Facial first impressions form two clusters representing approach-avoidance

Alex L. Jones[a,*], Robin S. S. Kramer[b]

[a] Department of Psychology, Swansea University, UK

[b] School of Psychology, University of Lincoln, UK

* Corresponding author at: Department of Psychology, Swansea University, Swansea, SA2 8PP, UK.

E-mail address: alex.l.jones@swansea.ac.uk (A. L. Jones).

Abstract

Existing models of facial first impressions indicate between two and four factors that underpin all social trait judgements. Here, we submitted several large databases of these first impression ratings to unsupervised learning algorithms with the aim of clustering together faces, rather than traits, to examine the ways in which impressions may be grouped together. Experiment 1 revealed two clusters of faces that exist in both a full-dimensional, and two- or three-factor representations, of social impressions, while Experiment 2 indicated that these clusters also emerged in additional datasets. In Experiment 3, using Bayesian modelling approaches, we extracted the impression profile of each cluster and also derived a vector that maximally separated the clusters. The resulting vector related strongly to the valence and approachability components in existing models. In a further test of our model, we showed in Experiment 4 that mere facial appearance, rather than perceptions, is sufficient to separate these clusters, demonstrating probabilistically that facial cues like smiling may drive the perceptual profile that gives rise to the perceptual clusters. Finally, Experiment 5 showed that observer responses to faces in these two clusters mapped closely on to approach-avoidance behaviour, with observers responding rapidly and without instruction to approach faces from one cluster over the other. Taken together, our findings provide compelling evidence, drawing upon both computational and behavioural approaches, that existing models of social impressions are realised practically in terms of basic approach-avoidance mechanisms.

Keywords: face perception, impression formation, statistical learning, clustering

1. Introduction

Facial appearance plays a crucial role in everyday life. From inferring basic information about individuals, such as their age (Porcheron et al., 2013) and sex (George & Hole, 1995; Russell, 2009), through to complex hiring and voting decisions (Chiao et al., 2008; Little, 2014; Marlowe et al., 1996), faces play a central role in human social interactions. Decisions about individuals based on facial appearance form without prompting (A. L. Jones et al., 2019; Ritchie et al., 2017) and may occur in less than 250 milliseconds (Borkenau et al., 2009). The persistent and ubiquitous nature of these judgements points to a fundamental role in human psychology – helping us to make survival-relevant decisions, such as identifying aggressive or unhealthy individuals (Haselhuhn et al., 2015; Henderson et al., 2017).

Significant research efforts have been dedicated to understanding the fundamental psychological processes and variations in facial appearance that give rise to first impressions. A critical finding is that perceptions based on appearance tend to be correlated – for example, judgements of trustworthiness are associated, to a greater or lesser extent, with almost every other social trait perception, and certain facial characteristics (e.g., a positive-looking neutral expression) seem to cause some perceptions more than others (A. L. Jones et al., 2018; Todorov, 2008). Put simply, there is no perception that cannot be explained in part by another, and these correlated perceptions are associated with statistical regularities in facial appearances (Todorov et al., 2013). In recent years, advancing knowledge in this area has abandoned more top-down, hypothesis-driven approaches – such as finding the facial attributes linked with health perception (A. L. Jones et al., 2016) – in favour of bottom-up, data-driven techniques (Holzleitner et al., 2019; Sutherland et al., 2017) that are capable of discerning the underlying structure of perceptions and associated facial appearances.

Initial efforts taking this approach utilised principal component analysis (PCA) and applied this technique to multiple trait dimensions, generating the well-known valence-dominance model of face evaluation (Balas et al., 2018; Morrison et al., 2017; Oosterhof & Todorov, 2008). In this approach, a set of orthogonal components are derived, with each explaining successively less variation in the original perceptual space, and the loadings of each impression on those components which explain substantial variance are calculated. In the original derivation of the model, high loadings for traits like trustworthiness and attractiveness were found on the first principal component, while dominance and aggressiveness loaded on to the second component (Oosterhof & Todorov, 2008). It is also possible to extract the appearance of faces with high or low scores on these components, indicating the morphology associated with these fundamental axes. This simple, two-dimensional model also aligns with social cognitive models of impression formation, such as the warmth-competence dimensions of intergroup perceptions (Fiske, 2018).

