Representation of facial identity includes expression variability

Annabelle S. Redfern*, Christopher P. Benton

School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK

ARTICLE INFO

ABSTRACT

In this study, we investigate the contribution of expression variability in the formation of face representations. We trained participants to learn new identities from face images either low or high in expressiveness, and compared their performance in a recognition test. After low expressiveness training, recognition of novel test images was modulated by image expressiveness: the more expressive the image, the slower the response. This differed from recognition after high expressiveness training, which showed little evidence of expression dependence. These findings are not readily explained by exemplar and prototype theories of face representation. However, we propose that our results can be explained by a combination of these theories, according to which average and exemplar representations co-exist – the latter of which preserve expressions and other within-person variability. We conclude that this study provides evidence that variability of expressions is, therefore, incorporated in the representation of an individual's face. Moreover, our results demonstrate that learning to recognise someone from their face entails learning how their face is changed by expressions.

1. Introduction

Faces constantly change. They change from moment to moment because of lighting, orientation and expressions; and over longer time periods because of ageing, facial hair, adiposity, health, cardio-vascular activity and cosmetics. How we deal with these changes in appearance, in order to recognise someone, seems to be partly governed by how well we know their face. When faces are familiar, recognition seems invariant to change and even degraded face images can be identified with ease (Burton, Jenkins, & Schweinberger, 2011). However when faces are *un*familiar, recognition is readily compromised by changes in appearance (Bruce, 1982).

How does a face become familiar? It makes intuitive sense that more experience of a face would assist recognition. However, it is not simply amount of exposure to a new face that determines how well it is learned, but also variability. For example, recognition of newly learnt faces from novel images has been shown to be better when the faces were learnt from multiple images, than when they were learnt from a restricted number of images (Murphy, Ipser, Gaigg, & Cook, 2015). Indeed, the emerging consensus is that experiencing the different ways in which a face can vary in appearance may be essential for learning that identity (Andrews, Jenkins, Cursiter, & Burton, 2015; Baker, Laurence, & Mondloch, 2017; Bindemann & Sandford, 2011; Burton, Kramer, Ritchie, & Jenkins, 2016; Dowsett, Sandford, & Burton, 2016; Menon, White, & Kemp, 2015; Ritchie & Burton, 2017). This raises the question of whether all types of variability help in the learning process,

or are some superfluous or even an impediment.

Precisely how variability contributes to face learning is uncertain, although perceiving a face across an extensive range of views and expressions may be important for the formation of face representations (Johnston & Edmonds, 2009). Bruce (1994) proposed the concept of 'stability from variation', which argues that it is the differences between instances of a face – its variability – that enables us to identify which aspects are unchanging and therefore stable, and which are transient. This facilitates discrimination between structural, more permanent features of the face and superficial, changeable aspects. Furthermore, within-person variations – such as from expressions – provide characteristic information about a particular face that helps define the 'possible and permissible' (Bruce, 1994; Vernon, 1952) ways in which it may change, thereby defining the boundary between possible instances of one face from those of another.

In this study, we investigate within-person variability in the formation of face representations. However, unlike most other studies that have looked at this question, we explore the specific contribution to this from a single source of variability: facial expressions. Our experimental approach is to do this through unfamiliar face learning. In the experiment reported below, we compare performance in a face recognition task, after participants have learned new faces with images that are low in expressiveness and after learning the faces from images high in expressiveness. The variability of expressions, although evident in both training conditions, is more pronounced in the high expressiveness condition because of the greater variability of facial distortions within

these images. Because the novel images we use in the test phase range extensively in expressiveness, we are able to measure the expression-dependence of test performance.

The approach described above is similar to that of Redfern and Benton (2017a), a study which showed that learning faces from neutral images led to slower and less accurate subsequent recognition of those faces, when they were expressive. However, the demonstration by Redfern and Benton (2017a) of expression dependence following neutral face-learning, left open the question of how performance after expressive face-learning may differ; which, in turn, may reveal how the underlying face representations, derived from these different training regimes, may also differ. The current study addresses this question, with emphasis on the comparison of expression-dependence in the two learning regimes.

In order to draw inferences about the underlying face representations, we need to understand how theories of face representation might account for a difference between the two learning conditions, when participants are tasked with recognising newly learned faces. We outline theoretical approaches and what they might predict for the current study. We then consider how expression variability may be important for face learning, according to these theoretical approaches.

Broadly we can conceive of two categories of face representation theory: exemplar and prototype. We describe and consider these theories with the assumption that expressions and facial identity are not analysed completely separately. We believe this to be reasonable, given the wealth of evidence suggesting that identity and expression processing interact (Campbell & Burke, 2009; Ellamil, Susskind, & Anderson, 2008; Fisher, Towler, & Eimer, 2016; Fox, Moon, Iaria, & Barton, 2009; Ganel & Goshen-Gottstein, 2004; Ganel, Valyear, Goshen-Gottstein, & Goodale, 2005; Hildebrandt, Schacht, Sommer, & Wilhelm, 2011; Yankouskaya, Booth, & Humphreys, 2012; Yankouskaya, Humphreys, & Rotshtein, 2014; Endo, Endo, Kirita, & Maruyama, 1992; Rhodes et al., 2015); and given evidence suggesting that expressions may indeed form an intrinsic part of identity representation (Burton et al., 2011; Calder, 2011; Calder & Young, 2005; Kaufmann & Schweinberger, 2004; Redfern & Benton, 2017a, 2017b).

According to exemplar theories, familiarity is achieved from storing multiple separate instances of a face across different views, more instances increasing the chance that one will be a close match when a novel instance is perceived (Longmore, Liu, & Young, 2008). This approach would argue that storing many instances of a face would thus enhance face recognition performance (Murphy et al., 2015). In terms of the current study, exemplar theory would predict that we should find a performance advantage with participants who have undergone low expressiveness training when they are tested with the more neutral, as opposed to expressive, novel images – because these would provide closer matches to the low expressiveness exemplars stored during training. For the same reason, there should be a performance advantage for the high expressiveness-trained participants with the expressive, compared to neutral, test images.

