOVA to test for overall sex differences. This analysis found no main effect of either subject sex [$F(1,18) = 0.65$, n.s] or stimulus sex [$F(1,18) = 2.25$, n.s.], and a significant interaction between these two factors [$F(1,18) = 11.04$, $p < .01$, $d = 1.2$]. Female participants produced significantly higher attractiveness scores for female faces ($M = 9.5$; SD = 4.6) than for male faces ($M = 6.2$; SD = 4.6) [$t(18) = 11.6$, $p < .01$, $d = 5.5$]. By contrast, male participants produced statistically equivalent attractiveness scores for females ($M = 8.6$; SD = 2.6) and for males ($M = 9.8$; SD = 4.3) [$t(18) = 1.7$, n.s.].

To test whether image variability made a significant contribution to overall variability in attractiveness scores, we compared the correlation between rank and identity score (the Rank-Identity correlation) with the correlation between rank and image score (the Rank-Image correlation). This analysis used the same procedure described in Experiment 3. Table 3 summarises the results of this analysis.

Significant differences between the correlations indicate variability in the attractiveness scores that is not accounted for by identity. Female raters tended to be rather harsh on the male faces, which somewhat compressed the distribution in that quadrant towards floor. In all other quadrants, the results confirm that facial attractiveness is not determined solely by the face, it is also determined by the photo. Indeed, for the faces used here, anyone could be more attractive than anyone else, depending on photo choice.