A criticism of this valence-dominance model is that while it was derived from real faces, it has been extensively validated and extended using computer-generated faces with neutral expressions, only varying on a constrained set of dimensions that may induce systematic biases in the perception of social traits (Balas & Pacella, 2017). To correct this, recent models have featured realistic, ‘ambient’ images – photographs of individuals of different ages and expressions, similar to those found on social media profiles. By collecting multiple impressions of these ambient images, and conducting a factor analysis (FA) on these impressions, an expanded three-dimensional model emerges (Sutherland et al., 2013). This model (approachability, youthful-attractiveness, and dominance) appears to better approximate how real faces are socially evaluated, and in contrast to PCA-derived solutions (which are orthogonal by definition), it captures correlations in underlying structure. For example, the youthful-attractiveness factor correlates with the approachability factor, mirroring how attractive faces induce approach behaviours in observers (Kramer et al., 2020). However, both of these components are almost independent of (i.e., are orthogonal to) the dominance component (Sutherland et al., 2013). Additionally, this three-factor structure seems to emerge at the within-observer level, as well as when aggregating ratings across observers (Sutherland et al., 2019).

However, more recent evidence has demonstrated that these two- and three-factor models may not fully capture the underlying structure underpinning social judgements due to an insufficient sample of impression ratings. The two- and three-dimensional models (Oosterhof & Todorov, 2008; Sutherland et al., 2013) were derived by asking observers to provide around 13 trait judgements for faces. When observers provided judgements regarding 100 traits, a four-factor structure emerged – warmth, competence, and both female and youth stereotypes (Lin et al., 2019), suggesting a more comprehensive set of evaluations leads to more factors. Other work has attempted to assess the universality of the two- and three-factor models by investigating the structure of impressions from many different global regions using 13 trait impressions (Jones et al., 2021). Interestingly, these comprehensive analyses found that the number of factors emerging from both PCA and FA approaches varied between two, three, and four factors, depending on the global region, but supported the general structure of a clearly low-dimensional representation of the psychology of social impressions. However, these analyses also show that the "true" number of factors that underpin social impressions shows significant variability.

Despite the inconsistent number of factors that emerge across studies, there is convergence in their typical content, particularly in the factor that explains the largest amount of variability which corresponds typically to positive versus negative valence (Oosterhof & Todorov, 2008; Sutherland et al., 2013), and researchers are in agreement that the wide range of impressions we form of others can be reduced to a smaller number of factors or core dimensions – a more compact representation of social perception does exist. However, there is an important distinction that the current literature has failed to address. While existing analyses demonstrate observers can make judgements of others based on a reduced number of factors, they do not speak to how observers use these factors. For example, it is unknown whether faces are judged mostly according to their positions along one of the primary factors, or whether impressions are arrived at by considering all factors simultaneously – what weightings are these factors given? We can cast the argument geometrically, treating the factors as vectors in face space, where each location or data point is a face, represented by a vector drawn from the origin. It is not known whether these vectors are short, indicating little variability in their use (i.e., most faces are judged similarly on the factor, but it is necessary to explain variance in judgements) or long, indicating greater variability (i.e., faces vary widely in their position on the factor). Given that the majority of research has focussed on the use of FA or PCA to reduce the number of variables into a smaller number of factors (e.g., distilling social perceptions into trustworthiness or aggression), we note the reverse perspective – clustering faces into groups based on similarities in perceptions – has been ignored and is arguably more relevant to our understanding of social perception formation. Existing methods point to some inferences about this, given that PCA solutions provide an explained percentage of variance in the original data. For instance, two-factor models tend to find that the first component ('valence') explains almost twice the amount of variance in comparison with the second component ('dominance'; Oosterhof & Todorov, 2008), indicating a substantially more important role of the former when judging faces.

Precedent for this approach can be found in neighbouring literature. Although substantial research supports the idea that an individual's personality can be summarised using five (John & Srivastava, 1999) or six (Lee & Ashton, 2004) underlying dimensions, more recent work has investigated how such a multidimensional space is populated. Analyses identified only four clusters or ‘personality types’ (Gerlach et al., 2018), demonstrating that large areas of this space remain unoccupied and that specific combinations of personality trait scores frequently co-occur. Indeed, if these four types are capable of describing how individuals vary with regard to personality, they may also represent candidates for the clusters we identify in the current study, where we explore the perceptions of others. Another promising possibility may be the biologically-inspired decision to approach or avoid (Gray, 1970). Evidence to support this classification is provided by numerous disciplines (Cacioppo & Berntson, 1994; Depue & Collins, 1999; Tellegen, 1985; Watson & Clark, 1997) and the motivation to approach or avoid a given target is thought to be fundamental to our personalities and decision-making processes (Elliot & Thrash, 2002). As such, it would make intuitive sense that this simple, binary decision might underlie more complex social trait judgements.