According to prototype (also called averaging) theories, we develop robust representations of facial identities by averaging instances of them to form a prototype for each person. With successive additions to the prototype, transient and superficial image properties that are identity-irrelevant become eliminated, while stable characteristics are reinforced (Burton, Jenkins, Hancock, & White, 2005). In this way, the prototype inherently prioritises constant facial aspects over those that change. Recognition is achieved when the viewed face is matched to a stored prototype. Murphy et al., (2015) reason that recognition accuracy has been linked to quality of the formed average (e.g. Jenkins & Burton, 2008), and that a better quality average is more likely when derived from many, as opposed to few, observations. The plausibility of this approach is demonstrated by Kramer, Ritchie, and Burton (2015); their study shows superior recognition for averaged (composite) faces derived from four previously-seen exemplars of an individual, compared to recognition for composite faces derived from four unseen

exemplars of that person.

For the current study, prototype theory would predict no performance difference between our two face learning conditions, since both training regimes would have resulted in very similar prototypes. This premise is supported by Jenkins and Burton's (2011) finding that a stable face average for an individual emerges from composites of only a dozen or so variable images. By 'stable', these authors explain that an average based on 10 variable images of an individual is much the same as one based on a different set of 10 variable images of that person, irrespective of the variability inherent in the images used. Further, averaging 20 variable face images was found to improve automatic face recognition from 54% to 100% accuracy (Jenkins & Burton, 2008) and an average from 20 images was robust to errors – so little changed by incorporating several images of different people (Jenkins, Burton, & White, 2006). In the present study, we use 35 different images of each face during each training regime, which should be sufficient to establish stable prototypes.

Both exemplar and prototype theories of face representation can explain how expression variability – indeed, within-person variability – contributes to learning to recognise a face; but only the former necessarily includes expression variability within the representation of facial identity. Exemplar theory incorporates the variability of facial expressions since it indiscriminately incorporates *all* variability. The face representation is enhanced by variability of expressions, because the differing expressions extend the range of aspects by which a face can be matched to a stored instance. Prototype theory does not readily incorporate expression variability, because expressions would be averaged out as successive instances are combined. For this account of face representation, the important factor is to view multiple instances of a face – irrespective of expression.

Exemplar and prototype theories are broad approaches that explain face representation in a general sense, describing the representation of multiple faces in a single system. These contrast with the person-specific coding space account of face representation, developed by Burton et al. (2016), which explains how an *individual* facial identity may be represented. The authors do not commit to a general recognition system that computes between-person variability in the same way, pointing out that this research is at an early stage. Their account builds on earlier work that proposed the variability of a face is a part of its representation (Burton et al., 2011). According to this concept, each facial identity is represented by its own coding space. Coding space is identity-specific, defined by "bespoke axes" (Burton et al., 2016, p. 207), which are PCA-computed dimensions of within-person variability – and therefore include expressions. Expression – and other – variability is therefore essential for a face representation to be generalizable, facilitating recognition of new instances of that face.

Importantly, Burton et al. (2011, 2016) advocate the notion that variability is not 'noise', but is informative – echoing the 'stability from variation' concept that we outlined earlier. This approach has been demonstrated in the growing number of studies that embrace variability in order to investigate it; and do so by using 'ambient' images (Burton et al., 2016; Jenkins, White, Van Montfort, & Burton, 2011; Murphy et al., 2015; Ritchie & Burton, 2017; Sutherland et al., 2013). These are unmanipulated, naturalistic face images taken from the environment, that incorporate extensive within-person variability and are of the sort we encounter every day. In the present study, we use ambient face images to explore the role of variability of expressions, in the formation of face representations. Selected from our own databases, these images encompass an extensive range of facial expressions.

Studies investigating the contribution of expressions to face recognition have tended to consider expressions as defined by the 'basic' emotion categories (Ekman, 1992; Ekman & Friesen, 1971) of happy, sad, anger, surprise, disgust and fear (e.g. Chen et al., 2015; Liu, Chen, & Ward, 2014, 2015). However, expressions need not be emotional. The tendency in the literature to conflate expressions with these 'basic' emotions, perhaps neglects the study of those many other expressions –

gesticulations and other transient face changes during social communication – that are so substantial a part of our day-to-day experience of faces. Therefore, for the purpose of this study, we use this wider definition of expressions, incorporating faces that have been judged 'expressive' but do not necessarily convey emotion.

We trained participants to learn facial identities from ambient images with low expressiveness and from those high in expressiveness. They were subsequently tested with novel images of the learned identities. In the experiment described below, we found that neutral training led to recognition responses that were modulated by expressiveness, with response times slowing as expressiveness increased. This contrasted with performance after expressive training, which showed little evidence of expression-dependence.

2. Method

2.1. Developing the stimuli databases

We created 2 databases of facial images that incorporate extensive variability: external variation (e.g. lighting), image capture variations (e.g. image resolution, camera type) and person-specific variation (e.g. expressions, age, hair style, facial hair, adiposity). Database 1 comprised 546 ambient facial images of 2 actors, Luigi Lo Cascio and Fabrizio Gifuni, actors with extensive filmographies but little known in the UK. Database 2 comprised 816 ambient facial images of the 2 actors, Christian Tramitz and Sven Nordin, also actors relatively little known in the UK.

Database 1 images were obtained from YouTube screenshots and the DVDs of 13 movies made between 2002 and 2014. For database 2, the images were from YouTube and the DVDs of 4 television series and 13 movies made between the years 1985 and 2012. As per the method used by Jenkins et al. (2011), images exceeded 150 pixels in height and showed faces of frontal or partial view that were free of occlusion. Images were cropped to portrait dimensions of 4:5 and sized to 320×400 pixels.

Images were collected in 'Image Groups', sets of 2–9 faces for which the camera, position and scene are the same. This ensures that aspects such as lighting and image capture are kept essentially constant so that Image Group faces differ only in expression. Author AR selected 'expressive' frames, attempting to find both those that showed the greatest facial distortion and/or affect, and those that were unexpressive but matched expressive screenshots in terms of other image variables. Copyright restrictions prevent us showing these images, however an illustrative example of a typical Image Group can be viewed in Redfern and Benton (2017a).

