CRediT author statement

Thomas Murray: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization

Justin O'Brien: Supervision, Funding acquisition

Noam Sagiv: Supervision

Lúcia Garrido: Conceptualization, Writing – Review & Editing, Supervision

The role of stimulus-based cues and conceptual information in processing facial expressions of emotion

Thomas Murray[1], Justin O'Brien[1], Noam Sagiv[1], Lúcia Garrido[2]

[1]Centre for Cognitive Neuroscience, Department of Life Sciences, Brunel University London

[2]Department of Psychology, City, University of London

Abstract

Face shape and surface textures are two important cues that aid in the perception of facial expressions of emotion. Additionally, this perception is also influenced by high-level emotion concepts. Across two studies, we use representational similarity analysis to investigate the relative roles of shape, surface, and conceptual information in the perception, categorisation, and neural representation of facial expressions. In Study 1, 50 participants completed a perceptual task designed to measure the perceptual similarity of expression pairs, and a categorical task designed to measure the confusability between expression pairs when assigning emotion labels to a face. We used representational similarity analysis and constructed three models of the similarities between emotions using distinct information. Two models were based on stimulus-based cues (face shapes and surface textures) and one model was based on emotion concepts. Using multiple linear regression, we found that behaviour during both tasks was related with the similarity of emotion concepts. The model based on face shapes was more related with behaviour in the perceptual task than in the categorical, and the model based on surface textures was more related with behaviour in the categorical than the perceptual task. In Study 2, 30 participants viewed facial expressions while undergoing fMRI, allowing for the measurement of brain representational geometries of facial expressions of emotion in three core face-responsive regions (the Fusiform Face Area, Occipital Face Area, and Superior Temporal Sulcus), and a region involved in theory of mind (Medial Prefrontal Cortex). Across all four regions, the representational distances between facial expression pairs were related to the similarities of emotion concepts, but not to either of the stimulus-based cues. Together, these results highlight the important top-down influence of high-level emotion concepts both in behavioural tasks and in the neural representation of facial expressions.

Introduction

Facial expressions are associated with specific musculature movements which physically change the appearance of the face and are consistent across individuals experiencing the same emotion (Ekman & Friesen, 1971). These movements allow for changes in the overall shape of the face, such that both the shape and position of individual features may vary. For example, a smile changes the shape of the mouth, whereas a surprised face may change the position the eyebrows. Muscular movements also create differences in what is referred to as surface information, i.e. the texture of the face (Bruce & Young, 1998). As the shape changes, the pattern of light reflectance and shadows across the face also changes due to differences in the protuberance of the shape. For example, an angry scowl is associated with a lowered brow which causes a darker shadow above the eyes than other expressions. Both shape and surface information play an important role in the perception of facial expressions.

Evidence for the role of shape and texture cues in the perception of facial expressions of emotion comes from studies that assess differences in performance when a cue is disrupted or removed. For example, some studies have used contrast negation to disrupt surface cues, while keeping shape cues intact. This contrast negation had no effect on error rates during same/different judgement tasks (Harris, Young, & Andrews, 2014; White, 2001), nor on psychophysical thresholds at which an emotional face could be discriminated from a neutral one (Pallett & Meng, 2013), suggesting that the perception of facial expression relies little on surface/texture information and instead depends mostly on shape information. Similarly, other studies have shown that participants are able to recognise emotions from line drawings of facial expressions, which removes the surface cues and leaves only shape cues available (Etcoff & Magee, 1992; Katsikitis, 1997; Mckelvie, 1973). While these studies suggest that shape cues are the most important in the perception of facial expressions, there is some evidence that surface cues may also play some role. For example, Sormaz, Young and Andrews (2016) created sets of stimuli in which either shape or surface information were removed, by warping faces to the average shape (removing shape information) or by applying the average surface texture to the original shapes (removing texture information). Recognition accuracy for both sets of stimuli was well above chance but was significantly improved for original unedited images (that varied in both shape and surface information). Additionally, Calder, Burton, Miller, Young, and Akamatsu (2001) showed that facial expressions could be classified from principle components derived from shape-free images (similar to those used by Sormaz, Young, and Andrews, 2016), suggesting that differences between facial expressions are conveyed by variation in surface information. These studies suggest that shape and surface cues play separate roles in the perception of facial expressions.

Visual perception, however, is not solely a feed-forward stimulus-driven process; top-down knowledge and predictions facilitate visual perception and aid in the recognition of stimuli across multiple domains, including social categories (O'Callaghan, Kveraga, Shine, Adams, & Bar, 2017; Stolier & Freeman, 2016). For example, judgements of gender and race can be biased by expectations based on internal concepts of stereotypes (Levin & Banaji, 2006; Macrae & Martin, 2007). Evidence for the role of top-down mechanisms in the processing of facial expressions of emotion comes from studies that manipulate access to internal concepts emotions via 'semantic satiation' (see Smith & Klein, 1990). Briefly, these studies rely on the assumption that when a word (e.g. an emotion label) is read aloud a high number of times, access to the meaning of the word becomes attenuated. After repeating emotion labels aloud (and thus disrupting access to the concept of the emotion), participants were less accurate at matching pairs of facial expressions (Lindquist, Barrett, Bliss-Moreau, & Russell, 2006), and experienced reduced effects of repetition priming (Gendron, Lindquist, Barsalou, & Barrett, 2012). Further evidence comes from studies showing that the presence of emotion labels can affect the perception of facial expressions. For example, some studies have shown that if an emotion label is presented alongside a facial expression, then participants' memory of the expression they had just viewed is biased towards the

emotion corresponding to the label (Halberstadt & Niedenthal, 2001; Nook, Lindquist, & Zaki, 2015). Additionally, participants were poorer at matching facial expressions to an emotion label than they were matching them to a prototypical facial expression (Fernández-Dols, Carrera, Barchard, & Gacitua, 2008). It has been argued that the presence of emotion labels activates concepts that guide the perception of emotions from facial expressions (Nook et al., 2015).

Research so far has therefore shown that both stimulus-based cues and conceptual information influence performance on tasks requiring the perceptual matching and categorisation of facial expressions of emotion. And what happens in the brain? One possibility is that different brain regions represent the different types of information that influence emotion processing. Research using functional magnetic resonance imaging (fMRI) with Multivariate Pattern Analysis (MVPA) has demonstrated that facial expressions are decodable from their patterns of activation (Harry, Williams, Davis, & Kim, 2013; Said, Moore, Engell, Todorov, & Haxby, 2010; Wegrzyn et al., 2015; Zhang et al., 2016) within each of the core face processing regions (the Fusiform Face Area (FFA), Occipital Face Area (OFA), and Superior Temporal Sulcus (STS); Haxby, Hoffman, & Gobbini, 2000). Furthermore, using an fMRI-adaptation paradigm, Andrews, Baseler, Jenkins, Burton, and Young (Andrews, Baseler, Jenkins, Burton, & Young, 2016) showed that the FFA and OFA are sensitive to changes in face shape and surface textures between different identities. It is possible, therefore, that these cues may influence how facial expressions are represented within these regions. In contrast, there is some evidence to suggest that regions typically involved in the processing of theory of mind may represent emotions in a way that is independent from any stimulus-properties. Machine learning classifiers trained to discriminate between patterns of activation in response different emotions within one stimulus modality (e.g. voices) can successfully decode different emotions within a different modality (e.g. faces), using voxels within the Medial Prefrontal Cortex (MPFC) and STS (Peelen, Atkinson, & Vuilleumier, 2010). Similar cross-modal classification of valence was found using activation patterns in the MPFC in response to positive/negative facial expressions and animations (Skerry & Saxe, 2014), suggesting supramodal mechanisms that are not dependent on stimulus properties. While the MPFC is not typically face-responsive, research using TMS has reported interference with a number of facial expression processing tasks (Gamond & Cattaneo, 2016; Harmer, Thilo, Rothwell, & Goodwin, 2001; Mattavelli, Cattaneo, & Papagno, 2011; Wölwer et al., 2014), suggesting that this region may have some involvement in the perception of facial expressions. These lines of research could suggest that the MPFC offers some guidance to facial expression perception by representing modality-independent emotion concepts.

Several studies have used Representational Similarity Analysis (RSA) to examine the roles of stimulus-based cues and emotion concepts in both behavioural tasks and neural representations of facial expressions. RSA allows the comparison of the structure or geometry of representations across different modalities, methods, or groups (Kriegeskorte, Mur, & Bandettini, 2008), which cannot be compared directly (such as comparing brain representations with computational models or with behaviour). In RSA, the similarities (or dissimilarities) between the responses to all pairs of stimuli or conditions in one modality are computed to construct a Representational Dissimilarity Matrix (RDM). RDMs can then be compared across modalities, typically by correlating the RDMs. RSA has been previously used to investigate the role of concepts in the perception of facial expressions (Brooks & Freeman, 2018), the similarity of emotion representations across sensory modalities (Kuhn, Wydell, Lavan, McGettigan, & Garrido, 2017) and the integration of high- and low-level representations in judgements of personality traits from faces (Stolier, Hehman, & Freeman, 2018; Stolier, Hehman, Freeman, Keller, & Walker, 2018).

RSA has also been used to assess the relationship between perceptual similarity of facial expressions and the similarities of face shapes, surface textures, and emotion concepts. The perceptual similarity of pairs of facial expressions is often measured using subjective judgements on a seven-point scale (Said et al., 2010; Sormaz, Watson, Smith, Young, & Andrews, 2016). Sormaz, Watson, et al. (2016) computed the similarities of pairs of emotional face shapes (using Procrustes analysis) and

similarities of surface textures (using the correlation between pixel intensities) and found that the similarities of these stimulus properties predicted subjective ratings of perceptual similarity. Furthermore, this measure of perceptual similarity predicted representational similarity of expressions in the OFA, STS, but not the FFA. In another study, Said et al. (2010) also found that representational similarity of expressions in the posterior-STS was predicted by subjective ratings of perceptual similarity. Together, this research suggests a relationship between perceptual similarity and the similarity of shape and surface properties, and between perceptual similarity and representational similarity within the OFA and STS. Research has yet to examine whether shape and surface cues can directly explain the representational structures within the core face regions.

A recent study by Brooks and Freeman (2018) used RSA to assess the relationship between the perceptual similarity of facial expressions and pairwise similarities of emotion concepts. Perceptual similarity of the images was measured using two approaches. The first was the deviation in mouse-tracking trajectory during a two-choice categorisation task, where participants categorised facial expressions into one of two categories using a mouse to drag the face towards one of the category labels. The size of the deviation in trajectory towards the non-target category label is argued to reflect the degree to which that category was activated by the perception of the face (Freeman, 2018). The second approach used a reverse-correlation paradigm (Dotsch & Todorov, 2012), where participants chose which of two identical neutral faces, overlaid with different patterns of noise, displayed more of a target emotion. Patterns of noise that were chosen to display more of a given emotion were then averaged (for each emotion) and were rated by independent samples to provide the measure of perceptual similarity for each emotion pair. Conceptual similarity was measured using subjective ratings of the similarity of emotion categories, and the Euclidean distance between vectors of ratings of each emotion category on its association with a set of traits. Across all experiments, each measure of conceptual similarity predicted each measure of perceptual similarity, even when controlling for the similarities of several image properties. These results suggest that emotion concepts may influence the perception of facial expressions.

In a subsequent study, Brooks, Chikazoe, Sadato and Freeman (2019) found that conceptual similarity predicted representational similarity of facial expressions in the brain. Using a whole-brain searchlight, results showed that the representational structure of expressions within a region of the right fusiform gyrus was explained by the similarity of emotion concepts, after controlling for three measures of visual similarity of the stimuli (the similarity of face silhouettes, similarity of pixel intensity maps, and similarity of 'higher-level visual features' output from a computational model of object recognition). The searchlight analysis showed no other regions that survived correction for multiple comparisons. These results suggest that representations of facial expressions in the right fusiform gyrus may be structured according to the similarity of emotion concepts. Together, these studies show that the perceptual similarity of facial expressions can be predicted by the similarities of face shapes, surface textures, and emotion concepts, suggesting that these cues may influence the perception of facial expressions. Furthermore, they suggest that the representation of facial expressions within the core face regions may reflect the perceptual similarity of expressions (Said et al., 2010; Sormaz, Watson, et al., 2016) and the conceptual similarity of emotions (Brooks et al., 2019).

While these studies provide an understanding of the independent roles of stimulus-based cues and conceptual information in the perception of facial expressions, there are several unanswered questions. The first is whether the conceptual similarity of emotions still explains perceptual similarity when emotion labels are not present in the perceptual task. The explicit use of emotion labels in tasks designed to measure perceptual similarity could have led participants to rely more on, or make explicit use of, conceptual information (Fernández-Dols et al., 2008; Halberstadt & Niedenthal, 2001; Nook et al., 2015). For example, in the mouse tracking task from Brooks and Freeman (2018), the changes in mouse trajectories could be affected by the conceptual similarity of the emotion labels. As perception likely involves the interaction between multiple bottom-up and

top-down processes, it is unlikely that any perceptual task (e.g. a 3AFC task) will fully isolate the use of perceptual processes from any higher-level cognition. Despite this, ideally a perceptual task designed to assess the ability to discriminate between facial expressions using perceptual processes would avoid the use of any emotion labels. Such a task would provide a more rigorous test of whether emotion concepts influence the perception of facial expressions.

The second question that remains unanswered concerns the roles of these cues in the explicit categorisation of facial expressions. The research outlined above has mostly measured perceptual similarity using 'perceptual tasks', although commonly used tasks in emotion processing research are categorisation tasks. Palermo, O'Connor, Davis, Irons and McKone (2013) distinguish between perceptual tasks (e.g. same/different, 3 alternate forced choice, etc.) and labelling tasks, by suggesting that labelling tasks require initial perception plus additional cognitive processes needed to assign an emotion label. Palermo et al. (2013) reported that performance on a perceptual matching task was correlated with performance on a labelling task. Furthermore, performance on the labelling task was correlated with performance on a vocal emotion labelling task, whereas performance on the perceptual matching task was not. Together, this suggests that the perceptual and labelling tasks share perceptual processes, whereas the labelling task also taps into a cognitive process used to assign emotion labels to a stimulus, regardless of the stimulus modality. These results suggest that perceptual tasks and labelling tasks may recruit different processes, or rely to different extents on different processes. It may therefore be of interest to examine the role of conceptual and stimulus-based cues in a task in which participants are required to explicitly assign an emotion label to the stimulus.

Finally, a further question that research has yet to address is the roles of these cues in explaining how facial expressions are represented in the brain. While previous research shows that perceptual similarity of expressions can be explained by the similarity of stimulus-properties (Sormaz, Watson, et al., 2016), and that representational similarity in the OFA and STS can be explained by perceptual similarity (Said et al., 2010; Sormaz, Watson, et al., 2016), research has yet to examine the direct role of the stimulus-properties in explaining representational similarity. Furthermore, as conceptual cues also play a role in explaining representations of facial expressions in the FFA (Brooks et al., 2019), we do not know the relative roles of these different sources of information, in different regions of the brain.