It is worth considering whether a limited number of clusters exist within the unending list of social trait perceptions that one might apply as it relates directly to the reality of these perceptions. Given that trait perceptions are correlated, it is likely that many kinds of judgements are similarly low for a given face (e.g., if a face is judged as aggressive, it is also likely to be judged as low on trustworthiness, warmth, and approachability; Todorov, 2008) and there may be a typical appearance that elicits this pattern of trait judgements. This approach proceeds differently to the construction of models with an interchangeable number of dimensions that attempt to capture all the possible ways in which we judge others, with such models giving little priority to the importance of certain factors. Instead, we focus on determining groups or clusters of faces that share similar evaluative profiles, along with any potential associated appearances.

2. Experiment 1: Cluster analysis of social trait ratings in full and theoretical face-space

Our first aim was to test for the presence of distinct clusters of identities that emerge from multidimensional trait impression data. To examine this, we employed mean shift, a non-parametric, unsupervised feature space algorithm (Comaniciu & Meer, 2002), and applied this to a large, open access set of faces – the 10k US Adult Faces Database (Bainbridge et al., 2013), examining these clusters in the full set of data, as well as splitting the data by female and male faces. We also sought to examine the robustness of the resulting clusters by testing for the presence of clusters in the two- (Oosterhof & Todorov, 2008) and three-dimensional (Sutherland et al., 2013) factor representations of this multidimensional trait space.

2.1. Method

2.1.1. Stimuli

The 10k US Adult Faces Database (Bainbridge et al., 2013) comprises 2,222 adult face images taken from real world scenarios, mirroring the unconstrained, 'ambient' images used in three-factor models (Sutherland et al., 2013). The set consists of 57.1% men and 83.7% White faces (9.9% Black, 3.1% Asian, and 3.2% Hispanic). Sex and ethnicity data were previously determined by a set of Amazon Mechanical Turk (MTurk) workers. Each of the face photographs was rated on 14 trait dimensions generated from free-flow descriptions of a set of faces (Oosterhof & Todorov, 2008) – *emotionally stable, attractive, sociable, confident, boring, aggressive, weird, caring, unhappy, responsible, intelligent, trustworthy, mean*, and *egotistic*. Faces were also rated for their antonyms (e.g., happy vs unhappy, unattractive vs attractive), but only the original 14 traits were retained for analysis. Faces were previously rated by 1,274 MTurk workers, and trait ratings were averaged across workers to provide a score for each face, and these were *z*-scored before subsequent analyses. Thus, each face is represented by a single data point within this 14-dimensional space.

2.1.2. Clustering

Rather than employing dimension reduction techniques such as PCA or FA in order to compress the 14 traits into a lower-dimensional space, here we are interested in extracting clusters of individual faces that occupy particular regions within this space. To do this, we used the mean shift algorithm (Comaniciu & Meer, 2002). Designed for applications in image segmentation and computer vision (Comaniciu & Meer, 1999), mean shift proceeds by moving a flat (i.e., non-Gaussian) circular kernel of a given radius through a multidimensional space. The window assesses neighbourhoods of points within the circle, computing a candidate centroid for that neighbourhood through deriving Euclidean distances between points. On each iteration, the algorithm moves towards regions of the highest density of points, updating the candidate centroid to represent the average of the points in its neighbourhood. The algorithm stops when the change in candidate centroids is small, and very similar candidate centroids are then discarded. The result is a set of centroids, along with clusters of identified points that are nearest to them.

Mean shift has significant benefits over other clustering algorithms and dimension reduction techniques. It is non-parametric and thus requires no assumptions of linearity (unlike PCA or FA), and importantly, it requires no researcher decisions on the number of clusters to search for (or equivalently, the numbers of factors to retain). Mean shift has a single setting that must be optimised – the radius of the kernel, which determines the amount of points the algorithm considers at once. Here, to select the radius, we used a uniform prior (from zero to a maximum that encompassed all points at once, thus making the clustering redundant), evaluating the performance of the clustering algorithm at each radius using the

Silhouette Coefficient (*SC*; Rousseeuw, 1987). Silhouette scores compare the mean distance between a particular data point and all other points within its assigned cluster against the mean distance between that data point and all other points in its next-nearest cluster. These ratio scores are then averaged across all data points to provide the *SC*. Here, the radius that produced the highest *SC* was selected to derive the number of clusters (radius = 3.2). We also used a far simpler clustering approach, K-Means, to confirm the presence of clusters and *SC* scores.