As described by Redfern and Benton (2017a), we collected expressiveness ratings for all Database 1 images. The 546 images were printed in greyscale and laminated, and participants were tasked with placing each into 1 of 5 boxes labelled from 1 ('neutral') to 5 ('very expressive'), with the box number therefore the score. These individual scores were summed and rescaled to give a percentage expressiveness rating for each image. We followed the same procedure for Database 2 but with a different 40 participants.

Participants were not provided with a definition of expressiveness or neutrality, but asked to use their judgement. This was so that their ratings would reflect and encompass a layperson's understanding of these descriptions, rather than an imposed definition. Using Spearman's rho, we compared the scores of each of the 40 raters of Database 1, with every other rater of that database; this resulted in 780 correlations. We did the same for Database 2, resulting in a further 780 correlations. The distributions of these correlations are shown in Fig. 1.

Of these comparisons, a substantial majority indicated a moderate or strong correlation; for Database 1, the number of correlations exceeding 0.4 was 95% (741/780) and for Database 2, 93% (727/780). Of these, 99% (771/780 in Database 1) and 100% (in Database 2) were statistically significant, $p < .05$. This clearly indicates that our
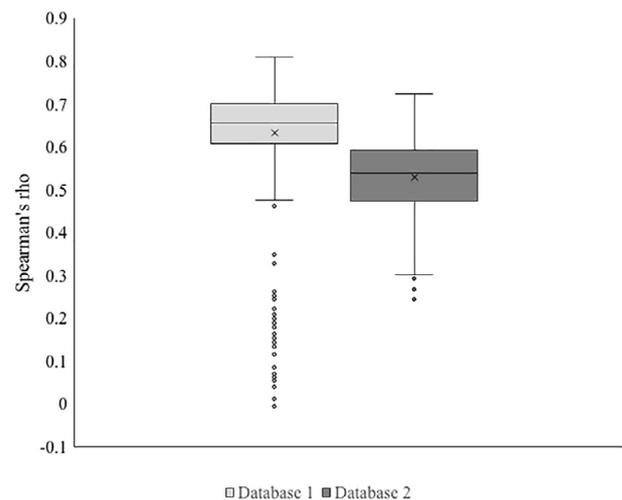

Fig. 1. Box-and-whisker plot of Spearman's rho correlations of between-rater image expressiveness scores, by database. Boxes depict median, 1st and 3rd quartiles, 'X' denotes mean; error bars and circles indicate variability outside the upper and lower quartiles and outliers.

participants did not classify randomly. However, it is also clear that our images elicited varying degrees of consensus from our raters. This implies that, with some images, people were working from different definitions of expressiveness. This, in turn, points to a larger problem of defining expressiveness in ambient images such as those used in the present study. Expressions gathered from the environment show far greater complexity and variability than the carefully constrained and manipulated expressions that can be found in many studies (e.g. Skinner & Benton, 2010).

2.2. Stimuli selection

From both databases we selected neutral and expressive training images comprising 70 image pairs (35 for each actor), from Image Groups with the highest range of expressiveness. In each pair, one image was low in expressiveness, the other high. For each database we split the 70 pairs into two sets: a neutral training set comprising images of < 50% expressiveness, and an expressive training set of images > 50% expressiveness. Therefore the training sets were matched for all variation types except for expressiveness. That is to say, the images of the two sets differed only in how expressive they were. See Fig. 2 for an illustrative example.

Author AR looked through the images to ensure our selection of high expressive training images included the 6 universal expressions (Ekman & Friesen, 1971) and as equal a balance as reasonably possible in the ratio of positive to negative affect expressions (Database 1, 50:20 images; Database 2, 36:34 images). Since we had conducted this selection on Database 1 some time before Database 2 was created, we selected images from Database 2 that resembled the expressiveness ratings of the Database 1 sets as closely as possible. Each 'test' set comprised 208 images (104 of each actor) ranging in expressiveness. All test images were selected from Image Groups other than those used as the source of stimuli for training, so as to ensure that they did not closely resemble those images. Table 1 shows the descriptive statistics of the selected stimuli by Database.

2.3. Participants

We tested 85 naïve participants and rejected the data of 5 (see 'data analysis' below). The remaining 80 participants, of whom 16 were male, had a mean age of 19 years (range 18–31 years). All were undergraduates who received course credit for participating and all gave informed consent. None were familiar with the actors whose images we

Fig. 2. Illustrative example low expressiveness stimuli (upper row) and matched high expressiveness stimuli (lower row) of actors in the motion picture, *"Suddenly"* (Basseler & Allen, 1954). Images on the left are of Frank Sinatra, and those on the right are Sterling Heyden. Expressiveness ratings (from left to right from top): 14.4%, 31.3%, 51.9%, 35.0%, 53.8%, 65.6%, 73.8%, and 82.5%.

Table 1
Descriptive statistics of stimuli.

	Perceived expressiveness%			
	Minimum	Maximum	Mean	SD ±
Database 1				
Low expressiveness training images × 70	9.38	49.38	22.64	9.18
High expressiveness training images × 70	53.75	100.00	79.22	11.18
Test images × 208	6.25	100.00	42.08	19.52
Database 2				
Low expressiveness training images × 70	5.63	46.88	24.72	9.83
High expressiveness training images × 70	51.25	98.75	75.46	11.80
Test images × 208	9.38	100.00	45.46	20.21

used as stimuli, which was confirmed during debrief. The work was carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). We had calculated a sample size of 76 + participants based on an alpha of 0.05, beta of 0.8 and effect size of $d = 0.46$. This was the size of the effect reported in Experiment 2 in a similar study by Redfern and Benton (2017a), for the difference in accuracy between responses to low and high expressiveness images. This experiment used the same face classification task and database 1 stimuli as the current study and, although this effect size is based on accuracy, whereas here we analyse differences in regression slopes, both are a measure of the extent to which expressiveness modulates face classification performance. We tested more than 76 participants because we recruited more, anticipating attrition.