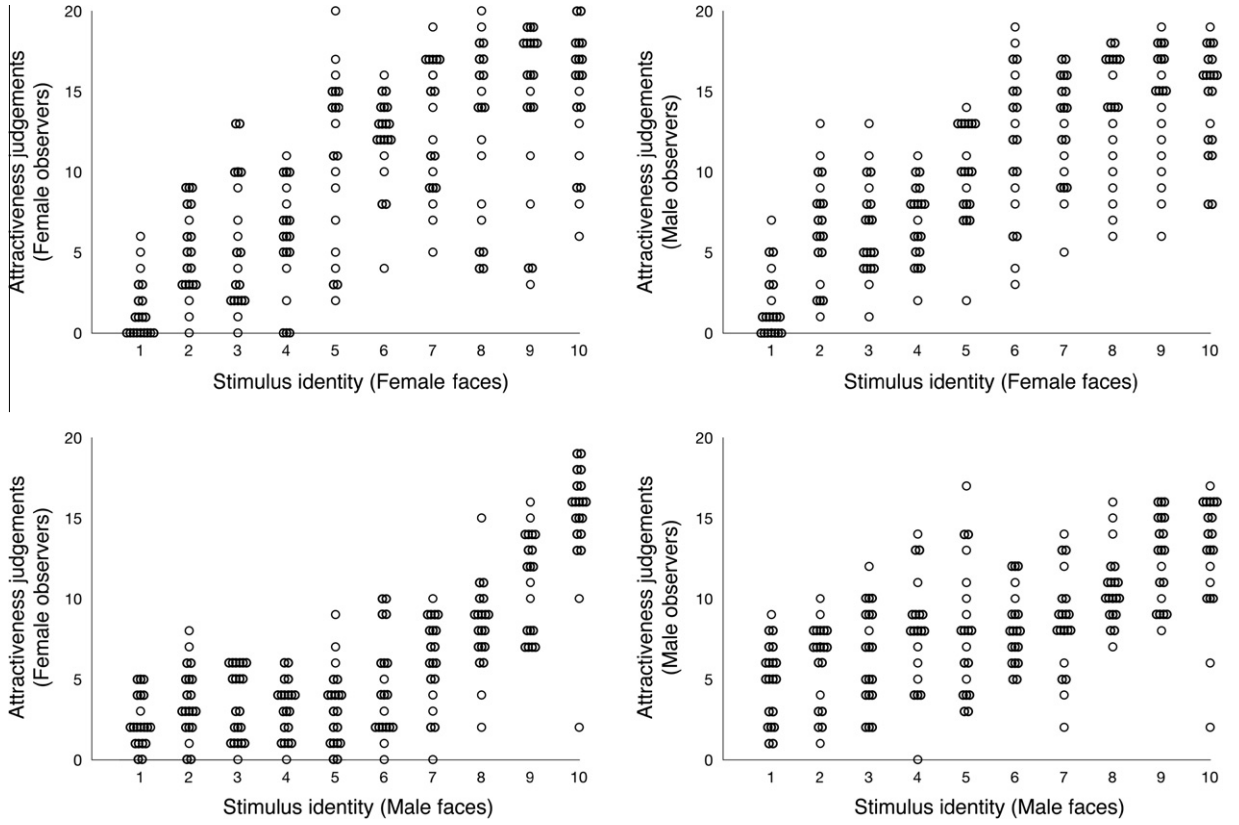

Fig. 7. Attractiveness data from Experiment 4, shown separately for female raters (Left) and male raters (Right), and for female stimuli (Top) and male stimuli (Bottom). Each column represents a single identity, and each point represents a single photograph. The *y*-axis shows aggregated attractiveness judgements (i.e. the number of observers who judged the face to be attractive). Points are spread horizontally if they would otherwise overlap. Identities are ranked on the *x*-axis by overall attractiveness scores from female raters.

Table 3
Variability analysis in Experiment 4. Correlation coefficients (*r*), Fisher's *z*, and *p*-values, are shown separately for male and female raters, and for male and female faces. See text for details of this analysis.

Statistic	Male participants		Female participants	
	Male faces	Female faces	Male faces	Female faces
Rank-Identity *r*	0.969	0.963	0.886	0.983
Rank-Image *r*	0.623	0.748	0.742	0.522
Fisher's *z*	3.5	2.64	1.17	4.68
p	<0.01	<0.01	n.s.	<0.01

6. General discussion

Four studies using ambient face photographs revealed unexpected within-person variability in appearance. In Experiment 1, photos of the same face were seen as different people. Experiment 2 confirmed that this was not due to unrepresentative photographs. In Experiment 3, multiple photographs of any individual formed a continuum of likeness, which was highly sensitive to familiarity level. Finally, in Experiment 4, within-person variability exceeded between-person variability in attractiveness: For any pair of faces, it was possible to choose photographs that reversed underlying person-level preferences. Presumably, the same might apply to other social judgements, although we did not test those here.

Everyone knows that faces vary. However, there is nothing in the psychological literature that addresses within-person variability of this scale. On the contrary, most experimental work treats *faces* and *face photographs* as interchangeable. That is a misleading oversimplification. As the present findings show, to ignore within-person variability is to miss most of the action. In light of these findings, we now consider the nature of within-person variability itself, and how it informs our understanding of face processing.

One clear implication of our card sorting data is that variability in photos of the same face greatly exceeds the level of variability expected by observers, when the face is unfamiliar. Without exception, observers mistook photos of the same person as photos of different people, often

subdividing each individual into several perceived identities. This very striking finding is beyond the scope of current theorising. Historically, face perception experiments have either ignored within-person variability completely or sought to control it away. As a result, many studies have only examined processing of 'neutral' photos, in which each photographic subject affects a blank expression, and is captured using the same camera under matched environmental conditions. We suggest that in seeking to minimise within-person variability, this convention controls away the core problem. Computer-based models of face space have allowed stimulus control to be taken to its logical conclusion: In many influential instantiations of face space (e.g. Blanz & Vetter, 1999; Todorov, Said, Engell, & Oosterhof, 2008), every point in the space represents a distinct person, so any possible move in the space is a change of identity. In this situation, variability within a face simply cannot arise. Importantly, these models are not merely theoretical – they have been used to generate experimental stimuli for scores of psychological experiments (e.g. Leopold, O'Toole, Vetter, & Blanz, 2001; Todorov et al., 2008); and data from these experiments have been used to inform theoretical development.

We propose that rather than trying to eliminate within-person variability, we should try to understand it, and incorporate it into our theorising (Burton, Jenkins, & Schweinberger, in press). A complete theory of face recognition should thus explain not only how we tell people apart, but also how we tell people *together*. We are not the first to argue that within-person variability should be taken seriously. In a pioneering paper, Bruce (1994) suggested that exposure to such variability may be necessary for building up a stable representation of a person's appearance. Our own position is very much in tune with this proposal (Jenkins & Burton, 2011; Burton et al., in press). To date however, there has been rather little experimental work investigating links between variability in input images and acquisition of stable face representations.