2.1.3. Theoretical social perception spaces

To test for the presence of clusters in existing theoretical social face perception spaces, we generated the models from the data according to the descriptions provided by the original authors. For the two-dimensional valence/dominance model, we subjected the 14 trait ratings of all the faces to a principal components analysis, and projected the data onto the first two components, as described by Oosterhof and Todorov (2008). We label this as principal component space (PCS). For the three factor model, we used factor analysis, retaining three components, fit with maximum likelihood estimation and with oblimin rotation (Sutherland et al., 2013), which we term factor analysis space (FAS). Note that this approach should, if existing models are robust, provide an excellent instantiation of both models – we use the same trait impressions for the PCA model as the original authors, and the assumptions of FA include that observed variables (e.g., meanness, trustworthiness) are linear combinations of underlying factors, and therefore should be robust to slight differences in trait impressions. After creating these reduced spaces, we applied the mean shift algorithm to resulting values as before.

2.1.4. Sample size considerations

Clustering is an exploratory approach, and where we conduct tests for differences between clusters, our sample size is limited to the number of faces in the stimulus sets (here, n = 2,222). The sensitivities of our tests are reported where applicable.

2.2. Results

2.2.1. Full Dataset Clusters

Applying mean shift to the full 2,222 data points, with all traits z-score standardized, revealed two underlying clusters, *SC* = .44. Cluster zero contained 1771 (80%) faces and cluster one contained 451 faces. Before applying further analysis to these clusters, we also applied K-Means clustering - a computationally far simpler procedure - to the same data, by testing the *SC* across two-, three-, and four-cluster solutions. The highest K-Means *SC* was observed with two clusters (*SC* = .41), dropping off sharply thereafter (three clusters *SC* = .26, four clusters *SC* = .18). Cluster zero of the K-Means solution contained 1629 faces (73%) of the dataset. Ninety-three percent of faces were allocated to the same cluster by K-Means as mean shift, but K-Means allocated an additional 142 faces to cluster one.

Before examining the clusters in more detail, we tested for an association between mean-shift cluster placement and the sex of the face, to understand whether a simple sex difference may be associated with a difference in clusters in social perceptions. A chi-square test showed a statistically significant association, $\chi^2(1) = 145.81$, $p < .001$, $V = 0.25$. Given the sample size, the test is sensitive to effects as small as $V = 0.07$, with 95% power and an alpha level of .05. Out of the 451 assigned to cluster one, 371 were male. As such, we tested for the presence of cluster formations within each sex separately.

2.2.2. Sex Separated Clusters

For the female dataset (*n* = 955), a two-cluster solution (mean shift bandwidth = 3.8), produced the highest silhouette coefficient, *SC* = 0.48. Using K-Means as a robustness check also revealed two clusters produced the highest score (two clusters *SC* = 0.37, three *SC* = 0.21, four *SC* = 20). Mean shift assigned a small proportion of female faces to cluster one (*n* = 84, 9%), while K-Means assigned 224 (23%). Despite this, the overall agreement between the clustering solutions was relatively high, with 85% of faces being assigned to the same cluster by either approach.

For the male dataset (*n* = 1,267), a two-cluster solution also emerged (mean shift bandwidth = 2.9), producing the highest silhouette coefficient, *SC* = .42. A robustness check with K-Means also revealed two clusters as the optimal solution (two clusters *SC* = 0.39, three *SC* = 0.24, four *SC* = .19). Mean shift assigned a larger minority of male faces to cluster one (*n* = 331, 26%), while K-Means assigned 378 (30%). The agreement between clustering approaches was very high, with 97% of faces being placed in the same cluster.

2.2.3. Clusters in existing theoretical social perception space

Given the consistent emergence of two clusters across male and female spaces, we generated the two- and three-factor models (PCS and FAS, respectively) on the entire dataset. Applying mean shift to the two-dimensional valence/dominance model, or PCS, revealed again that two clusters provided the highest silhouette score, *SC* = .56. Applying K-Means also confirmed two clusters resulted in the highest SC (two clusters *SC* = .54, three *SC* = .42). Similarly, for the three-factor space, or FAS, two clusters also emerged after applying mean shift (*SC* = .48), which was also supported by K-Means (two clusters *SC* = .44, three *SC* = .30).

We next examined the consistency with which faces were clustered in the full dataset, and the respective two- and three-factor models. Consistency was high for the FAS, with

98% of faces being assigned to the same cluster in both the full and FAS representation of the data. For the PCS, this was 97%. A visual comparison of each clustering solution by mean shift, for all three representations of the data, is shown in Figure 1.