The aim of the current research is therefore to address these three questions. We conducted a behavioural experiment (Study 1) and an fMRI experiment (Study 2), using RSA in both. Model RDMs were constructed measuring the similarities of face shapes, surface textures, and emotion concepts, and were used to predict perceptual similarity, patterns of categorisation, and the similarity of neural representations of facial expressions. In Study 1 we addressed the first two research questions, assessing the role of these cues in explaining perceptual similarity of facial expressions (measured using a task with no emotion labels), and patterns of categorisation in an emotion labelling task. As shape, surface, and conceptual cues have all been shown to play a role in the perception and recognition of emotions, we expect all three of these cues to explain behaviour in the perceptual and categorisation tasks. Given that the presence of emotion labels may activate emotion concepts, we expect that conceptual cues may play more of a role in the categorical task than in the perceptual task. In Study 2, we addressed the third research question, using RSA and fMRI to examine the ability of these three models to explain the representational structure of facial expressions in different brain regions. We have chosen to examine three core face regions (the FFA, OFA, and STS), and a region that may represent stimulus-independent emotion concepts (the MPFC). Given the research showing that representational similarities within the core face regions is explained by perceptual similarity (which is itself explained by the similarity of stimulus properties; Said, Moore, Engell, et al., 2010; Sormaz, Watson, et al., 2016), we expect that the shape and surface models can predict representational similarities within these regions. As representations of emotions in the MPFC are likely modality independent (Peelen et al., 2010; Skerry & Saxe, 2014), we

expect that the similarities of emotion concepts will explain representational similarities in this region. We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

Study 1

Methods

Participants

Fifty participants (40 women, mean age: 19.02 years, S.D: 1.70, range: 18-27) were recruited using the Brunel University London Psychology's Participant Pool, and by word of mouth. Sample size was determined based on previous studies with similar participant sizes, although no formal power analyses were conducted. All participants reported normal or corrected-to-normal vision, no history of stroke, other neurological conditions, or diagnosed emotion processing disorders. Ethical approval to conduct the research was granted by the College of Health and Life Sciences Research Ethics Committee at Brunel University London.

Design

This study used a within-subjects design, so all participants completed both the Perceptual and the Categorical Tasks. The Perceptual Task was conducted first, so that the participants were only exposed to the emotion labels during the Categorical task.

Perceptual task

Materials

We selected pictures of facial expressions of emotion from the Radboud Faces Database (Langner et al., 2010). Legal copyright restrictions prevent public archiving of the stimuli used in this study, which can be obtained from the Radboud Faces database: http://www.socsci.ru.nl:8180/RaFD2/RaFD. Three male identities and three female identities were chosen based on mean ratings on genuineness and intensity for the 6 basic emotions (using validation data from Langner et al., 2010). Identities with above average means (across all 6 emotions) and below average standard deviations for both genuineness and intensity ratings across the 6 expressions were selected. Only Caucasian adult identities were used (RaFD model numbers = 1, 8, 9, 23, 30, 58). In total, there were 36 different pictures (6 identities x 6 expressions).

For each identity, we created morphed continua between each pair of six emotions. Each face was marked with 112 fiducial points (using the positions in the FantaMorph software; www.fantamorph.com) to allow for linear interpolation between every pair of expressions, within identity (Figure 1A). Although FantaMorph was used to position the points, custom MATLAB scripts were used to create the continuum, which contained 100 discrete steps. Each face image was tessellated using Delaunay Triangulation to maximise the internal angles across triangles. Morphing between images was achieved by weighting the triangular mesh coordinates of the two images and determining, by interpolation, the colour values of the source images. Images were cropped to a square containing the whole face. All images were presented in full colour.

Procedure

During each trial, participants simultaneously viewed 3 faces aligned horizontally (Height/Width = 8.0°, Width of 3 images = 28.4°, with an average viewing distance of 60cm), each displaying one of two expressions. Beneath each face were numbers 1 to 3, and participants were required to indicate which one displayed a different expression from the other two (odd one out), by pressing the corresponding key (1-3) on a standard QWERTY keyboard (Figure 1B). The faces were presented for a maximum of 5 seconds, or until the participant made a response. There was an ITI of 300ms. All faces in each trial were of different identities, and the position of the target was random.

To measure the perceptual similarity of pairs of expressions, we calculated discrimination thresholds using a psychophysical method. Psychophysical methods have been used in previous research to calculate sensitivity at which emotions can be detected in warped neutral-expressive pairs (Calvo, Avero, Fernández-Martín, & Recio, 2016; Marneweck, Loftus, & Hammond, 2013; Suzuki, Hoshino, & Shigemasu, 2006). Each trial was used in the threshold estimation for one of 15 possible pairs of emotions. The emotion in each pair that was the odd-one-out was selected randomly for each trial (for example, a trial in the Happy-Disgust threshold estimation might contain two examples of Happy and one of Disgust, or two of Disgust and one of Happy). Each of the three pictures in each trial came from continua between the same pair of emotions (but with different identities), but two were from one end of the continua and one was from the other end. The first trial for each pair of emotions started with the 100% versions of stimuli (the end points of the continua). For example, in the first trial of the Happy-Disgust threshold estimation, two faces displayed 100% Disgust (and 0% Happy), and the other face displayed 100% Happy (and 0% Disgust). A staircase procedure (Cornsweet, 1962) was used to adjust the weighting of the expressions in each image by 10% of the remaining distance in each step. For example, faces presented in the next step of the Happy-Disgust staircase contained 90% Happy (and 10% Disgust) and 90% Disgust (and 10% Happy). This meant that the fiducial points of one expression were moved towards the corresponding points of the other expression (by 10% of the distance) for correct responses, and were moved away from the points of the other expression for incorrect responses (Figure 1C).

A 1-up-2-down design was used such that the participants were required to consecutively respond correctly 2 times for the weight to be adjusted downwards. No exaggerations were shown during the task (e.g. if a participant responded incorrectly on the first trial of one threshold estimation, weighting did not increase above 100%). The threshold estimation for each pair of emotions was terminated after 8 reversals in performance, and the discrimination threshold for a given pair of emotions was taken as the mean weighting at the last 7 reversal points. A perceptual matrix for each participant was constructed from the discrimination thresholds for each pair of emotions. Emotion pairs were intermixed and presented in a random order. After all pairs had been presented, the weighting was adjusted, and all (remaining) emotion pairs were presented again in a random order.

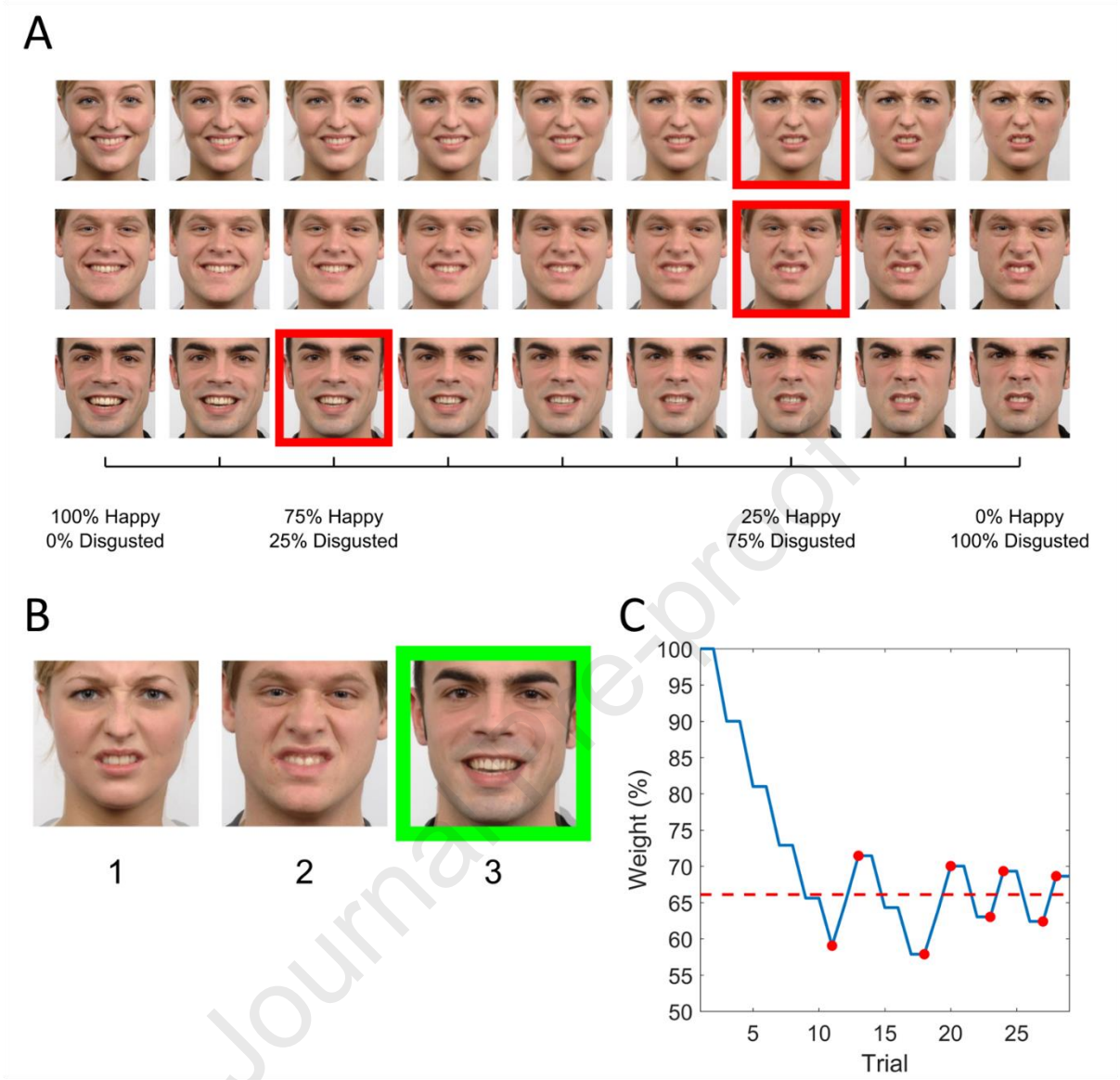

Figure 1: (A) Examples from the Happy-Disgusted continua, for three identities. Faces with red borders show an example of a trial at 75% weight. (B) An example trial at 75% weight. The odd-face-out in this trial is indicated with a green border. (C) An example of the staircase adjustment for the weighting of trials. Weight (as a percentage of the distance between the two prototype faces) is shown on the y-axis, trial number is shown on the x-axis. Red markers indicate reversals in performance. The dashed red line indicates the threshold, calculated as the mean weighting at the last 7 reversals.

Categorical task

Materials

We selected pictures of facial expressions of emotion from the Radboud Faces Database (Langner et al., 2010). Ratings of genuineness (using validation data from Langner et al., 2010) were used to select the top 40 most genuine examples of each emotion. In total, there were 240 images (40 x 6 emotions). 28 of these images were used to create the morphs in the Perceptual Task. Images were cropped to a square containing the whole face. All images were presented in full colour.

Procedure

During each trial participants were presented with a single picture of a facial expression in the centre of the screen (Height/Width = 9.6°, with an average viewing distance of 60cm). To avoid ceiling effects and ensure that the task was sufficiently challenging, each stimulus was presented for 200ms. This ensured that participants did not study the image in too much detail. Brief presentation times have been used in validated facial expression recognition tests to reduce ceiling effects (Ekman & Friesen, 1974; Matsumoto et al., 2000). After the image, participants were presented with 6 emotional labels with numbers 1-6 and were required to label the facial expression using the corresponding number keys on a keyboard. We recorded the number of correct responses to each emotion, as well as the number of instances each emotion was mislabelled with another.

The categorical RDM was constructed from the raw confusion matrix generated by each participant. Although this RDM does not measure 'categorical similarity' of emotions, each subject's pattern of categorisation encodes the similarity of emotions as reflected by the behavioural confusions between a facial expression and an emotion label. Raw confusion matrices have previously been used with RSA (Skerry & Saxe, 2015). As the confusion matrix is not necessarily symmetrical along the diagonal, mirroring cells in the upper and lower triangles were averaged, to produce a categorical matrix for each participant (e.g. the sad-anger cell in the categorical matrix contains the average number of instances a participant mislabelled sad faces as angry, and angry faces as sad).

Model construction

We constructed three models of the types of information that could explain the participants' pattern of categorisations in the Categorical Task and perceptual discrimination sensitivity in the Perceptual Task. A similarity matrix was constructed for each model, where each cell of the matrix represented the discrimination between the corresponding pair of emotions, using a different source of information for each model. Two models showed the similarity between pairs of expressions using stimulus-based cues (Shape and Surface information), whereas another model showed the similarity between the emotion concepts (Conceptual).

Shape

Similarity of face shape was measured by performing Procrustes analysis on the fiducial points of the faces used in the Perceptual Task, as this allowed for the measurement of within-identity changes in face shape. This analysis allows for the comparison of any two shapes, and has previously been used to calculate the similarity of facial expression shapes (Kuhn et al., 2017; Sormaz, Watson, et al., 2016). We used 112 fiducial points corresponding to the x and y coordinates of 112 landmarks in the face. Procrustes analysis computes the average squared distance between each pair of corresponding fiducial points, after correcting for size and position in 2D space by allowing shape translation, rotation, and scaling without morphing or non-linear distortion. The distance measure is then scaled such that the value of the output lies between 0 and 1. This analysis was performed between every within-identity pair of facial expressions as used in the Perceptual Task, to construct a shape dissimilarity matrix for every identity. To construct the Shape Model used in this experiment, we averaged these identity-level matrices. The resulting matrix was subtracted from 1 to keep the direction consistent with the other matrices: as such, higher values indicate greater similarity between the shapes of the corresponding facial expressions.

Surface

To measure surface similarity between the pictures used in the perceptual task, we computed the Fisher's Z-transformed Pearson's correlation coefficient between the pixel intensities for within-identity pairs of greyscale facial expressions. Previous studies have used similar measures of the similarity of surface textures to explain the perceived similarity of facial expressions (Kuhn et al., 2017; Sormaz, Watson, et al., 2016). First, we found the average face shape across all facial expressions, by averaging the locations of the corresponding fiducial points. Then, we warped all face stimuli to this average face shape using Psychomorph (Tiddeman, Stirrat, & Perrett, 2005) to remove shape cues and converted all images to greyscale. We then calculated the correlation coefficient of the pixel intensities, for pixels falling within a mask that excluded all non-face pixels, and transformed them using Fisher's Z-transformation. A matrix was constructed for each identity, then averaged across identities to create the Surface Model.