2.4. Equipment

The experiment was conducted in a quiet dark room in which participants sat at a computer. Stimuli were presented on a monitor, screen resolution 1280 × 1024 and a refresh rate of 85 Hz. The stimuli were centrally displayed on a 39.3 cd/m² grey background and, from the viewing distances of ~100 cm, subtended 5.6° × 7.0°. Responses were made on a Microsoft SideWinder gamepad. The experiment was coded in Matlab and used the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

2.5. Design and procedure

The experiment was in 4 parts, with a word search filler task performed for 5 min or so between parts. The 1st part was a training session in which the identities of 2 actors were learned. The 2nd part was a test session on those 2 identities. The 3rd and 4th parts comprised another training and test session, but on the Database not used for parts 1 and 2. Therefore both Databases were used and participants learned all 4 identities. Each participant learned one pair of identities from training with low expressiveness face images (the 'low' condition) and the other pair from training with highly expressive face images with variable expressions ('high' condition). The order of database and condition was counterbalanced across participants. The 'test' sessions were the same for all participants, irrespective of whether they had learned the faces in the low or the high expressiveness condition.

For the training sessions, the participants were presented with an image of a face on the computer screen and their task was to respond quickly and accurately with a right key-press if they thought the image was of 'Louis' (for Database 1 images, 'Chris' for Database 2), or a left key-press if they thought the face belonged to 'Rob' (for Database 1, 'Steve' for Database 2). Although their first trial was inevitably a guess, after every response they received feedback in the form of a tick or cross, which remained on the screen for 0.4 s. There were a total of 8 blocks of 70 trials, with each image of the training set presented once in each block. To prevent sequential presentation of the same image, the set of 70 was randomised in the following way: for each participant, the set was randomly split into halves, each half containing 18 images of one actor, 17 of the other. With each half set randomised for each training block, and presented such that one half always preceded the other, a minimum of 35 images between 2 presentations of each image was ensured. There were opportunities for breaks every 20 trials.

The test sessions employed the same task except without feedback, consequently each response triggering the next trial. In both test sessions the 208 test stimuli were presented once, and their order was randomised. There were break opportunities after every 26 trials. In total, the experiment took about an hour to run.

2.6. Data analysis

For each participant we took the proportion correct data for block 8 of the training phase and converted it to z-scores. We excluded the data of 5 participants who had accuracy z-scores lower than -2, which corresponds to 71% accuracy. Data were trimmed as follows: for each participant, the mean reaction time ("RT") was calculated, then RTs that were more than $2 \pm$ standard deviations from the mean were excluded. RTs to incorrect responses were also excluded from the RT data.

Normality tests on the mean RT data indicated that 3 of the 8 variables were non-normal (in the Training phase: Database 1 expressive training; and in the Test phase: Database 1 neutral, and Database 2 expressive data). We addressed this by inverse-transforming the mean RT data. Subsequent tests on the transformed data indicated normality, with non-significant Shapiro-Wilk tests for all 8 variables. Data analyses were conducted on the inverse-transformed data; when graphed, these are transformed back for interpretability. Analyses and graphs are by database so as to ensure that we are comparing faces belonging to the same actors and are, therefore, measuring only the manipulation of neutral versus expressive faces. Consequently, all comparisons are between-subject and by training condition.

3. Results

3.1. Training phase

Training phase results are shown in Fig. 3, which plots accuracy data to indicate that there was no speed-accuracy trade-off. We conducted a 2×8 mixed ANOVA of the RT data for each Database, in which the within-subjects factor was training block (1 to 8) and between-subjects factor was training type (low, high expressiveness). In summary, these analyses showed that RT performance improved across the course of training and that, although there was no difference in Database 2 performance between training conditions, Database 1 training performance was inferior in the high expressiveness condition compared to low.

For Database 1 (Fig. 3, upper panel) there was a main effect of training block indicating significant performance improvement across blocks, $F(7, 546) = 73.24, p < .001, \eta_p^2 = 0.484$. The main effect of training type indicates that the performance with low expressiveness training images was superior to training with high expressiveness images, $F(1, 78) = 7.63, p = .007, \eta_p^2 = 0.089$. There was a non-significant interaction between training condition and block where partial eta squared indicated a small effect size, $F(7, 546) = 1.87, p = .072, \eta_p^2 = 0.023$.

For Database 2 training data (Fig. 3, lower panel), there was a main effect of training block, $F(7, 546) = 66.37, p < .001, \eta_p^2 = 0.460$, indicating that participant performance improved significantly across the blocks. There was no main effect of training type however, indicating that overall performance did not differ between training conditions, $F(1, 78) = 0.42, p = .518, \eta_p^2 = 0.005$. There was no interaction, $F(7, 546) = 0.30, p = .952, \eta_p^2 = 0.004$.

The significant difference in training condition performance for Database 1, but not Database 2, raises the possibility of some difference in performance between the two databases; however, this difference may not itself achieve statistical significance (Nieuwenhuis, Forstmann, & Wagenmakers, 2011). If there were such a difference, we would be uncertain of its cause. However, for the present study, the purpose of

the Training Phase was to train participants to learn the facial identities. Whilst this may have been slower in the Database 1 high expressiveness condition, the overall pattern of the Training Phase data is that, in both conditions and for both databases, performance steadily improves over the course of training.

3.2. Test phase

Fig. 4 (upper panel) shows RT and accuracy results of the test phase. For each database, we compared the variables: test phase RT following low expressiveness training, and test phase RT following high expressiveness training, using a between-samples *t*-test. These revealed that RT performance for the conditions was not significantly different, for both Database 1, $t(78) = 1.20, p = .233, d = 0.27$; and for Database 2, $t(78) = 0.23, p = .818, d = 0.05$.