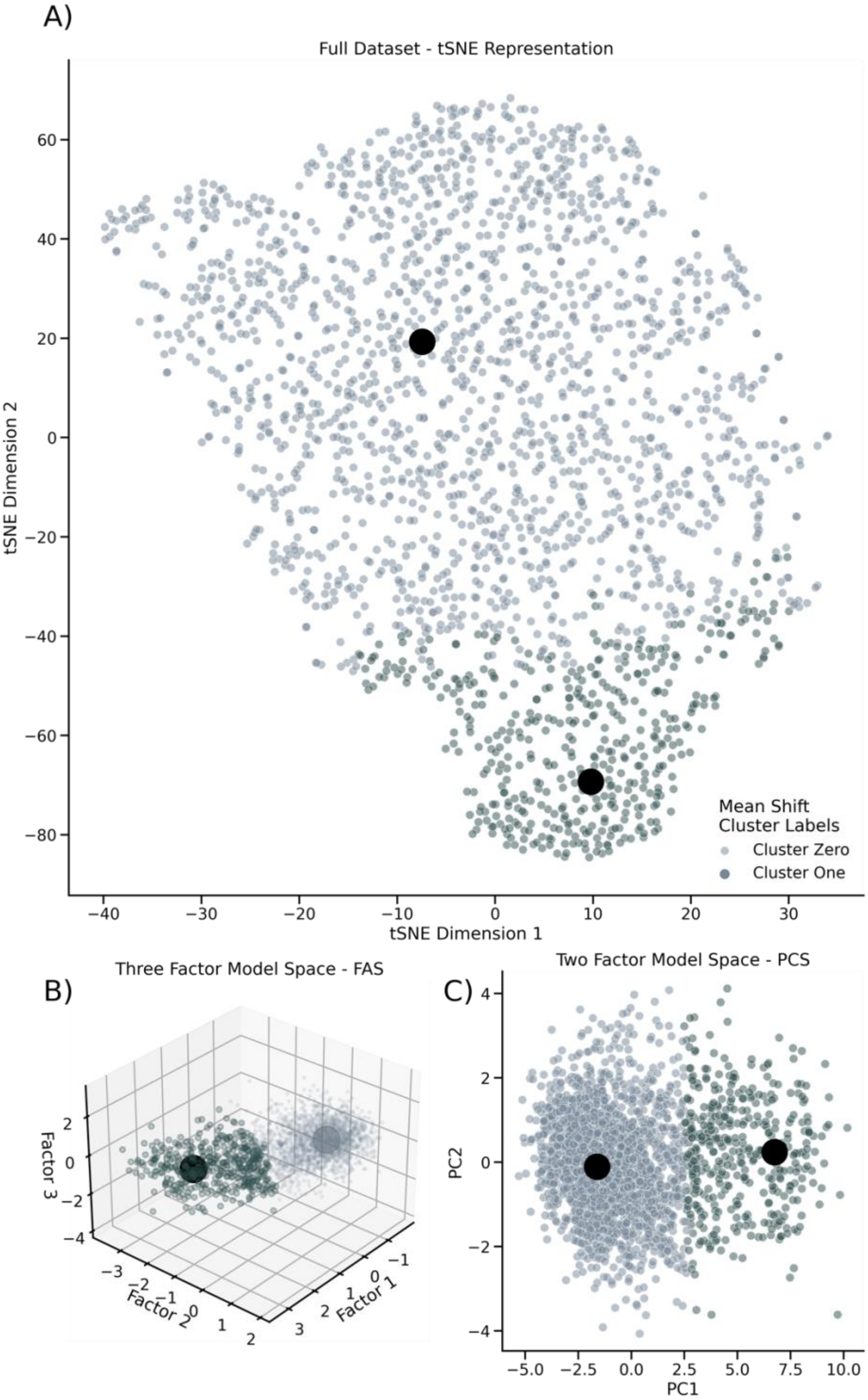

Figure 1. Summary of the cluster analysis resulting from the ratings included in the 10k US Adult Faces Database (Bainbridge et al., 2013). A) The 14-dimensional space reduced to two dimensions via t-SNE (t-

Distributed Stochastic Neighbour Embedding), coloured for each cluster. Black circles represent the mean shift derived centroids. B) Mean shift clustering on the three-factor model, and C) on the two-factor model. In both cases, mean shift demarcates almost entirely the same set of faces into cluster one. Black circles indicate the obtained centroids of the clusters with mean shift.