Conceptual

The conceptual model was computed based on data form Skerry and Saxe (2015), which was generously provided by the authors. In that study, online participants rated 200 short stories describing an emotional event that happened to a character (2-3 sentences each, mean length = 50.68 words) on the extent to which the character was experiencing each of the six basic emotions, on a scale from 1 to 10. In each story, the character experienced one of 20 'fine-grained' emotions (e.g. a story about a character experiencing the emotion loneliness was about them finding it difficult to make friends after moving to a new city). Other examples of these fine-grained emotions included 'disappointment', 'nostalgia', and 'embarrassment', although these labels were never shown to participants. Ten stories were written for each of the 20 'fine-grained' emotions. Full details of stimuli and procedure can be found in Skerry and Saxe (2015). A total of 1556 responses were provided (for this task) by 250 participants, (with each response providing one rating for each of the six emotion dimensions, for one story). From this data, we calculated the mean ratings for the six basic emotions for each of the 20 fine-grained emotion categories (i.e., we found the mean ratings for 'anger', 'disgust', etc. across all stories in which a character was feeling 'nostalgic', 'disappointed' etc.), resulting in six vectors (one vector for each of the six basic emotions) of 20 mean ratings.

To generate Conceptual Model, we calculated the pairwise Pearson correlation coefficients between the vectors for the six basic emotions, then transformed them using Fisher Z-transformation. Pairs of the six basic emotions with higher correlations showed then higher similarity. As such, cells with higher values suggest a greater conceptual overlap between the corresponding emotions than cells with lower values.

Data Analysis

For each task, we first assessed the relationship between the behaviour and each of the three models using Spearman's correlations. The correlation was performed between the behavioural matrices and each of the three models separately, for each participant. The array of coefficients for each model was then tested against 0 using a one-sided Wilcoxon signed rank test. Noise ceilings were computed in a similar manner to Nili et al. (2014), for both the Perceptual and Categorical tasks. The upper bound of the noise ceiling was calculated as the average Spearman's correlation coefficient between each participant's matrix and the average of all participants' matrices (e.g. the average coefficient between each participant's perceptual matrix, and the average of perceptual matrices across all participants), after rank transforming all matrices. The lower bound of the noise ceiling was calculated as the average Spearman's correlation coefficient between each participant's matrix and the average of all *other* participants' matrices (e.g. the average coefficient between each

participant's perceptual matrix and the average perceptual matrix across all *other* participants), after rank transforming all matrices.

To assess the relative contribution of each model in explaining the behavioural tasks, we used multiple linear regression. The three models were entered into a multiple linear regression model (with a constant of ones) to explain each behavioural task, for each participant. The behavioural measures and the models were z-scored to calculate the standardised beta weights, allowing us to examine the relative contribution of each model. The arrays of standardised beta weights were tested against 0 using a one-sample t-test. All analyses were conducted using only the lower triangle (15 off-diagonal cells) of each matrix.

To investigate whether the association of each cue with the behavioural tasks differed between each task, we compared the percentage of variance that was accounted for by each predictor between the two tasks, following analysis from Mur et al. (2013). In that study, the researchers compared the proportion of variance of behavioural and brain RDMs that was accounted for by several object category models. Similarly, in this study, for each regression model for each participant, we estimated the percentage of variance accounted for by each predictor, by calculating the squared standardised beta weight as a percentage of the sum of squared standardised betas for all predictors. Doing so provided us with a normalised measure of the variance that each predictor accounted for. These percentages were then compared between the two tasks across subjects, using a paired samples t-test for each predictor.

Results

Behavioural matrices

For each participant, we constructed a confusion matrix from their choice of labels in the categorisation task, and a perceptual similarity matrix from the pairwise discrimination thresholds in the perceptual task. Figure 2A shows the categorical matrix and the perceptual matrix, averaged across all participants (just for visualisation). Calculation and results of the inter-subject reliability are presented in the supplementary material (Figure S1).

Models

Figure 2B shows the matrices for the three models used in the analysis. We then investigated the correlations between the different models. Pairwise Spearman's correlations showed no significant correlation between Conceptual and Shape ($rho(13) = .043$, $p = .883$), a small non-significant correlation between Conceptual and Surface ($rho(13) = .354$, $p = .196$), and a large and significant correlation between Shape and Surface ($rho(13) = .707$, $p = .004$).

To check whether the estimation of regression coefficients would be substantially affected by any multicollinearity among the models we calculated the Variance Inflation Factor (VIF) for each model. The VIF for a predictor in a multiple linear regression model is an estimation of the factor by which the variance of the coefficient of the predictor has increased due to multicollinearity between the predictors (Thompson, Kim, Aloe, & Becker, 2017), where a large VIF (exceeding a recommended threshold of 5) indicates that the multicollinearity among the predictors may have substantially affected the estimation of the associated regression coefficients (Montgomery, Peck, & Vining, 2012). VIFs have been calculated to assess the degree of multicollinearity in a similar study using multiple linear regression with RSA (Sormaz, Watson, et al., 2016). The VIF for the Conceptual model was 1.22, for the Shape model was 1.97, and for the Surface model was 2.20. None of the VIFs exceed the recommended threshold of 5 (Montgomery et al., 2012), suggesting that the estimated regression coefficients have not been substantially affected by multicollinearity among the predictors.

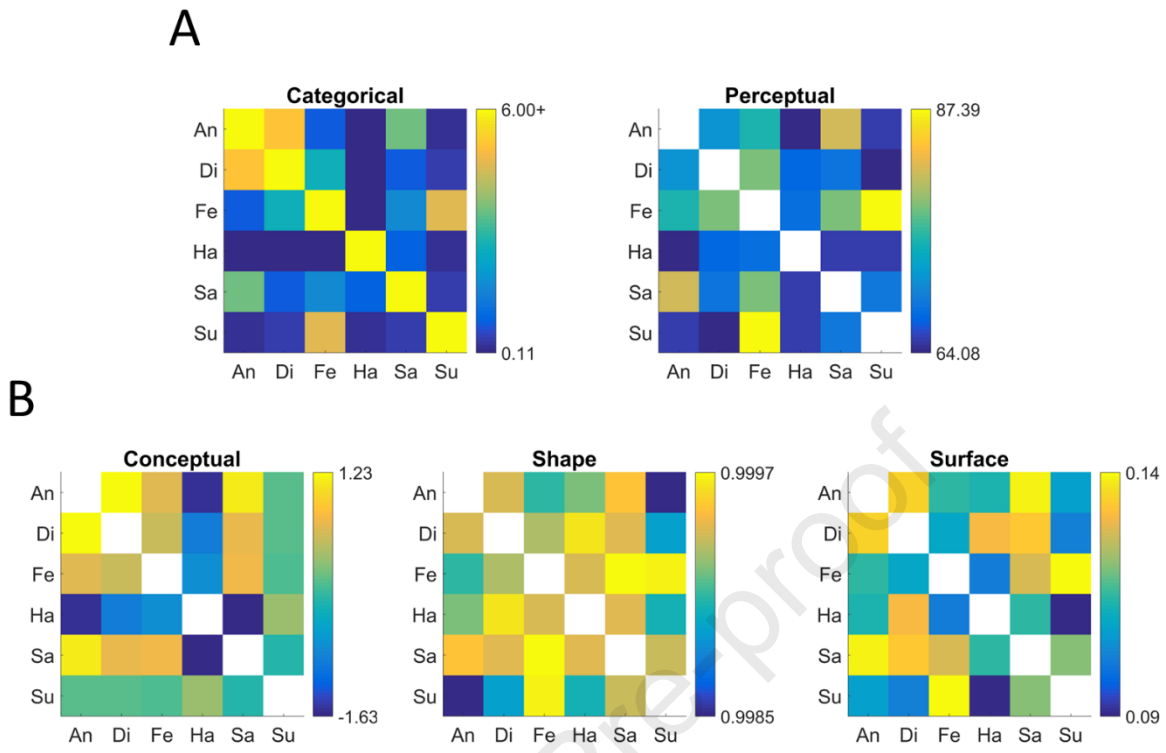

Figure 2: (A) The mean categorical matrix (left), which displays the mean number of instances pairs of emotions were confused with each other. The mean perceptual matrix (right), which displays the perceptual discrimination thresholds for each pair of expressions in the perceptual task. The colour bars are scaled to the minimum and maximum values in the off-diagonal triangles. (B) Conceptual (Fisher Z-transformed correlation coefficients; left), Shape (1-Procrustes distance; centre), and Surface (Fisher Z-transformed correlation coefficients; right) models.

Explaining perceived similarities in the Perceptual Task

To investigate the role of the Conceptual, Shape, and Surface cues on the Perceptual Task, we first computed the Spearman's correlation coefficient between each participant's perceptual matrix, and each of the three models. The correlation coefficients across all participants were significantly higher than zero for all models (Conceptual: mean rho = .361, Z = 1258.5, $p < .001$; Shape: mean rho = .373, Z = 1273, $p < .001$; Surface: mean rho = .430, Z = 1275, $p < .001$), after correcting for multiple comparisons with the Bonferroni adjustment ($\alpha = .0167$). These results show that the perceptual discriminations of facial expressions were significantly correlated with each of the models that we used here.

Figure 3A shows the correlation coefficients, and the estimated noise ceiling. The proximity of the upper and lower bounds of the noise ceiling suggest that all participants were behaving very similarly in the perceptual task. These results also show that, while the correlations were significantly above zero for each model, none of the correlations reached the lower bound of the noise ceiling, so no individual model could explain most of the variance in participants' perceptual dissimilarities. Pairwise comparisons (using Wilcoxon rank-sum tests) revealed no significant differences between the distributions of correlation coefficients (Conceptual-Shape: Z = -0.141, p = 0.888; Shape-Surface: Z = 1.565, p = 0.118; Conceptual-Shape: Z = 1.534, p = 0.125).

We then used multiple linear regression to estimate the relative contribution of each model for the similarities in the Perceptual judgments. The beta weights for each model were significantly higher than zero (Conceptual: $t(49) = 11.72$, $p < .001$; Shape: $t(49) = 8.72$, $p < .001$; Surface: $t(49) = 3.40$, $p = .001$), after correcting for multiple comparisons with the Bonferroni adjustment ($\alpha = .0167$). The mean R^2 value across participants was 0.386 (S.D. = 0.158), so approximately 38.6% of the variance was explained by all three models. Pairwise comparisons between the arrays of standardised regression coefficients revealed greater contribution of conceptual than surface ($t(49) = 2.96$, $p = .005$, $d = 0.74$), no significant difference between the contributions of Shape and Surface ($t(49) = 2.23$, $p = .030$, $d = 0.60$), and no significant difference between the contributions of Conceptual and Shape ($t(49) = 0.67$, $p = .504$, $d = 0.10$), after correcting for multiple comparisons with Bonferroni adjustment ($\alpha = .0167$). Distributions of standardised beta-weights are presented in Figure 3B.

Explaining judgments in the Categorical Task

To investigate the role of the conceptual, shape, and surface cues on the confusions between emotion pairs in the Categorical Task, we first computed the Spearman's correlation coefficient between each participant's categorical matrix, and each of the three models to evaluate their individual contribution. The arrays of correlation coefficients were significantly higher than zero for all three models (Conceptual: mean rho = .508, $Z = 1275$, $p < .001$; Shape: mean rho = .379, $Z = 1274$, $p < .001$; Surface: mean rho = .557, $Z = 1275$, $p < .001$), after correcting for multiple comparisons with the Bonferroni adjustment ($\alpha = .0167$). These results show that judgments in the Categorical Task could be explained by each of the models that we used here. Figure 3C shows the correlation coefficients, and the estimated noise ceiling. These results show that, while the correlations were significantly above zero for each model, none of the correlations reached the noise ceiling, and so none of the models could fully explain the variance in the confusion matrices in the Categorical Task. Pairwise comparisons (using Wilcoxon rank-sum tests) revealed higher correlation coefficients for the Conceptual Model than the Shape Model ($Z = 4.009$, $p < .001$), higher coefficients for the Surface Model than the Shape Model ($Z = 5.636$, $p < .001$, but no difference between the coefficients for the Conceptual and Surface Models ($Z = 1.610$, $p = .107$).

We used multiple linear regression to estimate the unique contribution of each model, in order to predict each participant's Categorical matrix. All arrays of standardised betas were significantly higher than 0 (Conceptual: $t(49) = 18.23$, $p < .001$; Shape: $t(49) = 7.54$, $p < .001$; Surface: $t(49) = 12.18$, $p < .001$), after correcting for multiple comparisons with the Bonferroni adjustment ($\alpha = .0167$). The mean R^2 value across subjects was 0.418 (S.D. = 0.125), so approximately 41.8% of the variance was explained by all three models. Pairwise comparisons between the arrays of standardised regression coefficients revealed a greater contribution of Conceptual than Shape ($t(49) = 10.50$, $p < .001$, $d = 1.66$), greater contribution of Surface than Shape ($t(49) = 4.50$, $p < .001$, $d = 1.17$), but no significant differences between the contributions of Conceptual and Surface cues ($t(49) = 0.838$, $p = .406$, $d = 0.21$) after correcting for multiple comparisons with the Bonferroni adjustment ($\alpha = .0167$). Distributions of standardised beta-weights are presented in Figure 3D.