For each database, we investigated the relationship between image expressiveness and mean RTs of correct test phase responses, comparing these for the low and high expressiveness training conditions. For each participant, we conducted an ordinary least squares linear regression, to estimate straight-line fits of the test phase RT data against image expressiveness. Fig. 4 (lower panel) shows the mean regression slopes by training type. We compared the regression slopes, to ascertain if they differed according to training type (high expressiveness, low expressiveness), and whether performance was expression-dependent. As a cautionary measure, we also conducted a robust linear regression using Matlab's *robustfit* function to minimise the effects of outliers (Holland & Welsch, 1977; Huber, 1981; Street, Carroll, & Ruppert, 1988). We conducted this for the purpose of verification, so that we could be satisfied that outliers did not drive our outcomes.

For Database 1, a between-samples *t*-test that compared the ordinary regression slopes of the low and high expressiveness training conditions, showed that there was a significant difference, $t(78) = 2.06, p = .043, d = 0.46$. The robust regression gave the same pattern of results, $t(78) = 2.14, p = .036, d = 0.48$. When considered with Fig. 3 (lower panel), this difference can be interpreted as showing that, after low expressiveness training, test phase responses were more affected by image expressiveness than they were after high expressiveness training, with RTs slowing when image expressiveness increased.

We found the same pattern of results with Database 2 ordinary regression slope data. For Database 2, a between-samples *t*-test that compared the ordinary regression slopes of the low and high expressiveness training conditions, showed that there was a significant difference, $t(78) = 2.87, p = .005, d = 0.64$. The robust regression result followed the same pattern, $t(78) = 2.68, p = .009, d = 0.60$. This difference between conditions, in combination with the Database 2 columns in Fig. 4 (lower panel), indicates that low expressiveness training, more than high expressiveness training, led to performance more affected by image expressiveness, with RTs slowing when image expressiveness increased.

In sum, participants are slower to recognise expressive faces when they learned those identities from images low in expressiveness, as opposed to high. We considered the possibility that the training phase results might underlie the test phase results. That is to say, might the difference in the slopes between the low and high expressiveness conditions be explained by the performance levels attained during training? However, this explanation is flawed; it would be unable to account for the same pattern of test phase results we found with both databases, given the evident lack of any substantial difference between the conditions in the Database 2 training phase.

For the high expressiveness conditions, the 95% confidence intervals in Fig. 4 includes zero for Database 1, and comes very close to zero for Database 2. Looking across these high expressiveness results there seems little evidence for any substantial effect of test expressiveness on RT. Considered together, these comparisons converge to demonstrate that responses following low expressiveness training were more sensitive to image expressiveness, whereas high expressiveness training led
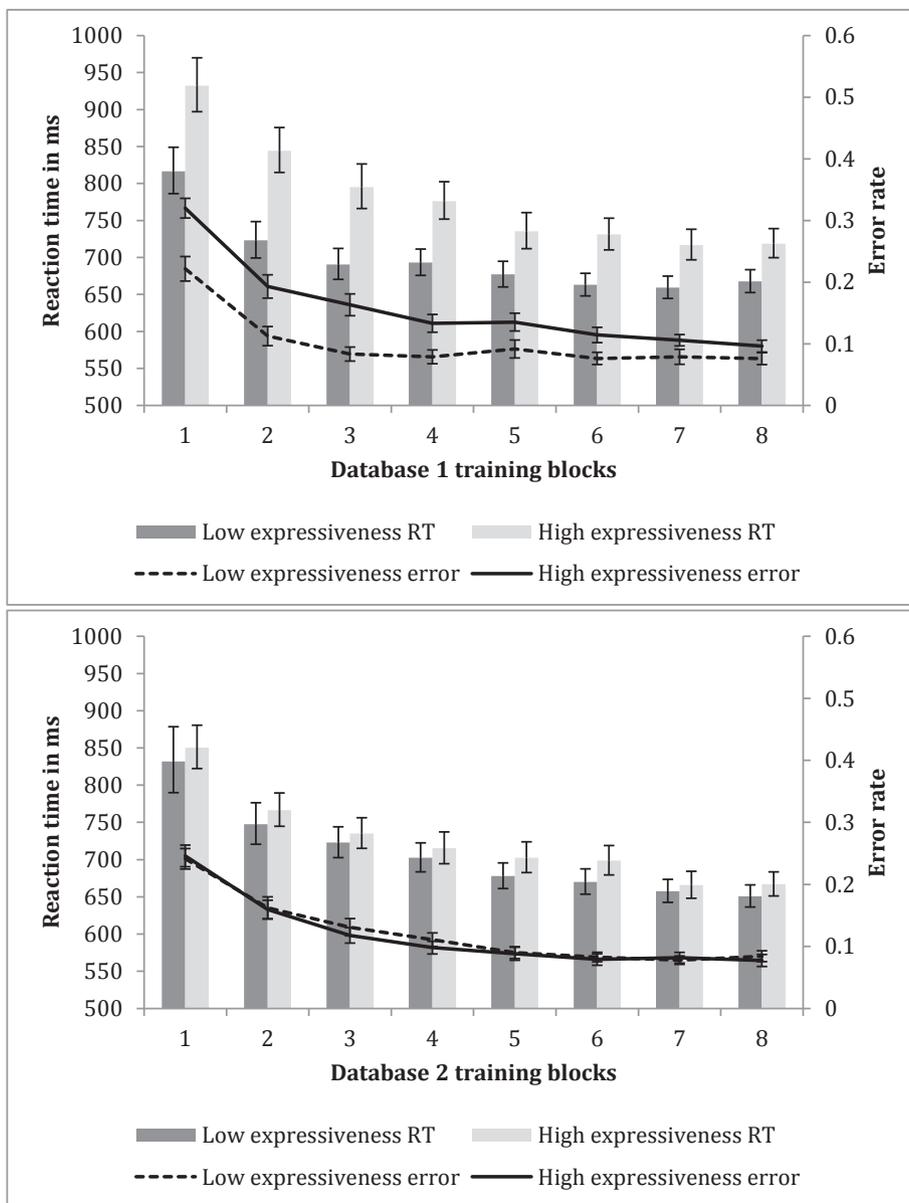
Fig. 3. Training phase results. *Upper panel:* Reaction time and accuracy data for Database 1. *Lower panel:* Reaction time and accuracy data for Database 2. Error bars denote standard error.

to a more stable response pattern.