The confusion matrices for each clustering solution (comparing the full-dimensional clustering result to the FAS, and the full-dimensional clustering to the PCS) are shown in Table 1. To assess their similarities more fully, we computed a chi-square statistic and Cramer's V as an effect size. For the full dimensional and FAS clustering, the association was significant, $\chi^2(1) = 1849.46$, $p < .001$, $V = 0.91$, as was the case with the PCS clustering, $\chi^2(1) = 1927.10$, $p < .001$, V = 0.93. Given the sample size, the test is sensitive to effects as small as $V = 0.07$, with 95% power and an alpha level of .05.

Table 1. Confusion matrices of clustering solutions in reduced spaces when compared with the full space.

	FAS		PCS	
Full data	Cluster Zero	Cluster One	Cluster Zero	Cluster One
Cluster Zero	1761	10	1771	0
Cluster One	51	400	48	403

Note. FAS = factor analysis space; PCS = principal component space.

2.3. Discussion

We demonstrate the existence of two distinct clusters within a large set of trait impressions and faces. These clusters existed both within the full-dimensional space and within the representations of social impressions that have become central in this research area. Mean shift, an algorithm ideally suited to exploring high-dimensional spaces for groups of observations, allocated 20% of individuals into a separate and distinct cluster when dealing with the full-dimensional representation of the dataset, and faces assigned to cluster one were similarly assigned when mean shift was used on the two- and three-dimensional models of social trait space. Further analysis suggested that more male than female faces were placed into cluster one, which suggested that differences in perceptions of men versus women could explain the cluster solution. However, two clusters also reliably emerged when considering men and women separately, and maintained the general pattern seen within the full clustering solution. Using K-Means as an alternative, simpler clustering algorithm supported the initial results.

These initial results indicate that social perception of faces is ‘clumpy’. Even when social perceptions are projected down onto theoretical factor spaces, we find that these spaces are not occupied uniformly – there seems to exist two general clusters that faces can be readily assigned to, one which is far smaller than the other. While we have shown that there exists a male bias in this smaller cluster, we have yet to investigate the kinds of perceptions that demarcate each cluster. Before investigating this question, we use additional data to test for the presence of a two-cluster solution to confirm the replicability of our initial findings.

3. Experiment 2: Cluster stability using other sets of ratings

Although the evidence presented in Experiment 1 strongly supports the existence of two clusters that account for the ratings given to faces, we initially remain skeptical with regard to the stability of these clusters. A significant amount of research effort has been focussed on reproducing two- and three-factor solutions, with sometimes mixed results, as discussed previously (Jones et al., 2021; Lin et al., 2019; Sutherland et al., 2019). Our aim here was to test for the presence of a two-cluster solution in datasets of social trait ratings that contain different social trait impressions. If the cluster solution is robust, then it should emerge regardless of the variation in measured social traits and the number of these traits used.

3.1. Method

3.1.1. Stimuli

We attempted to recover cluster solutions similar those found in Experiment 1 using two additional face databases.

Sutherland et al. (2013). The dataset used to generate the three-factor model of social perception contains 1,000 ambient face images, similar to the 10k US Adult Faces Database, varying naturally in pose, expression, and lighting. The set consists of equal numbers of men and women but contains only Caucasian individuals. This set also provides social impressions for 13 traits, with these traits mostly differing from those used in the 10k US Adult Faces Database. As such, these represent a useful test of stability in clustering solutions. Faces in this set are rated for *age, approachability, attractiveness, intelligence, babyfacedness, smiling, health, skin tone, trustworthiness, confidence, aggressiveness, dominance,* and *degree of sexual dimorphism*. Ratings of these traits were averaged across 50 participant raters, providing a rating of each trait per face.

Lin et al. (2019). The dataset previously used to support a four-factor model of social perception comprises 100 front-on, neutral face photographs of Caucasian individuals, with even numbers of men and women. Faces were selected from a wider pool of individuals by maximising differences in Euclidean distances of face vectors encoded by the Dlib (King, 2009) neural network. Thus, while limited to constrained images, these faces provide a great range of appearance within Caucasian individuals. These faces were rated for 100 traits that were generated from an agglomerative clustering solution of a set of descriptors (e.g., friendly, kind) embedded into a word vector generated by a neural network. Ratings included traits like *helpful, prudish, mature, independent, nosey*, etc. The full range of descriptors are available in the original study. Ratings of these traits were averaged across 1,500 participant raters, providing a rating of each trait per face. This set provides an additional, stringent test of the clusters we identified earlier as the feature space is an order of magnitude larger, and perceptions are made of neutral faces where no expression is visible.