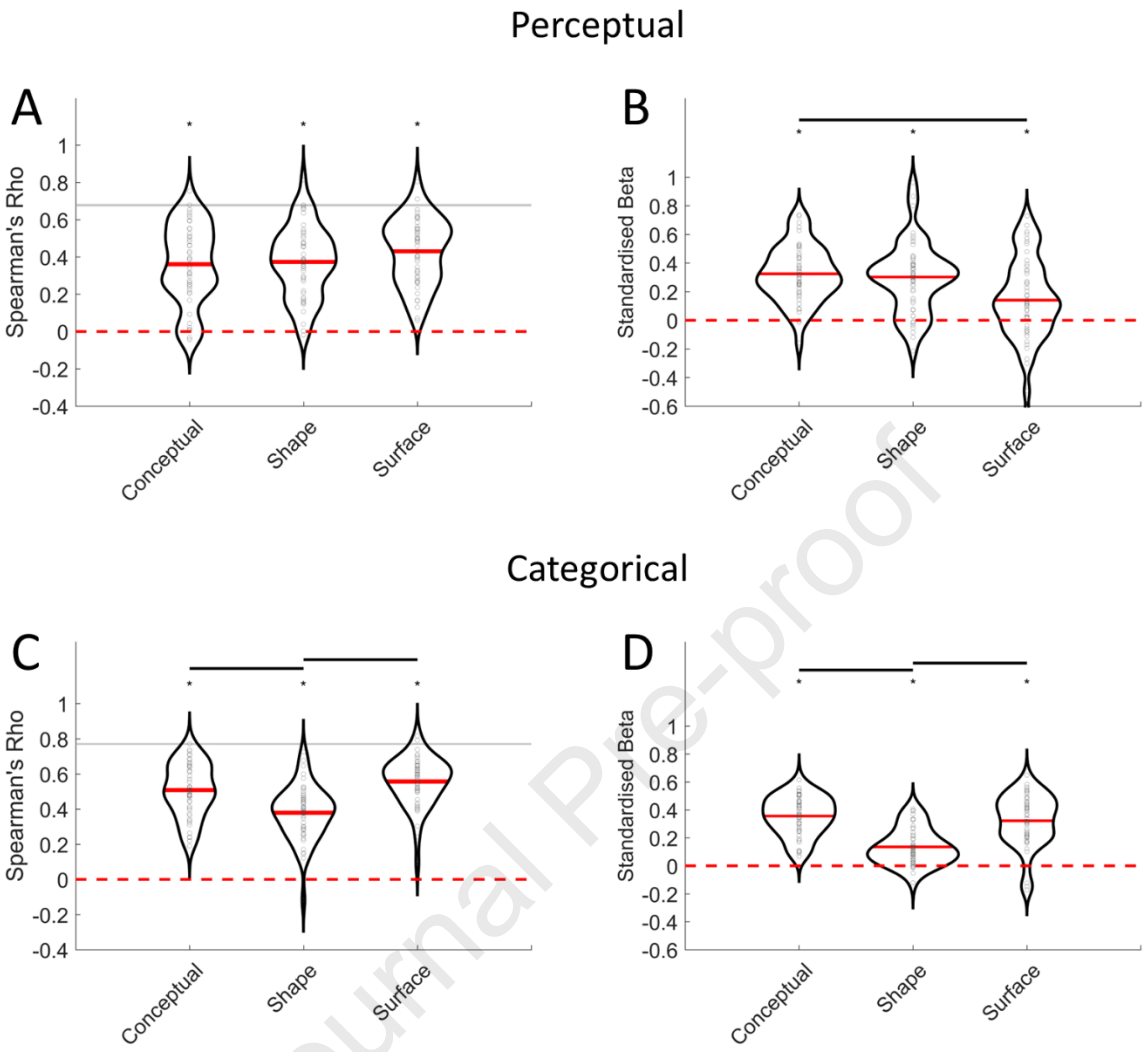

Figure 3: (A) Distributions of Spearman's correlation coefficients between each participant's perceptual matrix and each of the models. Red bars show the mean coefficient. Grey bar shows the upper and lower bounds of the noise ceiling. (B) Distributions of standardised beta weights for each model as a predictor of each participant's perceptual matrix. Red bars show the mean standardised beta. (C) Correlation coefficients between each participant's categorical matrix and each of the models. Red bars show the mean coefficient. Grey bar shows the upper and lower bounds of the noise ceiling. (D) Standardised beta weights for each model as a predictor of each participant's categorical matrix. Red bars show the mean standardised beta. Horizontal black bars indicate a significant difference for the pairwise comparison, and asterisks indicate the array is significantly larger than 0, accounting for an FDR of .05.

Differences between categorical and perceptual regression coefficients

After correcting for multiple comparisons with the Bonferroni adjustment (α = 0.0167), there was no difference between the percentage of variance that the Conceptual Model accounted for between the Perceptual (Mean = 36.1%) and Categorical (Mean = 45.0%) tasks ($t(49) = 1.730$, $p = .090$). The Shape Model accounted for more variance within the Perceptual (Mean = 37.5%) task than the Categorical (Mean = 10.6%) task ($t(49) = 6.710$, $p < .001$), whereas the Surface Model accounted for more variance within the Categorical (Mean = 44.4%) task than the Perceptual (Mean = 26.5%) task

($t(49) = 2.975$, $p = .005$). Percentages of variance accounted for are presented in the supplementary material (Figure S2).

Discussion

In Study 1, we aimed to explore the relative roles of conceptual, shape, and surface cues in the perceptual discrimination and explicit categorisation of facial expressions of emotion. First, as expected, we found that all three cues play some role in explaining behaviour during the two tasks. This is consistent with previous research showing that these cues might influence facial expression perception and recognition (Brooks & Freeman, 2018; Sormaz, Watson, et al., 2016; Sormaz, Young, & Andrews, 2016), but we furthered the research by using multiple linear regression to show that each cue plays a role even when controlling for the other cues.

A second main finding is that we found that conceptual information explains a similar amount of variance in both tasks. In fact, the current results suggest that emotion concepts seem to be as readily available when performing a perceptual discrimination task that required no explicit labelling of emotions as they are during a categorisation task that does require explicit labelling. These findings are not consistent with our predictions that conceptual information would explain the behaviour in the categorical task better than the perceptual. Previous research suggests that conceptual information influences both the perception and explicit labelling of emotions from facial expressions (Brooks & Freeman, 2018; Fernández-Dols et al., 2008; Gendron et al., 2012; Halberstadt & Niedenthal, 2001; Lindquist et al., 2006; Nook et al., 2015). However, the perceptual tasks in previous studies tended to still use verbal labels that we thought would be more likely to engage conceptual processing. We thus designed a measure of the perceptual similarity of facial expressions that attempted to reduce top-down influences during the task and tap 'solely' into perceptual discrimination processes. Despite this, we still found that conceptual information could explain perceptual discrimination well above chance, even when controlling for the similarity of stimulus-based cues, supporting the conclusions of previous research (Brooks & Freeman, 2018). The result that conceptual information can explain patterns of perceptual discrimination better than surface information is surprising, as we expected that processes involved in perceptual discrimination would rely more on stimulus-based cues than emotion concepts.

Finally, we found a dissociation between the role of shape and surface cues, in that shape cues explained more variance of the perceived similarities in the Perceptual Task than the confusions in the Categorical Task, and surface cues could explain more variance in the Categorical than Perceptual Task. Taking the view that perceptual processes occur before processes involved in the labelling of emotions (Palermo et al., 2013), this suggests that initial perceptual processes place greater weight on shape information, while surface information is required more so in the explicit categorisation of emotions.

Study 2

Methods

Participants

Thirty participants (19 women, mean age: 26.10 years, S.D: 6.45, range: 18-46) were recruited via posters, advertisements on the university participant pool, and word of mouth. Sample size was determined based on previous studies with similar participant sizes, although no formal power analyses were conducted. All participants had normal or corrected-to-normal vision, no history of stroke, other neurological conditions, or diagnosed emotion processing disorders. Seven participants

for Study 2 also participated in Study 1. Ethical approval to conduct the research was granted by the College of Health and Life Sciences Research Ethics Committee at Brunel University London.

Multiple choice task

Materials

Pictures of facial expressions of anger, disgust, fear, happiness, sadness and surprise were selected from the Radboud face database (Langner et al., 2010). We used the validation data from Langner et al. (2010) to rank the front facing images for each emotion by genuineness of expression, then selected the top five pictures for each emotion. Every image appeared in the Categorical Task (Study 1) and four of the images were used to create the morphs in the Perceptual Task (Study 1). Images were cropped to a square containing the whole face but removing the neck and the top of the head. All images were presented in full colour.

Procedure

Before participants entered the scanner, they completed a brief multiple-choice style expression categorisation test. Participants viewed each face for 5000ms in the centre of a laptop screen, with 6 labels (state the labels) presented underneath. Each label had a corresponding number (1-6), and participants were required to indicate which of the labels best described the emotion in the presented face. The purpose of this task was to encourage participants to discriminate between the faces in terms of the 6 basic emotions.

MRI data acquisition

MRI data was collected with a 3T TIM Trio MRI scanner (Siemens, Erlangen) using a 32-channel head array coil at the Combined Universities Brain Imaging Centre (CUBIC). Functional images were acquired for the experimental and localiser runs with an echo planar imaging sequence with a multiband acceleration factor of 2, with 46 axial slices aligned with the ventral surface of the temporal and occipital lobes, covering the whole brain excluding the cerebellum in most participants (TR = 2000ms, TE = 34ms, Flip angle = 76°, voxel size = 2.5mm x 2.5mm x 2.5mm). For participants with larger brains, the upper most part of the parietal lobe was excluded. A high-resolution T1-weighted MPRAGE anatomical scan was also acquired for each participant (TR = 1830ms, TE = 3.03ms, Flip angle = 11°, Voxel size = 1mm x 1mm x 1mm). Runs were divided into experimental runs and functional localiser runs. Participants completed ten experimental runs, two theory of mind area localiser runs, and one face area localiser run. Experimental runs were stopped after eight and nine runs for two participants due to discomfort, and one participant did not complete the theory of mind localiser. The data provided by these participants were still used in the analysis.

Experimental fMRI task

Materials

We selected the top 11 most genuine examples of each of the six basic emotions (Anger, Disgust, Fear, Happiness, Sadness, and Surprise) from the Radboud face database, using the validation data from Langner et al. (2010). Participants had previously viewed five examples of each expression category in the multiple-choice task. Every image appeared in the Categorical Task (Study 1) and eight of the images were used to create the morphs in the Perceptual Task (Study 1). During each run, participants viewed all 66 images. All images were presented in full colour.

Procedure

An event-related design was used for the experimental runs. The 66 pictures of facial expressions were presented once each in each experimental run. Pictures were presented sequentially for 1000ms each (height = 12.5°, width = 16.7°, at a viewing distance of 80cm). After each presentation, a blank fixation screen was presented at a jittered duration of between 1000ms and 3000ms. Participants performed a one-back task, by pressing a button whenever there was a consecutive repetition of any facial expression – the repetition was always of a different identity (i.e. it was a different picture), encouraging participants to attend to the expression rather than just the image.

During each run, the 66 faces were presented in a pseudo-random order: initially the order of 60 (ten examples of six expressions) was randomised until there were no consecutive repetitions of the same facial expression. Each of the six remaining images of facial expressions were inserted into this sequence after the position of a random example of the same expression, to provide a repetition of each expression for the one-back task. This randomisation procedure was conducted independently for each run and each participant. A blank fixation screen was presented for the first four seconds of each run.

Functional localisers

To localise the face responsive regions, participants viewed 16 second blocks of expressive faces, neutral faces, and scrambled faces from the Radboud face database (Langner et al., 2010), with additional 16 second blocks of rest (no stimuli). Blocks of expressive faces comprised a randomly selected mix of angry and fearful faces from the front facing angry and fearful faces of the database. Each face was also scrambled using MATLAB, by overlaying the image with a 20x20 grid and randomly rearranging the position of the 'tiles'. During each block, 16 faces (or scrambled faces) were selected at random. In each block, each stimulus was presented for 900ms (ISI = 100ms). Blocks of stimuli were presented in a pseudo-random order. Each block occurred four times with no separation between them, and there were never consecutive repetitions of the same block.

To localise the MPFC, we used stimuli from Dodell-Feder, Koster-Hale, Bedny and Saxe (2011) Participants were presented with short textual scenarios that require inferences about either the mental state of an individual (belief condition) or physical state of a scene (photo condition). Full stimuli are available at http://saxelab.mit.edu/superloc.php. Participants completed a true/false task in response to a statement presented beneath the text, by pressing a button with their left hand for true and right hand for false. The text and true/false statement appeared on screen for 18 seconds, followed by a blank screen for 12 seconds. Please note that, due to an error, we used timings that differed from those used in the original paper (which presented the text for 10 seconds, followed by the true/false statement for 4 seconds, and 12 seconds of rest).

Image preprocessing and general linear model

For each participant, all functional images were realigned (registered to the mean of the whole session using 2nd degree B-spline interpolation) and resliced (using 4th degree B-spline interpolation). Each participant's structural image was segmented and co-registered to their mean functional image, then functional images were normalised to MNI space (voxel size = 2mm x 2mm x 2mm) using the deformation field output from warping the structural to the MNI template. Functional images for the localiser tasks were smoothed with a gaussian kernel at FWHM = 8mm. All preprocessing was performed in the SPM12 toolbox in MATLAB 2016b. Runs were excluded if the participant moved more than 2.5mm in any direction, or rotated more than 1 degree along any axis.

To estimate betas for each event of the experimental runs (i.e. each of the 66 pictures of facial expressions), we used the LS-S approach (Mumford, Turner, Ashby, & Poldrack, 2012), where each

event was fitted with a separate linear model containing a regressor for the event of interest and a separate regressor that models all other events, along with 6 regressors for the realignment parameters within each run. This approach aims to lead to more accurate estimates of activation than modelling each trial with a separate regressor within the same model (Mumford et al., 2012), and has previously been used to estimate patterns of activation in response to facial expressions (Wegrzyn et al., 2015). The GLMs were convolved with a standard haemodynamic response function and high-pass filtered at 128s. Due to technical issues during scanning, the scanner stopped before all stimuli had been presented for some participants (a total of 6 stimuli across all runs across all participants were presented after the scanner stopped). Betas for any event presented within 3 seconds of the scanner finishing were not estimated (18 stimuli across all participants).

Region of interest definition

To define the face responsive regions and the MPFC, a group-constrained subject-specific approach was used (Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010; Julian, Fedorenko, Webster, & Kanwisher, 2012). This method allows us to define these regions without experimenter bias. Briefly, we used group masks for each region of interest from previous studies (see below) and then intersected each group mask with each participant's activation map to define individual ROIs.

Group level maps from previous research were used as masks for the middle-MPFC (Dufour et al., 2013), and three core face-responsive regions, namely the FFA, OFA, and STS (Julian et al., 2012). For each participant, we performed standard univariate analysis for each of the two localiser tasks. To localise face responsive regions, we subtracted the response to scrambled faces from the response to both neutral and expressive faces. Contrasting faces to scrambled faces has been used to identify face-responsive regions in previous research investigating the representation of expression and identity (Sormaz, Watson, et al., 2016; Weibert, Flack, Young, & Andrews, 2018; Winston, Henson, Fine-Goulden, & Dolan, 2004). To localise theory of mind responsive regions, we subtracted the response to the photo conditions from the response to the belief conditions in the Theory of Mind localiser task. Using a liberal threshold of $p<.05$ (uncorrected), we found the peak voxel for each participant within each mask and used a sphere with a 6mm radius, centred on these coordinates, as the ROI for each participant. If no peak voxels were significant below $p<.05$, we used the peak coordinates from second level random effects analysis as the centre the sphere. These coordinates were also used for the one participant who did not complete the theory of mind task. The second-level analysis was conducted at the group-level, using a one-sample t-test for each contrast at $p<.001$ (uncorrected), within the same group level maps from previous research (Dufour et al., 2013; Julian et al., 2012) as used in the subject-specific approach.