4. Discussion

We used ambient images to train participants to learn new facial identities under two conditions: from face images low in expressiveness, or from highly expressive face images. We compared how these training regimes affected subsequent recognition of the faces, and we measured whether recognition performance was modulated by the expressiveness of the test images. We found that after training with low expressiveness faces, performance was affected by image expressiveness: responses became slower as expressiveness increased. In contrast, expressive training led to performance that was significantly less dependent on the expressiveness of the test images, with reaction times showing little response to it.

Before discussing the implications of these findings for face representation, we consider the possibility that other factors might be driving our outcomes. It could be the case that the high variable expressions of our stimuli capture more attention than those of low

expressiveness, either directly, from the expression itself, or indirectly through the emotion elicited by the expression. The difference in attentional response may underlie the difference between the training conditions at test. Of course, different types of expression intensity have different affects. Wilson and MacLeod (2003) found that attention was oriented away from mildly threatening faces, but towards those that were strongly threatening. Moreover, different types of expression may have different affects. D'Argembeau and Van der Linden (2011) found that angry faces, but not happy or neutral, had a disruptive effect on face recognition; and Gallegos and Tranel (2005) found that recognition of familiar famous faces was faster when they had happy compared to neutral expressions. Given the balance in our expressive training stimuli between positive and negative emotional affect, as well as the variable nature of the expressions themselves, attentional responses could plausibly vary between one expressive image and the next.

However, in their review paper on face perception and attention, Palermo and Rhodes (2007) summarise evidence that strongly suggests emotional facial expressions, particularly threatening ones, receive enhanced processing. Setting aside the complication that our variable
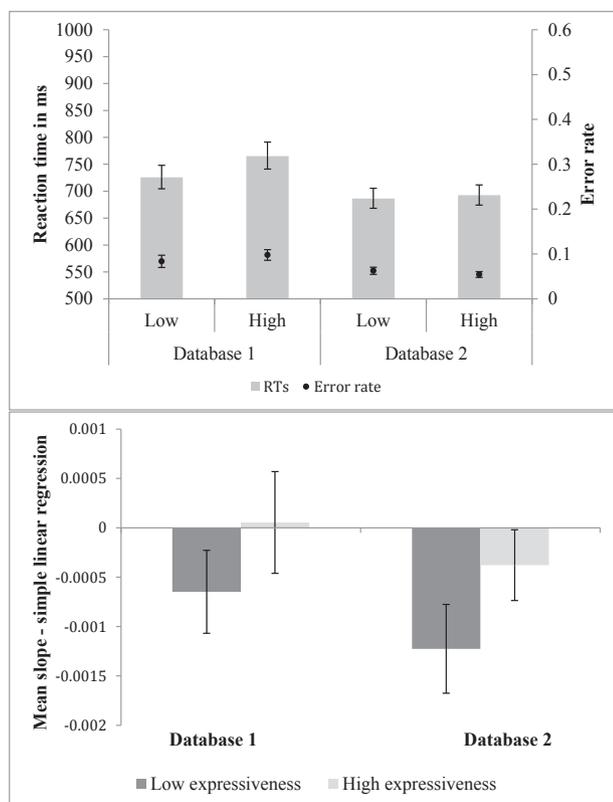
Fig. 4. Test phase results. *Upper panel:* Reaction time and accuracy data by database, using the same axes ranges as Fig. 3 for ease of comparison. Error bars denote standard error. *Lower panel:* Mean slopes of simple linear regressions. Data analysis of the slopes was conducted on inverse-transformed RT data; therefore the more negative the mean slope, the greater the increase in RT as a consequence of image expressiveness. Error bars denote 95% confidence limits.

expressions may have varying affects, what might enhanced processing of expressions predict for our experiment? If expressions enhance the initial perceptual encoding of faces, we would expect high expressiveness training to result in superior recognition of the faces at test, compared to low expressiveness training. However, we measured no difference between overall reaction times between conditions, which is inconsistent with this prediction. More troubling for this explanation, it does not account for our finding of expression-dependence. Because this followed *low expressiveness* training – and was almost entirely absent after high expressiveness training – it cannot be attributed to emotion or attention to expressions because – by their very nature – the low expressiveness training images displayed expressions mildly, if at all.

We now turn to consider theoretical approaches, which may provide a more compelling explanation. Earlier we outlined exemplar and prototype theories of face representation, and the predictions they might make for this study. However, returning to these, we see that prototype theory does not readily explain our data and exemplar theory can only partially explain it.

Prototype theory would predict no difference between the performance of the low and high expressiveness training groups, which was clearly not the case. Indeed, the experiment yielded the same outcome pattern from both face databases, with significantly different low expressiveness condition responsiveness to test image expressiveness, compared to the high expressiveness condition. An exemplar theory of face representation cannot fully explain our results. It would propose that after low expressiveness training, the less expressive test images – closer in expression to the stored instances – would be quicker to recognise than the high expressiveness test images. Our results are compatible with this prediction. However, an exemplar account would also predict that high expressiveness training would result in low

expressiveness test images being more slowly recognised, for the same reason. Our findings did not confirm this, as we found little evidence of expression dependence following high expressiveness training.

There is an alternative theoretical position that explains our findings. This is a combination approach, in which both averaging of instances of a face takes place, as well as the storage of individual exemplars. As an idea, this is not new. Bruce and Young (2012, p. 299) speculate that a combination of averaging, and storage of separate instances of faces, may "prove the best way forward" in terms of theoretical approaches to how faces become familiar. These authors suggest that similar images within views may be averaged, while separate instances are also retained – an approach that they compare to the distinction between structural and pictorial codes for familiar faces, drawn in their seminal paper on face recognition (Bruce & Young, 1986).

Simultaneous exemplar and average face representations have been demonstrated experimentally, with both familiar (Neumann, Schweinberger, & Burton, 2013) and unfamiliar faces (Kramer et al., 2015), although it remains unclear whether the demonstrated averaging is evidence of a general ensemble encoding mechanism, or the formation of stable identity representations. Although the idea of a face representation comprising both an average and exemplars is seemingly incompatible (Neumann et al., 2013), it is intuitively appealing; since our experience of a face can include memories of specific instances, as well as a general sense of how it looks.