As with our initial clustering analysis (Experiment 1), we applied mean shift to each of these *z*-score standardised datasets, configuring kernel radius before extracting clusters.

3.1.2. Sample size considerations

Clustering is an exploratory approach, and where we conduct tests for differences between clusters, our sample size is limited to the number of faces in the stimulus sets (here, n = 1,000 and n = 100). The sensitivities of our tests are reported where applicable.

3.2. Results

3.2.1. Sutherland et al. (2013)

For this dataset, mean shift (radius = 3.0) produced two clusters, although with a somewhat lower silhouette score than was observed in Experiment 1 with the 10k Face Database; $SC = 0.24$. As before, using K-Means, a simpler clustering technique, also produced two clusters as the best fit ($SC = .24$), confirming the result.

Cluster one contained 326 (33%) faces, indicating again the presence of a smaller cluster, but a somewhat larger minority than shown in Experiment 1. Testing for an association between cluster labels (zero or one) and the sex of the face again revealed a significant association, $\chi^2(1) = 1849.46$, $p < .001$, $V = 0.27$. Given the sample size of 1,000, the test is sensitive to effects as small as *V* = 0.11, with 95% power and an alpha level of .05. Of the 326 faces in cluster one, 227 (69%) were male. As before, we examined for the presence of the two clusters within each sex, separately.

For women (*n* = 500), mean shift (radius = 2.6) produced two clusters (*SC* = .26), which was also supported by K-Means (two clusters *SC* = .24, three *SC* = .19, four *SC* = .19). Cluster one was again smaller than cluster zero, with 116 faces (23%) allocated to it, which was a significantly larger proportion than that seen in Experiment 1 when splitting the dataset by sex.

For men (*n* = 500), mean shift (radius = 2.48) also produced two clusters (*SC* = .24), which was also supported by K-Means (two clusters *SC* = .23, three *SC* = .21, four *SC* = .19). Cluster one was also smaller here, with 145 faces (29%) allocated to it, very similar to what was observed in Experiment 1.

3.2.2. Lin et al. (2013)

For this dataset, mean shift (radius = 8.9) produced two clusters, with a similar silhouette score to the Sutherland et al. dataset, *SC* = 0.23. K-Means also produced two clusters as the best fit (two clusters, *SC* = .25, three clusters *SC* = .22).

Cluster one contained 19 (19%) faces, indicating again the presence of a smaller cluster. Testing for an association between cluster labels (zero or one) and the sex of the face here revealed no significant association, $\chi^2(1) = 0.26$, *p* = .610, *V* = 0.05. With a sample size of 100, the test is sensitive only to effects as small as *V* = 0.11, with 95% power and alpha of 0.05, indicating this test is likely underpowered. Of the 19 faces that were in cluster one, 11 were male.

3.3. Discussion

Across two additional datasets that previously proposed influential models of person perception, cluster analyses consistently revealed two clusters. The data of Sutherland et al. (2013) showed a similar pattern to that of the 10k Adult Faces Database in that a smaller cluster of faces was apparent, and within this cluster was a higher number of male faces. However, two clusters also emerged within both female and male ratings only. The data of Lin et al (2019) showed a similar, smaller cluster, though no preponderance of male faces was observed. Taken together, these results support those of Experiment 1 – there exists ‘clumps’ of faces in social perception which can be classified into two clusters, and these clusters seem to emerge reliably despite variations in the number and kind of social traits that are used to generate the solutions.

Of note here was the noticeably weaker cluster solutions than those observed in Experiment 1, with silhouette coefficients around half as large as initially shown. This may be due to two factors – the data of Sutherland et al. (Sutherland et al., 2013) contain only Caucasian faces, while the 10k Faces set contains a multitude of ethnicities. This might suggest that ethnicity plays a role in the formations of these clusters. For the data of Lin et al., which did indeed contain a multitude of ethnicities, the weaker association may be due to the size of the dataset being ill-suited to cluster analysis, containing only 100 observations. Guidelines for cluster analysis propose the use of 70 times the number of samples to variables (Dolnicar et al., 2013). Second, the dataset provided a fundamentally different sort of perception as compared to that of Sutherland et al. and the 10k Adult Faces Database, in that the perceptions were collected for standardised, neutral face photographs, which others have argued is an artificial representation (Satchell, 2019) – whether clusters of perceptions exist in these kinds of scenarios is therefore unclear. To summarise, the cluster solutions evident in the data appear to be fairly robust – there seems to exist a pair of clusters in social perception space (in either its full or compressed theoretical representations), with one cluster being significantly smaller than the other.