Model construction

We used the three models we used previously (Shape, Surface and Conceptual) and two models derived from the behavioural tasks (Perceptual and Categorical) from Study 1 for the analysis in Study 2. The behavioural models (Perceptual and Categorical) were created by averaging participants' individual matrices for each task (again mirroring averaging cells in the confusion matrices in the Categorical Task). The Shape and Surface Models were reconstructed using the stimuli presented to participants in the scanner. Following Sormaz, Watson et al. (2016), the similarity of every pair of faces was calculated and averaged for each expression pair to create a single 6x6 matrix for each model, but otherwise using the same procedures as in Study 1. The Conceptual Model remained the same as in Study 1. Pairwise Spearman's correlations showed no significant correlations between Conceptual and Shape ($rho(13) = .311$, $p = .259$), no significant

correlation between Conceptual and Surface (rho(13) = .311, p = .259), and no significant correlation between Shape and Surface (rho(13) = -.068, p = .812).

Univariate analysis (experimental task)

For exploratory purposes, we assessed any differences between the univariate responses to each of the facial expression categories within each of the ROIs. Estimated betas for each emotion were averaged across all voxels within each ROI, then averaged across runs. To assess any differential response between them, paired t-tests were performed between every pair of betas across all participants, using FDR correction to account for multiple comparisons (Benjamini & Hochberg, 1995).

Representational similarity analysis

After betas were estimated for each event in each run, they were averaged for each emotion (creating six averaged response patterns per run), then these averaged response patterns within each region were vectorised and each pattern was z-scored (i.e. the pattern of activation for each condition across voxels was z-scored to mean of zero and standard deviation of one — see Goesaert & Op de Beeck, 2013). A 6x6 Representational Dissimilarity Matrix was constructed by taking the squared Euclidean distances between each z-scored vector, for each region of interest. We chose this distance measure (squared Euclidean) as we used multiple linear regression with RSA, which requires a distance measure that sums linearly (Brooks et al., 2019; Carlin & Kriegeskorte, 2017).

Rows and columns of the RDMs were sorted by expression consistently across runs (Angry, Disgusted, Fearful, Happy, Sad, and Surprised). For each participant, an average RDM was computed by averaging the RDMs from all runs. Models were reversed-coded (by subtracting each element from the maximum element in the array + 1) to keep the direction and size consistent with the neural RDMs. The relationship between each of the models and the neural RDMs was first examined using a Spearman's correlation, for each participant. The array of correlation coefficients was then tested against 0 (i.e. no correlation) using a one-sided Wilcoxon signed rank test.

In addition, we ran multiple regression to examine the unique predictor value of each model and whether a combination of models could better explain the brain representational distances. For each participant, multiple linear regression was performed using the Conceptual, Shape, and Surface Models as predictors and the brain RDMs (separately for each region of interest) as the outcome (Figure 7A). All predictors and the neural RDM were z-scored to calculate the standardised betas for each model. A constant of ones was also entered into each regression model. The distributions of standardised beta weights across participants were tested against 0 using a one-sample t-test, for each model, for each region of interest. To assess the relationship between the neural RDMs and perceptual and categorical models, these were each entered as predictors in separate regression models. As with the previous regressions, the predictor and the neural RDM were z-scored and entered into the model with a constant of ones, and the array of standardised betas were tested against 0 using a one sample t-test.

Preregistration

This study was pre-registered before any data collection. Methods and analysis plan were pre-registered on 08/11/18 and are available on the Open Science Framework (https://osf.io/34fm7/). Any changes to the study procedures analysis plan are transparently identified and the outcomes of pre-registered and post hoc analyses are distinguished. We note that we had originally planned to calculate representational distances between the responses to every pair of stimuli (resulting in a

66x66 RDM) rather than between the averaged responses to each emotion category. However, preliminary analysis showed that this substantially reduced the reliability of RDMs across participants. Furthermore, we had intended to use different contrasts to localise the face areas, by contrasting neutral faces to scrambled to localise the FFA and OFA, and contrasting expressive faces to neutral to localise the STS. We have instead chosen to contrast all faces to scrambled faces for all regions, as this identifies voxels that are responsive to both neutral and expressive faces. Additionally, we had not planned to use the group-constrained subject-specific approach to define the regions of interest, but rather the results of univariate analysis for each participant. This was changed as some participants did not show significant differential activation to the contrasts used in the univariate analysis of the localiser tasks, and using spheres centred around peak voxels allowed for a consistent number of voxels in the ROIs across all participants.

Results

Data exclusion

One participant moved more than our threshold (2.5mm in any direction) during the face localiser and 5 of the experimental runs, so we excluded this participant from any further analysis. One experimental run was also excluded from each of three additional participants due to movement, but we analysed the other runs for these participants. Finally, one participant was found to have a large structural abnormality, so was excluded from any analysis, leaving 28 participants for the main analysis (with 10 experimental runs for 23 participants, 9 runs for 4 participants, and 8 runs for 1 participant, including those who only completed 8 and 9 runs of the experimental task). These exclusion criteria were established after the univariate analysis, but prior to the RSA.

Multiple choice task

We calculated the mean and standard deviation for the accuracy at correctly recognising each emotion during the behavioural multiple-choice task conducted outside the scanner. Participants were most accurate at the recognition of happy faces (with an average accuracy of 98%) and were least accurate at the recognition of angry and fearful faces (with recognition accuracies of 61.33% and 66% respectively). Results of this task are presented in the supplementary material (Table S1). As this task was conducted to familiarise participants with the six basic expressions, these results were not used in any further analysis.

ROI definition

Using the approach described in the methods section, we used the localiser tasks to define the regions of interest (FFA, OFA, STS, and MPFC) for each participant. Table 1 shows the number of participants in which we found at least one voxel showing significant differential activation, and the mean coordinates in MNI space across these participants, for each region. For these participants, we placed a sphere (of 6mm radius) around the peak voxel for each contrast and extracted response patterns for each condition from these spheres. For the remaining participants in each region, the sphere was centred around the peak coordinates from second level analysis conducted across all participants (MNI coordinates [x, y, z]: FFA = [44, -52, -18]; OFA = [36, -82, -10]; STS = [52, -54, 6]; MPFC = [10, 48, 14]). We note that the contrast and method used to define these regions was different to the plan outlined in the preregistration (see preregistration section for more detail).

Table 1: Results from the group-constrained subject-specific approach to region of interest definition. Number of subjects in which at least one-voxel showing significant differential activation within each mask, and the average x, y, and z coordinates across these subjects are shown.

Region	**Number of participants (/28)**	**Mean coordinates**		
		x	*y*	*z*
FFA	28	40.7	-49.1	-18.7
OFA	28	43.1	-76.1	-9.9
STS	25	52.3	-46.4	5.58
MPFC	27	3.6	56.4	14.2

Univariate analysis (Experimental task)

To assess whether there was any differential univariate response to the facial expressions within each of the ROIs, the betas in response to each emotion were averaged across all voxels within each ROI, and then averaged across runs for each participant. Paired *t*-tests were then performed between every possible pairwise combination of these mean betas. To account for multiple comparisons, we controlled false discovery rate (FDR) at level .05 (Benjamini & Hochberg, 1995). Results are presented in Figure 4. A similar pattern of results was observed across the three face-responsive regions, where fearful, surprised, angry, and disgusted faces evoked the largest responses. Happy faces evoked the smallest response. There were also differences between expressions in the MPFC, but with a different pattern of responses.

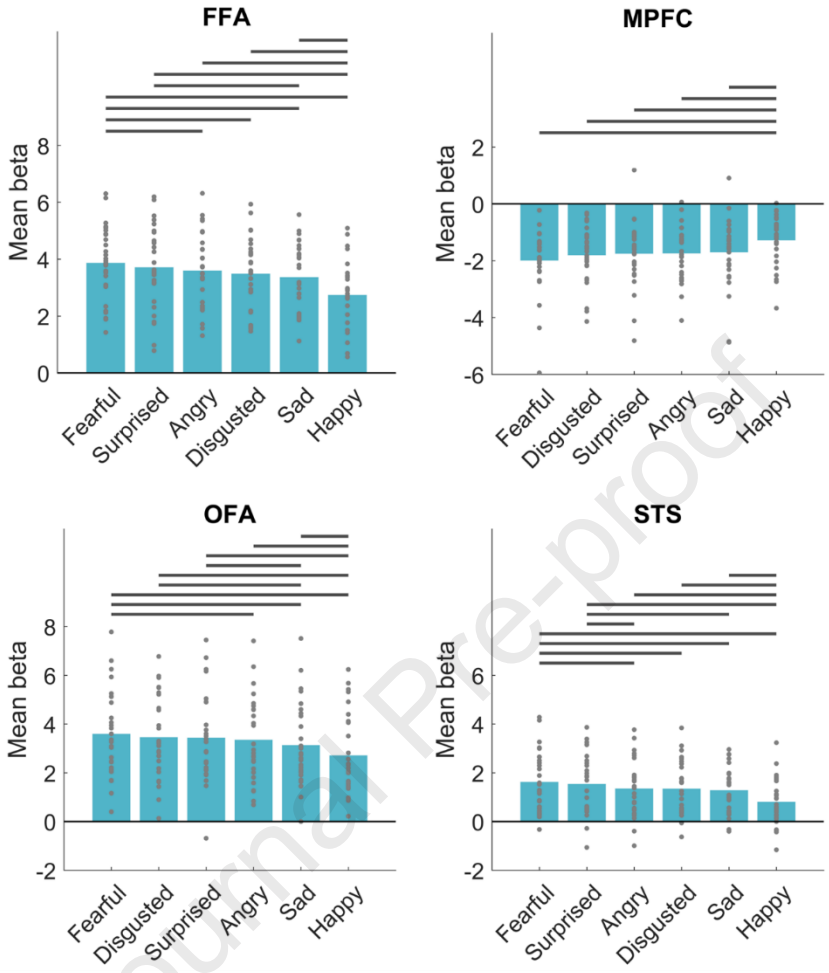

Figure 4: Univariate response to each expression category within each ROI. Horizontal black bars indicate a significant p-value for the pairwise comparison, accounting for an FDR of .05 (FFA all p < .010; OFA p < .022; STS p < .026; MPFC p < .013)

Representational similarity analysis

We first explored the relationship between each model and the brain representational geometries within each region. To do so, we calculated the Spearman's correlation coefficient between the lower diagonal cells of each model RDM and the corresponding cells in the brain RDM for each region and each participant. Brain RDMs are presented in Figure 5 (averaged across participants for the purpose of visualisation, although individual RDMs were used in the analysis). We note that we had originally planned to calculate the representational distances between every pair of stimuli rather than every emotion category (see preregistration section for more detail).

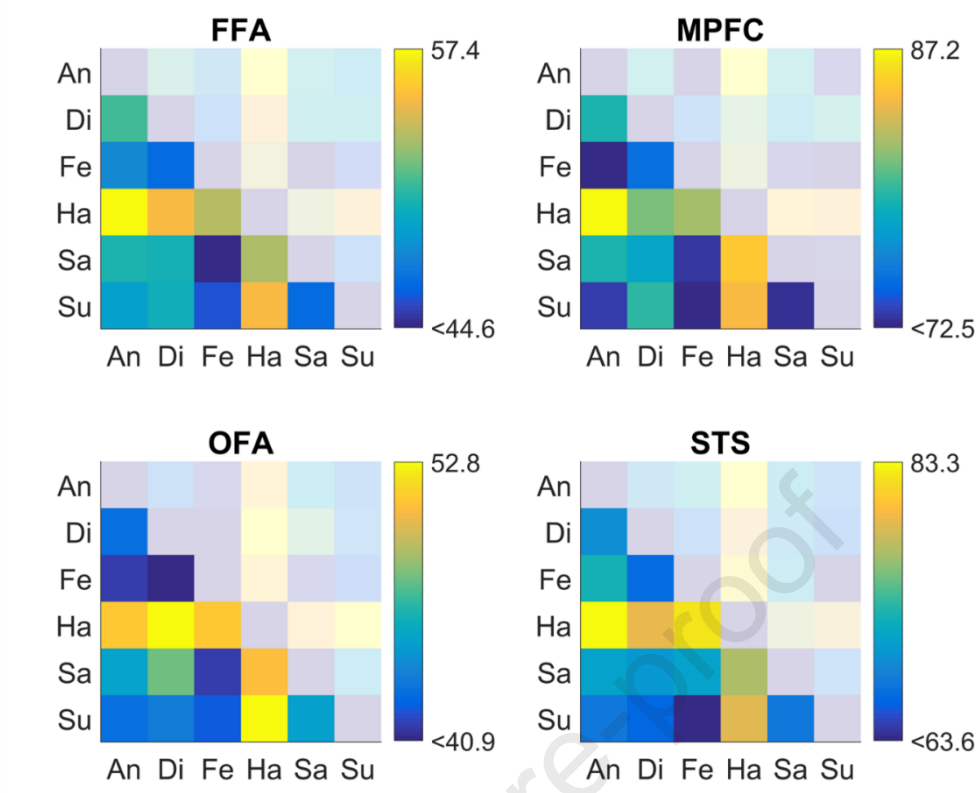

Figure 5: Brain RDMs for each ROI (averaged across participants for visualisation purposes). Cells of each matrix used in the analyses are highlighted. Colourbar values represent squared Euclidean distance.

The array of correlation coefficients (i.e. coefficients across all participants) was then tested against 0 using a one-sided Wilcoxon signed rank test. FDR correction was used to account for multiple comparisons within each ROI. The mean correlation coefficients, standard deviations, associated z-statistic from the Wilcoxon signed rank test, and p-values are reported in Table 2.

Table 2: The mean and standard deviation of the correlation coefficients, and the results of the Wilcoxon signed tank test, for each model, for each region.

ROI	Model	Mean Rho	S.D.	Z-stat	P
FFA	Conceptual	0.232	0.336	3.074	**2.11E-03**
	Shape	0.158	0.275	2.642	**8.25E-03**
	Surface	-0.018	0.258	-0.376	7.071E-01
	Categorical	0.282	0.254	4.111	**3.95E-05**
	Perceptual	0.312	0.262	4.099	**4.14E-05**
OFA	Conceptual	0.284	0.315	3.587	**3.35E-04**
	Shape	0.117	0.275	2.027	4.27E-02
	Surface	0.077	0.258	1.670	9.50E-02
	Categorical	0.279	0.271	3.928	**8.55E-05**
	Perceptual	0.275	0.252	4.076	**4.58E-05**
STS	Conceptual	0.249	0.273	3.530	**4.16E-04**
	Shape	0.015	0.287	0.108	9.14E-01
	Surface	0.054	0.342	0.911	3.62E-01
	Categorical	0.318	0.243	4.258	**2.06E-05**
	Perceptual	0.255	0.270	3.712	**2.06E-04**
MPFC	Conceptual	0.206	0.349	2.699	**6.97E-03**
	Shape	0.118	0.219	2.631	**8.51E-03**
	Surface	0.049	0.325	0.820	4.12E-01
	Categorical	0.206	0.302	2.915	**3.56E-03**
	Perceptual	0.250	0.266	3.632	**2.81E-04**

Conceptual, Categorical, and Perceptual similarities were related to the representational distances in all four ROIs. The structure of the Shape Model was only associated with the representational structure in the FFA and MPFC, whereas the Surface model was not associated with the representational structure in any region. Figure 6 shows the correlation coefficients, and the estimated noise ceiling, which was calculated following Nili et al. (2014). The upper bound of the noise ceiling for each region shows the average Spearman's correlation coefficient between each participants' RDM (for that region) and the average of all participants' RDMs (for the same region), after rank transforming all matrices. The lower bound of the noise ceiling is the average Spearman's correlation coefficient between each participant's RDM and the average of all *other* participants' RDMs, after rank transforming all matrices. The lower-bound of the noise-ceilings ranged from 0.243 in the MPFC to 0.413 in the STS. The only model to reach the noise ceiling in any region was the Perceptual Model in the MPFC, suggesting this model performs as well as any model can, given the noise in the data (Nili et al., 2014).