This combination approach can readily explain our results. Following low expressiveness training, participants would have formed a low expressiveness average, and have stored low expressiveness exemplars of the newly learned faces. At test, we would expect relatively faster recognition of low expressiveness images – being a closer match to both representation types – compared to the recognition of high expressiveness images; and this is what we found, for both face databases. Following high expressiveness training, participants would have formed a 'neutral' average, the expressions having been cancelled out; and they would have stored highly expressive exemplars. From this, we would predict that there would be no particular recognition advantage for either low or high expressiveness test images; and our findings are consistent with this. Therefore, we suggest this account can explain the difference between the two training regimes at recognising high expressiveness test images, and propose that high expressiveness training conferred 'stability from variation' (Bruce, 1994) of expressions.

A combined average-and-exemplar explanation preserves a degree of within-person variability in the representation of individual instances. In doing so, it is consistent with the concept that part of learning a new face entails learning how it varies (Young & Burton, 2017). This, in turn, dovetails with Burton et al. (2016) person-specific coding space account of face representation – outlined earlier – for which stability from variation is a fundamental concept. What is surely needed is a reconciliation of Burton et al. (2016) person-specific account of how a single identity representation may develop, with a general recognition system.

5. Conclusion

We investigated face learning using ambient face images of the type we encounter every day. We demonstrate that encountering a wide range of expressions and expressiveness in the learned faces, compared to learning with faces low in expressiveness, led to performance less affected by image expressiveness. This may be because the highly expressive images contain more identity-specific variation that acts as a cue to identity, enabling people to develop facial identity representations that are robust to the challenge of previously unseen expression variability. Our findings are not readily explained by exemplar and prototype theories of face representation. However, we propose that a combination of these theories can account for our results, according to which average and exemplar representations co-exist, the latter of which preserve expressions and other within-person variability. Our

interpretation is that our results demonstrate that the generalizability of a face representation is, at least partly, based on the variability it putatively incorporates. Showing this specifically with expression variability confirms how important it is as an identity cue. Furthermore, it suggests that learning *how* expressiveness changes appearance is an important part of learning to recognise new faces.

Author contributions

A. S. Redfern and C. P. Benton developed the study concept and study design. Programming code was written by A. S. Redfern and revised by C. P. Benton. Testing, data collection and data analysis was performed by A. S. Redfern. A. S. Redfern drafted the manuscript under the supervision of C. P. Benton, who provided critical revisions. Both authors approved the final version of the manuscript for submission.

Declaration of interest

Conflicts of interest: none.

Acknowledgment

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology, 68*(10), 2041–2050. http://dx.doi.org/10.1080/17470218.2014.1003949.

Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition, 161*, 19–30. http://dx.doi.org/10.1016/j.cognition.2016.12.012.

Basseler, R (Producer), and Allen, L (Director), (1954). Suddenly [Motion picture]. US: Libra Productions Inc.

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception, 40*(5), 625–627. http://dx.doi.org/10.1068/p7008.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. http://dx.doi.org/10.1163/156856897X00357.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*(1), 105–116. http://dx.doi.org/10.1111/j.2044-8295.1982.tb01795.x.

Bruce, V. (1994). Stability from variation: The case of face recognition the M.D. Vernon memorial lecture. *The Quarterly Journal of Experimental Psychology Section A, 47*(1), 5–28. http://dx.doi.org/10.1080/14640749408401141.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327. http://dx.doi.org/10.1111/j.2044-8295.1986.tb02199.x.

Bruce, V., & Young, A. W. (2012). *Face perception* (1st ed.). Hove: Psychology Press.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256–284. http://dx.doi.org/10.1016/j.cogpsych.2005.06.003.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology, 102*(4), 943–958. http://dx.doi.org/10.1111/j.2044-8295.2011.02039.x.

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*(1), 202–223. http://dx.doi.org/10.1111/cogs.12231.

Calder, A. (2011). Does facial identity and facial expression recognition involve separate visual routes? In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.). *Oxford handbook of face perception* (pp. 427–448). Oxford: Oxford University Press.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews. Neuroscience, 6*(8), 641–651. http://dx.doi.org/10.1038/nrn1724.

Campbell, J., & Burke, D. (2009). Evidence that identity-dependent and identity-independent neural populations are recruited in the perception of five basic emotional facial expressions. *Vision Research, 49*(12), 1532–1540. http://dx.doi.org/10.1016/j.visres.2009.03.009.

Chen, W., Liu, C. H., Li, H., Tong, K., Ren, N., & Fu, X. (2015). Facial expression at retrieval affects recognition of facial identity. *Frontiers in Psychology, 6*, 780. http://

dx.doi.org/10.3389/fpsyg.2015.00780.

D'Argembeau, A., & Van der Linden, M. (2011). Influence of facial expression on memory for facial identity: Effects of visual features or emotional meaning? *Emotion, 11*(1), 199–202.

Dowsett, A. J., Sandford, A., & Burton, A. M. (2016). Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. *Quarterly Journal of Experimental Psychology, 69*(1), 1–10. http://dx.doi.org/10.1080/17470218.2015.1017513.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200. http://dx.doi.org/10.1080/02699939208411068.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124–129. http://dx.doi.org/10.1037/h0030377.

Ellamil, M., Susskind, J. M., & Anderson, A. K. (2008). Examinations of identity invariance in facial expression adaptation. *Cognitive, Affective, & Behavioral Neuroscience, 8*(3), 273–281. http://dx.doi.org/10.3758/CABN.8.3.273.

Endo, N., Endo, M., Kirita, T., & Maruyama, K. (1992). The effects of expression on face recognition. *Tohoku Psychologica Folia, 51*, 37–44.

Fisher, K., Towler, J., & Eimer, M. (2016). Facial identity and facial expression are initially integrated at visual perceptual stages of face processing. *Neuropsychologia, 80*, 115–125. http://dx.doi.org/10.1016/j.neuropsychologia.2015.11.011.