The issue of exposure brings us to the second clear finding from the current experiments: Familiarity with a face completely transforms our ability to accommodate within-person variability. This is evident from the contrast between Experiment 1, in which unfamiliar observers erroneously perceived many identities in the card sorting task, and Experiment 2, in which familiar observers correctly perceived just two identities. Many previous studies have reported contrasting performance for familiar and unfamiliar face recognition (e.g. Jiang, Blanz, & O'Toole, 2007; see Johnston & Edmonds, 2009 for a review). To our knowledge however, this is the first directly to associate contrasting recognition performance with contrasting tolerance of within-person variability. The graded nature of this association was revealed by Experiment 3, in which likeness ratings were monotonically higher for better known celebrities. This pattern is consistent with previous demonstrations of dose effects of exposure on matching performance (Clutterbuck & Johnston, 2002, 2004, 2005). The present findings point to a mechanism for this improvement, by showing that familiarity increases the range of images that count as the individual concerned. In cognitive terms, this corresponds to increasing proliferation of many-to-one links from input images to face representations (e.g. FRUs); or funnel-like connectivity with a better catchment area.

A number of theoretical implications flow from these findings. Foremost, they suggest a specific formulation of familiarity, as understanding all the ways in which a particular face can vary. This formulation implies that variability must be understood for each face separately, rather than for faces as a unitary class of objects. By contrast, the debate in the literature concerning face expertise has tended to consider expertise for the entire class (Bukach et al., 2006; McKone et al., 2007). Our findings also demonstrate that variability is not just a problem of input, but also a problem of representation, as observers with contrasting levels of familiarity respond to the same range of variability very differently. Future accounts of face representation will have to accommodate this representational component.

The consequences of within-person variability are not confined to judgements of identity. They also extend to social signals and impression formation, as illustrated here for attractiveness. Experiment 4 revealed that within-person variability in attractiveness was large compared with between-person variability, such that ranking of faces by attractiveness could be reversed by appropriate photo selection. How generally this finding applies to other face sets an open question, but note that the faces in the current study were not chosen to be uniformly attractive. Indeed they included political commentators and sports personalities who are not necessarily famed for their good looks. Much of the influential research on facial attractiveness has emphasised anatomical predictors of attractiveness ratings, such as facial symmetry and averageness (e.g. Fink & Penton-Voak, 2002; Perrett et al., 1999; Rhodes, Zebrowitz, et al., 2001; Thornhill & Gangestad, 1999). Presumably, such anatomical differences account for some of the between-person variability in our data. However, they cannot account for the observed within-person variability, for which anatomy is held constant. Although some within-person comparisons can be gleaned from the literature, these typically involve simple binary or parametric manipulations of isolated variables, such as smiling versus neutral expression (Mehu, Little, & Dunbar, 2008), or direct versus averted gaze (Kampe et al., 2001). Such studies are designed to assess the effects of specific factors, rather than to characterise the full range of variability found in the real world. How best to achieve the latter is a matter for ongoing research. For now we simply note that some smiling images and some direct gaze images received low attractiveness ratings, while some unsmiling images and some averted gaze images received high attractiveness ratings (cf. Fig. 6). On the strength of these informal observations, we anticipate that other image factors will turn out to be at least as important in determining perceived attractiveness, as well as other socially significant attributes.

In saying this, it is important to emphasise that we are not advocating a systematic exploration of one parameter after another, as they come to mind. What we are advocating is a genuine sampling of the variability that occurs in the world, such that the eventual characterisation of face variability is shaped by statistical data, rather than by *a priori* assumptions. This is an important distinction for at

least two reasons. First, the readiness with which a parameter springs to mind is not necessarily proportional to the amount of image variability it explains. For example, changes in gaze direction are generally salient, but account for rather little image variability (Burton et al., in press). Second, parametric manipulation of a given variable may not reflect the actual distribution of cases. For instance, directions of gaze that are equally likely in an experiment may not be equally likely in daily life. The more general point is that *statistical analysis* of images should operate on a *statistical sample* of images, if it is to structure the variability that is actually encountered. Note that this emphasis on sampling is closely entwined with the graded nature of familiarity. Observers are not simply familiar or unfamiliar with a face, they are familiar over the range of variability that they have experienced. Accordingly, a colleague might recognise your adult face, but not your childhood face; a school friend might recognise your childhood face, but not your adult face; your parents might recognise both, and a customs officer neither. In this paper we have used rather an arbitrary sample of naturally occurring images. If our general approach is correct, then sampling will need to become a serious focus of future research. Indeed, it may prove fruitful to examine parallels between this research effort and statistical approaches to understanding expertise in other domains, such as language processing (e.g. Redington & Chater, 1998). For now, one interesting implication for the face domain is that observer familiarity, which has already been intensively studied in the context of identification, might turn out to be relevant for other aspects of face perception, such as perception of attractiveness. Although effects of familiarity on attractiveness have been reported before (Bornstein, 1989; Peskin & Newell, 2004; Rhodes, Halberstadt, & Brajkovich, 2001), previous studies have typically presented a single image of each face, and have not considered the within-person variability examined here.