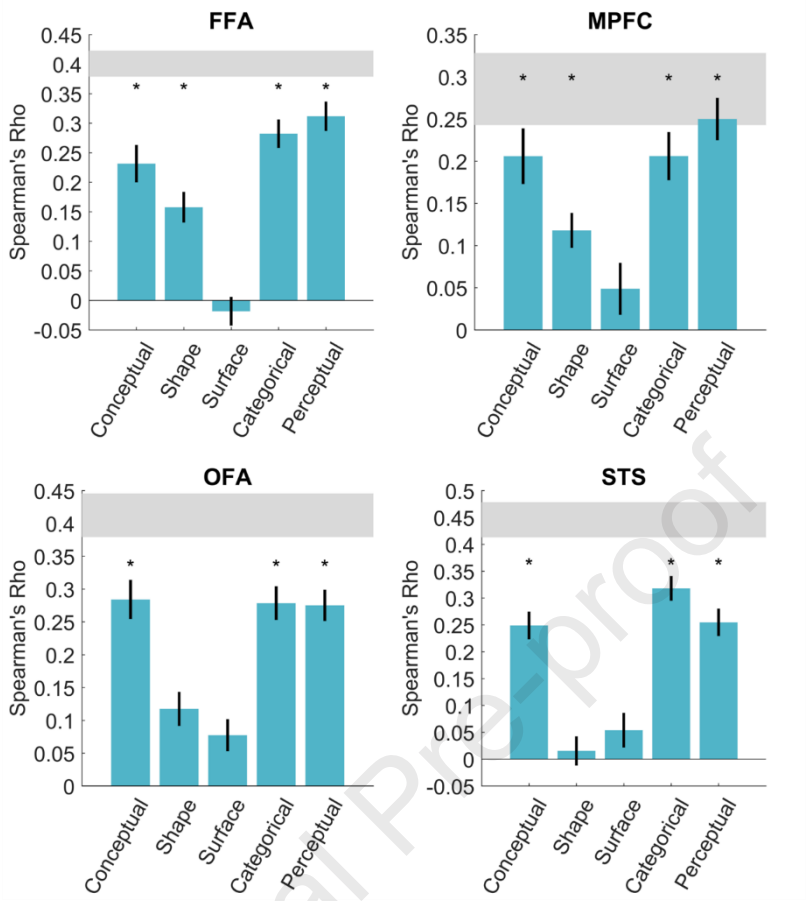

Figure 6: The mean correlation coefficient between each model and the RDM for each region, across participants. Error bars represent one standard error. Asterisks show that the array of coefficients is significantly larger than 0. Grey bars show the upper and lower bounds of the noise ceiling.

Multiple linear regression representational similarity analysis

To assess the relative contribution of the three models (Conceptual, Shape, and Surface) to the representational structure of emotions in each of the four regions, we used multiple linear regression with each of the models as predictors and each subject's RDM as the dependent variable. We first checked whether the estimation of the regression coefficients had been affected by multicollinearity by again calculating the VIFs. The VIF for the Conceptual model was 1.14, for the Shape model was 1.05, and for the Surface model was 1.09. None of the VIFs exceed the recommended threshold of 5 (Montgomery et al., 2012), suggesting that any multicollinearity among the predictors had not substantially affected the estimation of the coefficients. Models and RDMs were z-scored to calculate the standardised beta value for each predictor. Standardised betas were then compared against 0 using a one-sample t-test, and compared to each other using paired sample t-tests. FDR correction was used to account for multiple comparisons separately for each ROI, accepting an average FDR of .05. The mean standardised parameter estimates and standard error for each predictor are presented in Figure 7B. Results of the one-sample T-tests are presented in Table 3, and results of the pairwise comparisons are presented in Table 4. Conceptual similarity significantly explained some of the variance of the representational structures in all four regions after controlling for the similarities of face shapes and surface textures.

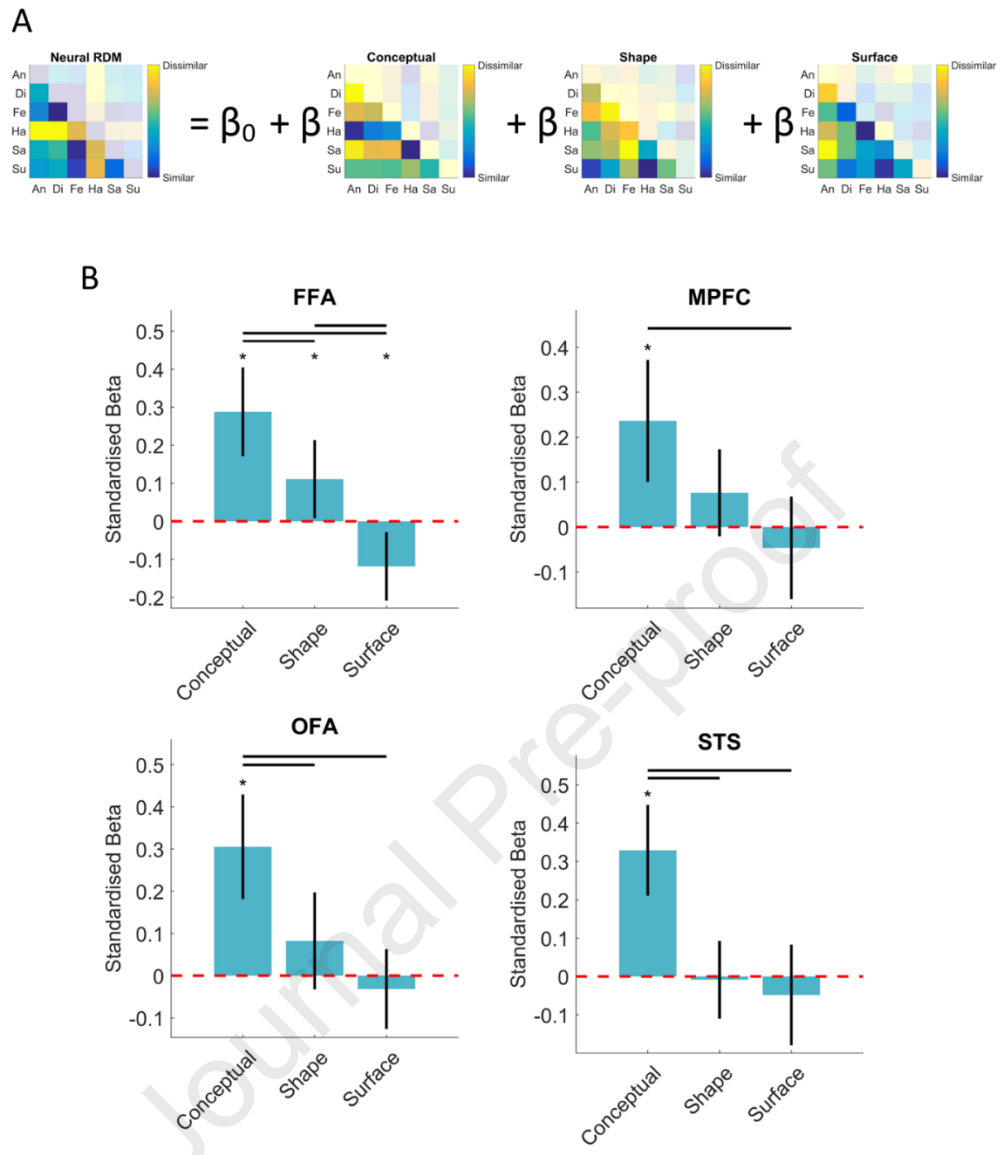

Figure 7: (A) Illustration of the regression model performed per participant, to assess the relative contribution of conceptual, shape, and surface information. (B) The mean standardised betas for each of the 3 models, for each region. Error bars show standard error. Asterisks indicate that the regression coefficients are significantly larger than 0, and horizontal bars indicate significant differences between the regression coefficients.

Interestingly, after controlling for conceptual similarities, shape and surface similarities did not share unique variance with representational similarity in the MPFC, STS, or OFA. In the FFA, these models explained some additional variance after controlling for conceptual similarity, although the variance explained by conceptual similarity was significantly larger across participants than the variance explained by the shape and surface models. The three models together explained most variance in the STS (mean R^2 = 0.354), followed by the OFA (mean R^2 = 0.343), MPFC (mean R^2 = 0.322), then the FFA (mean R^2 = 0.316).

Table 3: Mean, standard deviation, and the results of the one-sample t-test for the conceptual, shape, and surface models as predictors of the representational structure in each of the 4 regions. Significant t-tests are presented in bold.

Region	Model	Mean	S.D.	Mean R^2	T-test
FFA	Conceptual	0.288	0.300		t(27) = 5.071, p < .001
	Shape	0.111	0.265		**t(27) = 2.214, p = .035**
	Surface	-0.118	0.232	0.316	**t(27) = -2.708, p = .012**
OFA	Conceptual	0.305	0.319		**t(27) = 5.062, p < .001**
	Shape	0.082	0.296		t(27) = 1.471, p = .153
	Surface	-0.032	0.244	0.343	t(27) = -0.690, p = .496
STS	Conceptual	0.329	0.305		t(27) = 5.712, p = .001
	Shape	-0.009	0.261		t(27) = -0.175, p = .863
	Surface	-0.048	0.338	0.354	t(27) = -0.756, p = .456
MPFC	Conceptual	0.236	0.350		**t(27) = 3.570, p = .001**
	Shape	0.076	0.250		t(27) = 1.610, p = .119
	Surface	-0.046	0.294	0.322	t(27) = -0.834, p = .411

Table 4: Results of the pairwise comparisons between the standardised betas for each predictor, for each region. Significant t-tests are presented in bold.

Region	Model comparison	T-test
FFA	Conceptual-Shape	**t(27) = 2.338, p = .027**
	Conceptual-Surface	**t(27) = 6.342, p < .001**
	Shape-Surface	**t(27) = 3.615, p = .001**
OFA	Conceptual-Shape	**t(27) = 2.649, p = .013**
	Conceptual-Surface	**t(27) = 4.475, p < .001**
	Shape-Surface	t(27) = 1.479, p = .151
STS	Conceptual-Shape	**t(27) = 4.707, p < .001**
	Conceptual-Surface	**t(27) = 3.966, p < .001**
	Shape-Surface	t(27) = 0.442, p = .662
MPFC	Conceptual-Shape	t(27) = 1.690, p = .103
	Conceptual-Surface	**t(27) = 3.557, p = .001**
	Shape-Surface	t(27) = 1.724, p = .096

Perceptual and categorical models

We then assessed whether the representational structure within each region is best associated with the Categorical or Perceptual Tasks (derived from the results of Study 1). For this analysis, we conducted a separate regression for each of the two tasks, using the model as a predictor of the neural RDM for each ROI. As with the previous regressions, the neural RDMs and predictors were z-scored, and the array of resulting standardised beta weights were compared against 0 using a one-sample t-test. Both models were able to predict the representational structure in all regions above chance. The mean standardised beta, mean R squared value, and the results of the one-sample t-test for the categorical and perceptual models are presented in Table 5 below.

Table 5: The mean standardised beta across participants, the standard deviation, and the results of the t-tests (one sample t-tests comparing parameter estimates against zero) for the perceptual and categorical models as predictors of the representational structure in each region. Significant t-tests are presented in bold.

Region	Predictor	Mean standardised beta	S.D.	Mean R^2	T-test
FFA	Perceptual	0.310	0.242	0.153	**t(27) = 6.780, p < .001**
	Categorical	0.212	0.271	0.116	**t(27) = 4.144, p < .001**
OFA	Perceptual	0.286	0.235	0.135	**t(27) = 6.442, p < .001**
	Categorical	0.243	0.252	0.120	**t(27) = 5.093, p < .001**
STS	Perceptual	0.273	0.243	0.132	**t(27) = 5.938, p < .001**
	Categorical	0.262	0.209	0.111	**t(27) = 6.632, p < .001**
MPFC	Perceptual	0.232	0.275	0.127	**t(27) = 4.450, p < .001**
	Categorical	0.163	0.275	0.099	**t(27) = 3.143, p = .004**

We compared the overall fit of the two models, by performing a paired sample t-test between the arrays of R^2 values output from the regressions. We found no difference between the amount of variance explained by either model in any region (FFA: t(27) = 1.487, p = .149; STS: t(27) = 0.930, p = .360; OFA: t(27) = 0.848, p = .404; MPFC: t(27) = 1.290, p = .208).

Discussion

In this study, we aimed to explore the relative role of conceptual and stimulus-based cues in the neural mechanisms underpinning facial expression perception. The results highlight the particular role of conceptual information, as the conceptual model showed the highest correlations with the representational distances within three face-responsive regions and an MPFC region involved in the processing of theory of mind. Such results add to previous research into the role of these three cues in neural representations of facial expressions (Brooks et al., 2019; Sormaz, Watson, et al., 2016), by showing that representational distances in all regions we examined (the FFA, OFA, STS, and MPFC) are best explained by the similarities of emotion concepts, even after controlling for the variance shared by the other cues.