Fox, C., Moon, S., Iaria, G., & Barton, J. (2009). The correlates of subjective perception of identity and expression in the face network: An fMRI adaptation study. *NeuroImage, 44*(2), 569–580. http://dx.doi.org/10.1016/j.neuroimage.2008.09.011.

Gallegos, D., & Tranel, D. (2005). Positive facial affect facilitates the identification of famous faces. *Brain and Language, 93*(3), 338–348.

Ganel, T., & Goshen-Gottstein, Y. (2004). Effects of familiarity on the perceptual integrality of the identity and expression of faces: The parallel-route hypothesis revisited. *Journal of Experimental Psychology: Human Perception and Performance, 30*(3), 583–597. http://dx.doi.org/10.1037/0096- 1523.30.3.583.

Ganel, T., Valyear, K. F., Goshen-Gottstein, Y., & Goodale, M. A. (2005). The involvement of the "fusiform face area" in processing facial expression. *Neuropsychologia, 43*(11), 1645–1654. http://dx.doi.org/10.1016/j.neuropsychologia.2005.01.012.

Hildebrandt, A., Schacht, A., Sommer, W., & Wilhelm, O. (2011). Measuring the speed of recognising facially expressed emotions. *Cognition and Emotion, 26*(4), 650–666. http://dx.doi.org/10.1080/02699931.2011.602046.

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods, 6*(9), 813–827. http://dx.doi.org/10.1080/03610927708827533.

Huber, P. J. (1981). *Robust statistics.* New York: Wiley http://dx.doi.org/10.1002/0471725250.

Jenkins, R., & Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science (New York, N.Y.), 319*(5862), 435. http://dx.doi.org/10.1126/science.1149656.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 366*(1571), 1671–1683. http://dx.doi.org/10.1098/rstb.2010.0379.

Jenkins, R., Burton, A. M., & White, D. (2006). Face recognition from unconstrained images: Progress with prototypes. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)* (pp. 25–30). . http://dx.doi.org/10.1109/FGR.2006.45.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313–323. http://dx.doi.org/10.1016/j.cognition.2011.08.001.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*(5), 577–596. http://dx.doi.org/10.1080/09658210902976969.

Kaufmann, J. M., & Schweinberger, S. R. (2004). Expression influences the recognition of familiar faces. *Perception, 33*(4), 399–408. http://dx.doi.org/10.1068/p5083.

Kramer, R. S. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision, 15*(4), 1. http://dx.doi.org/10.1167/15.4.1.

Liu, C. H., Chen, W., & Ward, J. (2014). Remembering faces with emotional expressions. *Frontiers in Psychology.* http://dx.doi.org/10.3389/fpsyg.2014.01439.

Liu, C. H., Chen, W., & Ward, J. (2015). Effects of exposure to facial expression variation in face learning and recognition. *Psychological Research Psychologische Forschung, 79*(6), 1042–1053. http://dx.doi.org/10.1007/s00426-014-0627-8.

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34*(1), 77. http://dx.doi.org/10.1037/0096-1523.34.1.77.

Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception, 44*(11), 1332–1341. http://dx.doi.org/10.1177/0301006615599902.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577–581. http://dx.doi.org/10.1037/xhp0000049.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128*(1), 56–63. http://dx.doi.org/10.1016/j.cognition.2013.03.006.

Nieuwenhuis, S., Forstmann, B., & Wagenmakers, E. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience, 14*(9), 1105.

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia, 45*(1), 75–92. http://dx.doi.org/10.1016/j.neuropsychologia.2006.04.025.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. http://dx.doi.org/10.1163/156856897X00366.

Redfern, A. S., & Benton, C. P. (2017a). Expression dependence in the perception of facial

identity. *I-Perception, 8*(3), http://dx.doi.org/10.1177/2041669517710663 2041669517710663.

Redfern, A. S., & Benton, C. P. (2017b). Expressive faces confuse identity. *I-Perception, 8*(5), 1–21. http://dx.doi.org/10.1177/2041669517731115.

Rhodes, G., Pond, S., Burton, N., Kloth, N., Jeffery, L., Bell, J., ... Palermo, R. (2015). How distinct is the coding of face identity and expression? Evidence for some common dimensions in face space. *Cognition, 142*, 123–137. http://dx.doi.org/10.1016/j.cognition.2015.05.012.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70*(5), 897–905. http://dx.doi.org/10.1080/17470218.2015.1136656.

Skinner, A. L., & Benton, C. P. (2010). Anti-expression aftereffects reveal prototype-referenced coding of facial expressions. *Psychological Science, 21*(9), 1248–1253. http://dx.doi.org/10.1177/0956797610380702.

Street, J. O., Carroll, R. J., & Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician, 42*(2), 152–154. http://dx.doi.org/10.1080/00031305.1988.10475548.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A.

W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105–118. http://dx.doi.org/10.1016/j.cognition.2012.12.001.

Vernon, M. D. (1952). *A further study of visual perception* (1st ed.). Cambridge: Cambridge University Press.

Wilson, E., & MacLeod, C. (2003). Contrasting two accounts of anxiety-linked attentional bias: Selective attention to varying levels of stimulus threat intensity. *Journal of Abnormal Psychology, 112*(2), 212–218.

Yankouskaya, A., Booth, D., & Humphreys, G. (2012). Interactions between facial emotion and identity in face processing: Evidence based on redundancy gains. *Attention, Perception, & Psychophysics, 74*(8), 1692–1711. http://dx.doi.org/10.3758/s13414-012-0345-5.

Yankouskaya, A., Humphreys, G. W., & Rotshtein, P. (2014). The processing of facial identity and expression is interactive, but dependent on task and experience. *Frontiers in Human Neuroscience, 8*, 920. http://dx.doi.org/10.3389/fnhum.2014.00920.

Young, A. W., & Burton, A. M. (2017). Recognizing faces. *Current Directions in Psychological Science, 26*(3), 212–217. http://dx.doi.org/10.1177/0963721416688114.