Beyond these theoretical concerns, within-person variability has important practical implications. For example, attractiveness not only predicts mating success (Thornhill & Gangestad, 1999), it also influences evaluations of personality and performance (Dion, Berscheid, & Walster, 1972; Landy & Sigall, 1974), as well as employment prospects (Dipboye, Arvey, & Terpstra, 1977). Given that attractiveness varies widely from one photo to the next, it matters which photos we use to present ourselves to the world. Interestingly, a number of consultancies now offer professional advice on photo selection for commercial websites.

Our analysis of photographic likeness also questions the utility of photographs in proof of identity documents. Standards for passport photographs are set out by the International Civil Aviation Organisation (ICAO). In many countries, passport applications require a countersignature to certify that the photograph is a true likeness of the applicant. The countersignatory is required to have known the applicant for a minimum period (e.g. 2 years in the UK). This condition acknowledges that only somebody who is familiar with the applicant is qualified to judge a photographic likeness. However, our data suggest a catch: To a familiar observer virtually any photograph will be a good likeness, which rather defeats the purpose.

The celebrity photos that elicited the highest likeness ratings in Experiment 3 were those showing an open mouth smile. There are a number of reasons why a smile might have helped. One possibility is that our perceptual experience of celebrities is dominated by smiling images (see Table 2), so that their smiles are incorporated into our representations of their faces (Burton et al., in press; Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2008). On this account, a smiling photo is a good likeness because it is a close match to the stored representation. Of broader applied interest is whether this effect generalises beyond celebrities. Passport guidelines explicitly prohibit smiling, on the grounds that "Laughing or smiling distorts the normal facial features" (Identity & Passport Service, 2005). Our findings suggest that the opposite is sometimes closer to the truth: Faces usually smile, and posing a neutral expression distorts the normal facial features.

To summarise the situation for photo-ID: Some photos look like their subjects, but others do not. Smiling photos show some promise, but these are banned from identity documents. Certification of likeness is pointless, because the familiarity required to judge likeness elevates likeness ratings. ICAO guidelines stipulate that "passport photographs must meet internationally agreed standards and must be a true likeness" (Identity & Passport Service, 2005). The present findings suggest that it is difficult to satisfy both of these conditions simultaneously.

Within-person variability is a neglected topic in face perception research. As long as this continues to be the case, theories of face perception will be missing half the story, and experimental work will yield misleading results. That is not a good platform for explaining the cognitive bases of face perception, or for addressing their applied implications. We anticipate that a better understanding of within-person variability will lead to significant advances in both of these areas.

Acknowledgements

This research was supported by an ESRC grant to Jenkins & Burton (RES-062-23-0549), and an ESRC grant to Burton & Jenkins (RES-000-22-2519). We thank Rachael Main for assistance in collecting data for Experiment 4. Our thanks also to three anonymous reviewers for helpful comments on an earlier version of this paper.

Appendix I. Solution to Fig. 2

ABAAABABAB
AAAAABBBAB
BBBAAABBAA
BABAABBBBB

References

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194.