We found a similar pattern of results in all four regions that we examined, where the representational structure was more related to the structure of emotion concepts than to the presentational structures of the stimulus properties. This was unexpected, as we predicted that the similarities of the stimulus properties would best explain the variance in representational geometries in the core face regions (the FFA, OFA, and STS). However, the role of emotion concepts on representations in the FFA is consistent with the results of Brooks et al. (2019), who similarly reported that representational distances in the right fusiform gyrus are still explained by the similarities of emotion concepts after controlling for the similarities of several image properties (the similarity of face silhouettes, similarity of pixel intensity maps, and similarity of 'higher-level visual features' output from a computational model of object recognition). Our results expand upon these as we show that conceptual information can still explain representations after controlling for the similarities of both face shape and surface textures, two stimulus-based cues that play a considerable role in the perception of facial expressions (Sormaz, Watson, et al., 2016; Sormaz, Young, & Andrews, 2016). This pattern of results in the OFA suggests the accessibility of emotion concepts in the early perception of facial expressions. The result that the two stimulus-based cues did not explain the representational structure was unexpected as this region is involved in early stages of the visual processing of faces and facial expressions (Haxby et al., 2000). It may be the case that the representational distances of facial expressions within the OFA could have better been

explained by similarities of other low-level image properties. For example, Weibert et al. (2018) found that the correlations between GIST descriptors (spatial frequency distributions after passing images through a series of Gabor filters; Oliva & Torralba, 2001) explained the representational similarity of facial expressions in the three core face regions. Similarly, Brooks et al. (2019) used three measures of visual similarity as controls in the RSA regression model (the similarity of face silhouettes, similarity of pixel intensity maps, and similarity of 'higher-level visual features' as output from the HMAX model of object recognition). While the regression coefficients of these models were not reported, their measure of conceptual similarity did not explain representational distances within the OFA after controlling for these measures. Perhaps these low-level measures of image similarities can better explain the representational structure of expressions within the OFA than shape and surface similarities. For the MPFC these results were in line with our predictions as previous research has shown that representations of emotions within this area are modality independent (Peelen et al., 2010; Skerry & Saxe, 2014), and so are not structured around any property of the stimulus but rather conceptual knowledge of emotions.

To our knowledge, research has yet to assess whether shape and surface properties of facial expressions can explain representations in core face regions. Despite this, the result that these properties did not explain unique variance in the OFA and STS was surprising. Representational similarities of facial expressions within the OFA and STS are associated with perceptual similarity, as measured behaviourally (Said et al., 2010; Sormaz, Watson, et al., 2016), and research has also highlighted the important role of shape and surface properties in the perception of expressions (Sormaz, Watson, et al., 2016; Sormaz, Young, & Andrews, 2016). Similarly, in Study 1, we showed that pairwise perceptual similarities were explained above chance by shape and surface properties, even when controlling for conceptual information. Given these two lines of evidence, we expected to find that shape and surface similarities would explain representational similarities in these regions after controlling for conceptual similarity.

Additionally, we explored which of the perceptual and categorical tasks (from Study 1) was better associated with the representational structure of emotions within each brain region. The representational distances within all four regions were associated with both tasks above chance, and no differences were found between the models in any region. Sormaz, Watson, et al. (2016) found that perceptual similarity of expressions was associated with representational similarity in the OFA and STS, but not the FFA, whereas the results of our study suggested that the representational structure in all three of these regions is explained by perceptual similarity. A potential reason for the difference of results in the FFA is our measure of perceptual similarity, which does not require making subjective judgements of similarity.

In our study, we found that representations of facial expressions in the MPFC can be explained by conceptual knowledge of emotions after controlling for the similarities of the stimulus-properties. Additionally, perceptual similarity of expressions and categorisation errors explained representational similarity in this region in separate regression models. These results suggest that, while the MPFC may represent modality independent emotion concepts, it may also play a role in tasks requiring the perceptual discrimination and categorisation of facial expressions of emotion.

General Discussion

The aim of the present research was to examine the roles of emotion concepts and stimulus-based cues in the perceptual discrimination, categorisation, and brain representational structures of facial expressions of emotion. Study 1 showed that similarities of shape, surface, and conceptual cues are associated with both perceptual similarities of facial expressions and with patterns of categorisations. The results of Study 2 furthered this, by showing that the similarity of emotion concepts explains the similarity of neural representations of emotions in all four regions we

examined (the FFA, OFA, STS, and MPFC), after controlling for the same stimulus-based properties. Considering both the behavioural and neuroimaging studies together, our results suggest that conceptual knowledge of emotions impacts perception of facial expressions and representations of emotions within brain regions involved in the early perceptual processing of facial expressions of emotion.

Three questions were addressed within the research. The first was whether the conceptual similarity of emotions still explains perceptual similarity when emotion labels are not present in the perceptual task. Several studies have shown that the presence of emotion labels may influence the perception of facial expressions by activating emotion concepts (Fernández-Dols et al., 2008; Halberstadt & Niedenthal, 2001; Nook et al., 2015). Like Brooks and Freeman (2018), we have shown that the similarity of emotion concepts explains perceptual similarity, although we have furthered this by showing that this is still the case even when emotion labels are not present in the measurement of perceptual similarity. The second question asked what the role of these cues are in the explicit labelling of facial expressions. The results of Study 1 showed that the stimulus-based and conceptual cues explain the patterns of categorisation of facial expressions during an emotion labelling task, that is argued to recruit additional processes involved with assigning labels to stimuli (Palermo et al., 2013). The result that conceptual information explains these patterns of categorisation is comparable to the results of research showing that categorisations of gender and race is influenced by top-down expectations (Levin & Banaji, 2006; Macrae & Martin, 2007). Interestingly, the shape and surface cues explained behaviour during the two tasks to different extents, suggesting that perceptual tasks and labelling tasks may recruit different processes, or rely to different extents on different processes. Finally, we asked what the roles of these cues are in explaining the representational structures of facial expressions within several regions of the brain. While previous research had examined the independent contribution of stimulus-based and conceptual cues (Brooks et al., 2019; Sormaz, Watson, et al., 2016), we showed that the stimulus-based cues did not seem to explain unique variance of the representational structures in any of the brain regions after controlling for conceptual cues.

The role of emotion concepts in the perception, recognition, and neural representation of facial expressions is consistent with several models of social perception suggesting that top-down beliefs and prior expectations interact with bottom-up processes to shape our visual perception of others (Freeman & Johnson, 2016; Stolier, Hehman, & Freeman, 2018; Tamir & Thornton, 2018). Like research showing that the perception of race and gender categories is influenced by internal stereotype knowledge (Hugenberg & Bodenhausen, 2004; Macrae & Martin, 2007), the results of the current research support the suggestion that the perception and categorisation of others involves an interaction between lower-level features and higher-level conceptual information, and show that these interactions may extend into the domain of emotion recognition. One model (Freeman & Johnson, 2016) proposes a network of brain regions involved in social perception, suggesting that the anterior temporal lobe retrieves social-conceptual information, and the orbitofrontal cortex (OFC) implements top-down predictions that shape representations of faces in the fusiform gyrus. Given that a region within the OFC is involved in top-down processes during object recognition (Bar, 2003; Bar et al., 2006), and in predicting the affective value of stimuli (Shenhav, Barrett, & Bar, 2013), it is possible that this region may be involved in implementing the top-down cues that shape representations of facial expressions in the FFA. Future research could examine the role of this region within the perception and recognition of facial expressions.

The focus of this discussion thus far has been on the role of conceptual information, however we aimed to examine the relative influence of conceptual and stimulus-based cues. The results for the role of the stimulus-based cues, however, have not been quite as clear. In Study, 1 we found that the similarities of both shape and surface cues can explain perceptual similarity and categorisation errors, where shape information played a particular role in explaining behaviour in the Perceptual Task and surface information best explained behaviour in the categorical task. However, the results

of Study 2 are not entirely consistent with the behavioural results, as these cues did not share any unique variance with the representational distances in any region (after controlling for the similarity of emotion concepts). As we found that shape and surface information still play a role in explaining perceptual similarity after controlling for conceptual similarity, and there is research suggesting that perceptual similarity is associated with representational similarities within several face processing regions (Said et al., 2010; Sormaz, Watson, et al., 2016), it was unexpected that these cues did not explain unique variance in the structure of neural representations in the brain regions that we examined. We note that the shape and surface models used in each study were not identical, as they were derived from the faces used in the Perceptual task for Study 1, and from the faces presented in the fMRI task for Study 2. Additionally, shape and surface similarities were calculated within-identity for Study 1 and between-identity for Study 2, due to the stimuli that were used in each study. Perhaps the differences between the stimuli used to create the models, or the variability between the within- and between-identity measurements of shape and surface similarity, may have been one cause of the differences between the role of the models in the two studies.

Several studies have examined the contribution of face shape and surface texture to the processing of face identity, with results generally suggesting that explicit recognition of identity relies more on surface cues over shape (Andrews et al., 2016; Itz, Golle, Luttmann, Schweinberger, & Kaufmann, 2017). One study, however, used EEG to show that an early face-sensitive ERP component (the N170) is sensitive to changes in face shape but not changes in surface reflectance, suggesting that face shape contributes to identity processing before surface textures (Caharel, Jiang, Blanz, & Rossion, 2009). Our results are consistent with this finding, in that we showed that shape cues explain perceptual discrimination more than surface cues, but surface cues play more of a role than shape cues in explaining the patterns of explicit labelling of expressions. Studies using fMRI with adaptation paradigms showed that the FFA and OFA are sensitive to changes in both shape and surface cues (Andrews et al., 2016; Jiang, Dricot, Blanz, Goebel, & Rossion, 2009), while another study using RSA showed that the dissimilarities of representations of face identities in the OFA are explained primarily by lower-level image-based properties of the stimuli, whereas dissimilarities in the FFA are explained primarily by higher-level human-rated properties such as the similarities of social traits and attractiveness (Tsantani et al., 2021). Our results suggest that, in the domain of emotion processing, high-level conceptual cues may play more of a role than low-level stimulus properties in explaining the representational dissimilarities of facial expressions in these regions.

While our results provided an understanding of the relative influence of conceptual and stimulus-based information on behaviour and brain representations of emotion, there are some further limitations to consider. In particular, we cannot be sure of any causal relationship between any particular cue and behaviour or neural representations. For example, it would not be appropriate to infer that a given pair of facial expressions are perceptually more similar *because* the corresponding emotion concepts are more similar. This issue of causality raises a further conceptual issue, which is that we cannot be sure that participants are 'using' any of the three cues we examined. For example, in the regression model for the Perceptual Task in Study 1 the average beta for the conceptual model was larger than that of the surface model, but the claim that participants use conceptual information more than surface information would be unwarranted. Future work could perhaps introduce some form of experimental manipulation (e.g. by removing each stimulus-cue after Sormaz, Young, and Andrews (2016), or disrupt access to emotion concepts via semantic satiation (Gendron et al., 2012; Lindquist et al., 2006)) to assess behaviour and neural representations where the availability of each cue is removed. It could also be interesting to combine RSA with EEG to investigate the timing at which different types of information may come into play when processing facial expressions of emotion.

The low-to-moderate averaged R-squared values produced in the regression models (in Study 2) are also worth noting, as they suggest that there are additional sources of variance that can account for the representational structure within the four ROIs. Additionally, no model reached the noise ceiling

(with the exception of the Perceptual model in the MPFC), suggesting that they do not perform as well as any theoretical model could, given the noise in the data (Nili et al., 2014).

One important limitation to consider is that our measure of surface similarity used greyscale images while participants were presented with colour images. By converting images to greyscale, our model of surface similarity was not able to capture the information provided by colour cues. Previous research, however, has demonstrated the importance of colour in the recognition of facial expressions of emotion. In particular, Benitez-Quiroz, Srinivasan and Martinez (2018) demonstrated that facial expressions can be classified above-chance using only face colour features. Moreover, the authors showed that participants could reliably categorise emotions using only colour cues, independently of shape information. Thorstenson, Elliot, Pazda, Perrett and Xiao (2018) additionally showed that participants can reliably manipulate the colour of emotional faces to match held emotion-colour associations. In our study, the discrepancy between using colour images and colour-free surface models could also have contributed to the differences with the results of Sormaz, Watson et al. (2016), who presented participants with greyscale images and used greyscale images to calculate surface similarity. We think that it would be very interesting for future studies to build models based on colour cues of emotional faces (Benitez-Quiroz et al., 2018) and on the associations between emotions and colour cues (Thorstenson et al., 2018), and compare these models to the ones we included here.

More generally, the three sources of information we chose to examine are not an exhaustive list of factors that can affect the processing of facial expressions. Several low-level image properties are reported to explain representations of facial expressions of emotion within the core face regions. Weibert et al. (2018) found that the correlations between the spatial frequencies of pairs of images (measured using GIST descriptors; Oliva & Torralba, 2001) explained the representational structure within the three core face regions. Similarly, Brooks et al. (2019) used the similarities of the silhouettes, pixel-by-pixel intensities, and a measure output from a computational model of object recognition (HMAX; Serre, Oliva, & Poggio, 2007) as control measures in the regression models. Furthermore, the dissimilarity of representations of faces in face processing regions has been explained by high-level judgements of social traits, attractiveness, and race and gender categories (Stolier & Freeman, 2016; Tsantani et al., 2021). Perhaps future research could employ a wider range of models to provide a more complete picture of the sources of information that can explain how facial expressions of emotion are represented. In addition to the cues examined within this research, the processing of emotions from facial expressions is well documented to be affected by multiple social, cognitive, affective, personality, and hormonal measures, so it is feasible that these factors may account for additional variance of the representational structures we observed. Should future work include participant-level measures (e.g. personality traits), more complex multilevel regression models could be used to account for this variance at the participant-level (in addition to variance between emotion-pairs).

It is worth discussing the use of the six basic emotions in these two studies, following the work by Ekman and colleagues (e.g. Ekman, 1970; Ekman & Oster, 1979) and because these six categories have been used extensively in subsequent research. However, the experience of an emotion is clearly not confined to these six distinct categories, and is instead a much a more subjective and content-rich experience (Barrett, Mesquita, Ochsner, & Gross, 2007). It has been argued that emotions do not have such distinct boundaries, and that evidence that the six basic facial expressions can reliably be identified should not be taken as evidence of these boundaries for emotional experiences (Barrett, 2006). A data-driven study in which participants viewed emotionally evocative videos showed that participants reliably self-reported experiencing an emotion that falls into one of 27 distinct categories (Cowen & Keltner, 2017). In the domain of social perception, Skerry and Saxe (2015) showed that when identifying which of 20 emotions that characters in short stories were experiencing, dissociable patterns of activation for each emotional category were elicited in several regions involved with theory of mind. Together, these studies show that both the perception

and experience of emotions is not necessarily confined to the six basic emotions as has been used within this research. Similarly, the present work does not assume categorical separations between emotions, but investigates their representational structure. Moreover, although the emotion categories used in the present studies may not have captured a wider range of emotional experiences, similar procedures could be applied using a much wider range of labels.

In conclusion, these results show the roles of shape, surface, and conceptual cues in behavioural and neural representations of facial expressions. Behaviourally, we showed that all three cues explain patterns of perceptual discrimination and the explicit categorisation of facial expressions. Using fMRI, the important role of conceptual information was demonstrated in the representational structure of facial expressions within several regions involved in the processing of facial expressions. These studies pave the way for future research to examine facial expression processing from the perspective of an integration between top-down and bottom-up cues.

Acknowledgments

We would like to thank Rebecca Saxe for generously providing the data used to construct the conceptual model.

Funding: This work was supported by the College of Health, Medicine and Life Sciences, Brunel University London.