Bornstein, R. F. (1989). Exposure and affect: Overview and metaanalysis of research, 1968–1987. *Psychological Bulletin, 106*, 265–289.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*, 105–116.
Bruce, V. (1994). Stability from variation: The case of face recognition the MD Vernon memorial lecture. *Quarterly Journal of Experimental Psychology, 47A*, 5–28.
Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.
Bukach, C. M., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: The power of an expertise framework. *Trends in Cognitive Sciences, 10*, 159–166.
Burton, A. M., Jenkins, R., & Schweinberger, S. R. (in press). Mental representations of familiar faces. *British Journal of Psychology*.
Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*, 256–284.
Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*, 286–291.
Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor quality video: Evidence from security surveillance. *Psychological Science, 10*, 243–248.
Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception, 31*, 985–994.
Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition, 11*, 857–869.
Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology, 17*, 97–116.
D'Argembeau, A., Van der Linden, M., Comblain, C., & Etienne, A. (2003). The effects of happy and angry expressions on identity and expression memory for unfamiliar faces. *Cognition & Emotion, 17*, 609–622.
Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*, 285–290.
Dipboye, R. L., Arvey, R. D., & Terpstra, D. E. (1977). Sex and physical attractiveness of raters and applicants as determinants of resume evaluations. *Journal of Applied Psychology, 4*, 288–294.
Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*, 576–585.
Dyer, A., Neumeyer, C., & Chittka, L. (2006). Honeybee (*Apis mellifera*) vision can discriminate between and recognise images of human faces. *Journal of Experimental Biology, 208*, 4709–4714.
Endo, N., Endo, M., Kirita, T., & Maruyama, K. (1992). The effects of expression on face recognition. *Tohoku Psychologica Folia, 52*, 37–44.
Ewing, L., Rhodes, G., & Pellicano, E. (2010). Have you got the look? Gaze direction affects facial attractiveness. *Visual Cognition, 18*, 321–330.
Fink, B., & Penton-Voak, I. S. (2002). Evolutionary psychology of facial attractiveness. *Current Directions in Psychological Science, 11*, 154–158.
Gallegos, D. R., & Tranel, D. (2005). Positive facial affect facilitates the identification of famous faces. *Brain and Language, 93*, 338–348.
Golarai, G., Ghahremani, D. G., Whitfield-Gabrieli, S., Reiss, A., Eberhardt, J. L., Gabrieli, J. D., et al. (2007). Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Neuroscience, 10*, 512–522.
Gupta, R., & Srinivasan, N. (2008). Emotions help memory for faces: Role of whole and parts. *Cognition & Emotion, 23*, 807–816.
Identity and Passport Service (2005). Guidelines for passport photographs. <http://www.ips.gov.uk/cps/files/ips/live/assets/documents/photos.pdf>.
Jenkins, R., & Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science, 319*, 435.
Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B, 366*, 1671–1683.
Jiang, F., Blanz, V., & O'Toole, A. J. (2007). The role of familiarity in view transferability of face identity adaptation. *Vision Research, 47*, 525–531.
Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*, 577–596.
Kampe, K. K., Frith, C. D., Dolan, R. J., & Frith, U. (2001). Reward value of attractiveness and gaze. *Nature, 413*, 589.
Kottoor, T. M. (1989). Recognition of faces by adults. *Psychological Studies, 34*, 102–105.
Landy, D., & Sigall, H. (1974). Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology, 4*, 299–304.
Leopold O'Toole Vetter & Blanz (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience, 4*, 89–94.
McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences, 11*, 8–15.
Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*, 865–876.
Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*, 364–372.
Mehl, B., & Buchner, A. (2008). No enhanced memory for faces of cheaters. *Evolution & Human Behavior, 29*, 35–41.
Mehu, M., Little, A., & Dunbar, R. (2008). Sex differences in the effect of smiling on social judgments: An evolutionary approach. *Journal of Social, Evolutionary, and Cultural Psychology, 2*, 103–121.
Perrett, D. I., Burt, D. M., Penton-Voak, I. S., Lee, K. J., Rowland, D. A., & Edwards, R. (1999). Symmetry and human facial attractiveness. *Evolution and Human Behavior, 20*, 295–307.
Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., et al. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature, 394*, 884–887.
Peskin, M., & Newell, F. N. (2004). Familiarity breeds attraction: Effects of exposure on the attractiveness of typical and distinctive faces. *Perception, 33*, 147–157.
Redington, M., & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language and Cognitive Processes, 13*, 129–191.
Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*, 199–226.
Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects in social stimuli. *Social Cognition, 19*, 57–70.
Rhodes, G., Zebrowitz, L. A., Clark, A., Kalick, S. M., Hightower, A., & McKay, R. (2001). Do facial averageness and symmetry signal health? *Evolution and Human Behavior, 22*, 31–46.
Roberts, S. C., Havlíček, J., Flegr, J., Hruskova, M., Little, A. C., Jones, B. C., et al. (2004). Female facial attractiveness increases during the fertile phase of the menstrual cycle. *Proceedings of the Royal Society of London B, 271*, 270–272.
Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*, 252–257.
Sansone, S., & Tiberghien, G. (1994). Facial expression coding and face recognition: Independent or interactive processes. *Psychologie Francaise, 39*, 327–344.
Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences, 3*, 452–460.
Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*, 455–460.
Tsukiura, T., & Cabeza, R. (2011). Remembering beauty: Roles of orbitofrontal and hippocampal regions in successful memory encoding of attractive faces. *NeuroImage, 54*, 653–660.
Turati, C., Bulf, H., & Simion, F. (2008). Newborns' face recognition over changes in viewpoint. *Cognition, 106*, 1300–1321.
Winston, J. S., Henson, R. N., 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.