Data availability

Data (participant RDMs and models) and analysis code for reproducing the results for Experiment 1 are available here: https://osf.io/xgw5a/. Data (raw anonymised imaging data, participant RDMs and models) and analysis code for reproducing the results for Experiment 2 are available here: https://osf.io/34fm7/.

References

Andrews, T. J., Baseler, H., Jenkins, R., Burton, A. M., & Young, A. W. (2016). Contributions of feature shapes and surface cues to the recognition and neural representation of facial identity. *Cortex*, *83*, 280–291. https://doi.org/10.1016/j.cortex.2016.08.008

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600–609. https://doi.org/10.1162/089892903321662976

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 449–454. https://doi.org/10.1073/pnas.0507062103

Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science*, *1*(1), 28–58. https://doi.org/10.1111/j.1745-6916.2006.00003.x

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The Experience of Emotion. *Annual Review of Psychology*, *58*(1), 373–403. https://doi.org/10.1146/annurev.psych.58.110405.085709

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2018). Facial color is an efficient mechanism

to visually transmit emotion. *Proceedings of the National Academy of Sciences*, 201716084. https://doi.org/10.1073/pnas.1716084115

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brooks, J. A., Chikazoe, J., Sadato, N., & Freeman, J. B. (2019). The neural representation of facial-emotion categories reflects conceptual structure. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32), 15861–15870. https://doi.org/10.1073/pnas.1816408116

Brooks, J. A., & Freeman, J. B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature Human Behaviour*, *2*(8), 581–591. https://doi.org/10.1038/s41562-018-0376-6

Bruce, V., & Young, A. W. (1998). *In the eye of the behonder: The science of face perception*. Oxford University Press.

Caharel, S., Jiang, F., Blanz, V., & Rossion, B. (2009). Recognizing an individual face: 3D shape contributes earlier than 2D surface reflectance information. *NeuroImage*, *47*(4), 1809–1818. https://doi.org/10.1016/j.neuroimage.2009.05.065

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208. https://doi.org/10.1016/S0042-6989(01)00002-5

Calvo, M. G., Avero, P., Fernández-Martín, A., & Recio, G. (2016). Recognition thresholds for static and dynamic emotional faces. *Emotion*, *16*(8), 1186–1200. https://doi.org/10.1037/emo0000192

Carlin, J. D., & Kriegeskorte, N. (2017). Adjudicating between face-coding models with individual-face fMRI responses. *PLoS Computational Biology*, *13*(7), 1–28. https://doi.org/10.1371/journal.pcbi.1005604

Cornsweet, T. N. (1962). The Staircase-Method in Psychophysics. *The American Journal of Psychology*, *75*(3), 485. https://doi.org/10.2307/1419876

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). FMRI item analysis in a theory of mind task. *NeuroImage*, *55*(2), 705–712. https://doi.org/10.1016/j.neuroimage.2010.12.040

Dotsch, R., & Todorov, A. (2012). Reverse Correlating Social Face Perception. *Social Psychological and Personality Science*, *3*(5), 562–571. https://doi.org/10.1177/1948550611430272

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., … Saxe, R. (2013). Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE*, *8*(9), e75468. https://doi.org/10.1371/journal.pone.0075468

Ekman, P. (1970). Universal Facial Expressions of Emotion. *California Mental Health Research Digest*, *8*(4), 151–158.

Ekman, P., & Friesen, W. V. (1974). Nonverbal Behavior and Psychopathology. In R. J. Friedman & M. Katz (Eds.), *The psychology of depression: Contemporary theory and research* (pp. 3–31). Washington, DC: Winston & Sons.

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

Ekman, P., & Oster, H. (1979). Facial Expressions of Emotion. *Annual Review of Psychology*, *30*(1), 527–554.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227–240. https://doi.org/10.1016/0010-0277(92)90002-Y

Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fernández-Dols, J.-M., Carrera, P., Barchard, K. A., & Gacitua, M. (2008). False recognition of facial expressions of emotion: Causes and implications. *Emotion*, *8*(4), 530–539. https://doi.org/10.1037/a0012724

Freeman, J. B. (2018). Doing Psychological Science by Hand. *Current Directions in Psychological Science*, *27*(5), 315–323. https://doi.org/10.1177/0963721417746793

Freeman, J. B., & Johnson, K. L. (2016). More Than Meets the Eye: Split-Second Social Perception. *Trends in Cognitive Sciences*, *20*(5), 362–374. https://doi.org/10.1016/j.tics.2016.03.003

Gamond, L., & Cattaneo, Z. (2016). The dorsomedial prefrontal cortex plays a causal role in mediating in-group advantage in emotion recognition: A TMS study. *Neuropsychologia*, *93*, 312–317. https://doi.org/10.1016/j.neuropsychologia.2016.11.011

Gendron, M., Lindquist, K. A., Barsalou, L., & Barrett, L. F. (2012). Emotion Words Shape Emotion Percepts. *Emotion*, *12*(2), 314–325. https://doi.org/10.1037/a0026007

Goesaert, E., & Op de Beeck, H. P. (2013). Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. *Journal of Neuroscience*, *33*(19), 8549–8558. https://doi.org/10.1523/JNEUROSCI.1829-12.2013

Halberstadt, J., & Niedenthal, P. M. (2001). Effects of emotion concepts on perceptual memory for emotional expressions. *Journal of Personality and Social Psychology*, *81*(4), 587–598. https://doi.org/10.1037//0022-3514.81.4.587

Harmer, C. J., Thilo, K. V., Rothwell, J. C., & Goodwin, G. M. (2001). Transcranial magnetic stimulation of medial-frontal cortex impairs the processing of angry facial expressions. *Nature Neuroscience*, *4*(1), 17–18. https://doi.org/10.1038/82854

Harris, R. J., Young, A. W., & Andrews, T. J. (2014). Brain regions involved in processing facial identity and expression are differentially selective for surface and edge information. *NeuroImage*, *97*, 217–223. https://doi.org/10.1016/j.neuroimage.2014.04.032

Harry, B., Williams, M. A., Davis, C., & Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in Human Neuroscience*, *7*, 1–6. https://doi.org/10.3389/fnhum.2013.00692

Haxby, J. V, Hoffman, E. a, & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in Social Categorization. *Psychological Science*, *15*(5), 342–345. https://doi.org/10.1111/j.0956-7976.2004.00680.x

Itz, M. L., Golle, J., Luttmann, S., Schweinberger, S. R., & Kaufmann, J. M. (2017). Dominance of

texture over shape in facial identity processing is modulated by individual abilities. *British Journal of Psychology*, *108*(2), 369–396. https://doi.org/10.1111/bjop.12199

Jiang, F., Dricot, L., Blanz, V., Goebel, R., & Rossion, B. (2009). Neural correlates of shape and surface reflectance information in individual faces. *Neuroscience*, *163*(4), 1078–1091. https://doi.org/10.1016/j.neuroscience.2009.07.062

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*, *60*(4), 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055

Katsikitis, M. (1997). The classification of facial expressions of emotion: a multidimensional-scaling approach. *Perception*, *26*(5), 613–626.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4), 1–28. https://doi.org/10.3389/neuro.06.004.2008

Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar Representations of Emotions Across Faces and Voices. *Emotion*, *17*(6), 912–937. https://doi.org/10.1037/emo0000282.supp

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud faces database. *Cognition and Emotion*, *24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces: The role of race categories. *Journal of Experimental Psychology: General*, *135*(4), 501–512. https://doi.org/10.1037/0096-3445.135.4.501

Lindquist, K. A., Barrett, L. F., Bliss-Moreau, E., & Russell, J. A. (2006). Language and the perception of emotion. *Emotion*, *6*(1), 125–138. https://doi.org/10.1037/1528-3542.6.1.125

Macrae, C. N., & Martin, D. (2007). A boy primed Sue: feature-based processing and person construal. *European Journal of Social Psychology*, *37*(5), 793–805. https://doi.org/10.1002/ejsp.406

Marneweck, M., Loftus, A., & Hammond, G. (2013). Psychophysical measures of sensitivity to facial expression of emotion. *Frontiers in Psychology*, *4*, 1–6. https://doi.org/10.3389/fpsyg.2013.00063

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., … Goh, A. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian brief affect recognition test (JACBART). *Journal of Nonverbal Behavior*, *24*(3), 179–209. https://doi.org/10.1023/A:1006668120583

Mattavelli, G., Cattaneo, Z., & Papagno, C. (2011). Transcranial magnetic stimulation of medial prefrontal cortex modulates face expressions processing in a priming task. *Neuropsychologia*, *49*(5), 992–998. https://doi.org/10.1016/j.neuropsychologia.2011.01.038

Mckelvie, S. J. (1973). The meaningfulness and meaning of schematic faces. *Perception and Psychophysics*, *14*(2), 343–348.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis, 5th Ed*. New York: Wiley.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–

2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human Object-Similarity Judgments Reflect and Transcend the Primate-IT Object Representation. *Frontiers in Psychology*, *4*, 1–22. https://doi.org/10.3389/fpsyg.2013.00128

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Nook, E. C., Lindquist, K. A., & Zaki, J. (2015). A new look at emotion perception: Concepts speed and shape facial emotion recognition. *Emotion*, *15*(5), 569–578. https://doi.org/10.1037/a0039166

O'Callaghan, C., Kveraga, K., Shine, J. M., Adams, R. B., & Bar, M. (2017). Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and Cognition*, *47*(3), 63–74. https://doi.org/10.1016/j.concog.2016.05.003

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

Palermo, R., O'Connor, K. B., Davis, J. M., Irons, J., & McKone, E. (2013). New Tests to Measure Individual Differences in Matching and Labelling Facial Expressions of Emotion, and Their Association with Ability to Recognise Vocal Emotions and Facial Identity. *PLoS ONE*, *8*(6), e68126. https://doi.org/10.1371/journal.pone.0068126

Pallett, P. M., & Meng, M. (2013). Contrast negation differentiates visual pathways underlying dynamic and invariant facial processing. *Journal of Vision*, *13*(14), 1–18. https://doi.org/10.1167/13.14.13.doi

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal Representations of Perceived Emotions in the Human Brain. *Journal of Neuroscience*, *30*(30), 10127–10134. https://doi.org/10.1523/JNEUROSCI.2161-10.2010

Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision*, *10*(5), 11–11. https://doi.org/10.1167/10.5.11

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6424–6429. https://doi.org/10.1073/pnas.0700622104

Shenhav, A., Barrett, L. F., & Bar, M. (2013). Affective value and associative processing share a cortical substrate. *Cognitive, Affective and Behavioral Neuroscience*, *13*(1), 46–59. https://doi.org/10.3758/s13415-012-0128-4

Skerry, A. E., & Saxe, R. (2014). A Common Neural Code for Perceived and Inferred Emotion. *Journal of Neuroscience*, *34*(48), 15997–16008. https://doi.org/10.1523/JNEUROSCI.1676-14.2014

Skerry, A. E., & Saxe, R. (2015). Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology*, *25*(15), 1945–1954. https://doi.org/10.1016/j.cub.2015.06.009

Smith, L., & Klein, R. (1990). Evidence for Semantic Satiation: Repeating a Category Slows Subsequent Semantic Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(5), 852–861. https://doi.org/10.1037/0278-7393.16.5.852

Sormaz, M., Watson, D. M., Smith, W. A. P., Young, A. W., & Andrews, T. J. (2016). Modelling the perceptual similarity of facial expressions from image statistics and neural responses.

NeuroImage, *129*, 64–71. https://doi.org/10.1016/j.neuroimage.2016.01.041

Sormaz, M., Young, A. W., & Andrews, T. J. (2016). Contributions of feature shapes and surface cues to the recognition of facial expressions. *Vision Research*, *127*, 1–10. https://doi.org/10.1016/j.cortex.2016.08.008

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, *19*(6), 795–797. https://doi.org/10.1038/nn.4296

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A Dynamic Structure of Social Trait Space. *Trends in Cognitive Sciences*, *22*(3), 197–200. https://doi.org/10.1016/j.tics.2017.12.003

Stolier, R. M., Hehman, E., Freeman, J. B., Keller, M. D., & Walker, M. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*(37), 9210–9215. https://doi.org/10.1073/pnas.1807222115

Suzuki, A., Hoshino, T., & Shigemasu, K. (2006). Measuring individual differences in sensitivities to basic emotions in faces. *Cognition*, *99*(3), 327–353. https://doi.org/10.1016/j.cognition.2005.04.003

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Thompson, C. G., Kim, R. S., Aloe, A. M., & Becker, B. J. (2017). Extracting the Variance In flation Factor and Other Multicollinearity Diagnostics from Typical Regression Results. *Basic and Applied Social Psychology*, *39*(2), 81–90. https://doi.org/10.1080/01973533.2016.1277529

Thorstenson, C. A., Elliot, A. J., Pazda, A. D., Perrett, D. I., & Xiao, D. (2018). Emotion-Color Associations in the Context of the Face. *Emotion*, *18*(7), 1032–1042. https://doi.org/10.1037/emo0000358

Tiddeman, B. P., Stirrat, M. R., & Perrett, D. I. (2005). Towards realism in facial image transformation: Results of a wavelet MRF method. *Computer Graphics Forum*, *24*(3), 449–456. https://doi.org/10.1111/j.1467-8659.2005.00870.x

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). Ffa and ofa encode distinct types of face identity information. *Journal of Neuroscience*, *41*(9), 1952–1969. https://doi.org/10.1523/JNEUROSCI.1449-20.2020

Wegrzyn, M., Riehle, M., Labudda, K., Woermann, F., Baumgartner, F., Pollmann, S., … Kissler, J. (2015). Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. *Cortex*, *69*, 131–140. https://doi.org/10.1016/j.cortex.2015.05.003

Weibert, K., Flack, T. R., Young, A. W., & Andrews, T. J. (2018). Patterns of neural response in face regions are predicted by low-level image properties. *Cortex*, *103*, 199–210. https://doi.org/10.1016/j.cortex.2018.03.009

White, M. (2001). Effect of photographic negation on matching the expressions and identities of faces. *Perception*, *30*(8), 969–981. https://doi.org/10.1068/p3225

Winston, J. S., Henson, R. N. A., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, *92*(3), 1830–1839. https://doi.org/10.1152/jn.00155.2004

Wölwer, W., Lowe, A., Brinkmeyer, J., Streit, M., Habakuck, M., Agelink, M. W., … Cordes, J. (2014). Repetitive transcranial magnetic stimulation (rTMS) improves facial affect recognition in schizophrenia. *Brain Stimulation*, *7*(4), 559–563. https://doi.org/10.1016/j.brs.2014.04.011

Zhang, H., Japee, S., Nolan, R., Chu, C., Liu, N., & Ungerleider, L. G. (2016). Face-selective regions differ in their ability to classify facial expressions. *NeuroImage*, *130*, 77–90. https://doi.org/10.1016/j.neuroimage.2016.01.045