4. Experiment 3: Understanding the perceptual characteristics of cluster profiles

Broadly, Experiments 1 and 2 indicate a ‘clumpiness’ of social perception from faces, revealing two distinct clusters that are apparent across varying datasets, theoretical component representations of the data, and choice of clustering algorithms. While these clusters seem to emerge reliably, it is unclear which perceptual profiles give rise to them – that is, which combinations of perceptions are more likely to result in a face being clustered into the larger or smaller category? A primary aim of the following experiment is to estimate the kinds of patterns in ratings that result in placement in cluster zero or one.

It is also unclear how the two clusters may sit alongside existing two- and three-factor models of social perception. Existing models suggest that an overall impression of an individual can be determined by identifying their position along each of the factors, and the results of Experiment 1 demonstrated that even in these models, the clusters were apparent. However, it is difficult to compare component-based models (Oosterhof & Todorov, 2008; Sutherland et al., 2013) with a cluster solution, and thus a direct comparison with existing models is untenable.

There are two objectives in this experiment. The first is to identify a profile of the impressions that determine the assignment of a face to cluster zero versus one, to uncover how faces are generally perceived within each cluster. To do this, we use a Bayesian logistic regression with a train and test split of the data, using highly restrictive priors to guard against overfitting. The second aim is to describe a ‘cluster axis’, which is a component that separates each cluster, and can thus be used to contrast with existing two- and three-factor solutions, which we estimate using linear discriminant analysis. For the following analysis, we focused on the use of the 10k Adult Faces Database, due to the sample size and variations in expression, ethnicity, and gender.

4.1. Method

4.1.1. Estimating cluster profiles – Bayesian logistic regression

To estimate the different cluster impression profiles, we used Bayesian inference to fit a logistic regression that classified faces into cluster zero or one based upon the trait ratings assigned to each face, as well as whether the face is male or female. Our reason for using Bayesian inference is to guard against circular analysis (Kriegeskorte et al., 2009) – for a model to classify clusters (and thus provide theoretically important information about what discriminates them), the clusters must be generated from a dataset, and that same dataset must necessarily be used to train the classifier, with a consequence being that results will appear much stronger than they actually are. Bayesian methods can reduce this problem by incorporating extremely skeptical priors, which force any estimated coefficients down towards zero – more simply, using Bayesian approaches, it is possible to make the differences more difficult to estimate. This is entirely analogous to regularisation methods that prevent overfitting in frequentist models, such as ridge regression (Marquardt & Snee, 1975).

We also take additional steps to guard against overstating any effects. Our analysis pipeline proceeds as follows. First, we randomly split the 10k Face Database into a training (comprising 75% of the data, *n* = 1,666 observations) and testing set. Both sets were *z*-score standardised separately. The Bayesian logistic regression was fitted to the training data using strongly regularised priors for each of the fourteen coefficients (one categorical sex predictor, and thirteen trait ratings), using a normal distribution with scale zero and a standard deviation of 0.05, ~ N(0, 0.05), with the aim of separating the mean shift assigned cluster label. A standard normal prior was also used for the intercept, ~ N(0, 1), and a Bernoulli distribution was used for the likelihood. As logistic regression coefficients are estimated on the log-odds scale, our priors reflect that the odds of assignment to cluster zero or one are most probable between 0.95 and 1.05; close to no effect. Once the model was fitted, we used it to predict the test score cluster labels, assessing its accuracy (the number of cluster one and cluster zero labels predicted correctly, divided by *n*), and recall (the number of cluster one faces assigned as cluster one, out of the total number of cluster one faces in the test data) to evaluate its performance.

In making decisions about the importance of parameters, we set a null region of the posterior distribution of the odds as between 0.95 and 1.05 (Kruschke, 2018). Posterior 95% highest-density intervals (which represent a 95% probability that the true effect is within the given bounds) overlapping or falling within this region we consider having insufficient evidence to determine assignment into cluster zero or one. Models were estimated using the PyMC3 package with Markov Chain Monte Carlo methods (Salvatier et al., 2016).

Statistical power is appropriate within frequentist paradigms, representing the probability of correctly rejecting a null hypothesis when an alternative is true. The Bayesian analysis conducted here focuses on the 95% highest density interval of parameter coefficients as a way to interpret the coefficients and whether they have any influence on belonging to a given cluster. As we use very strict priors, our model is conservative – any effects must necessarily be present to overwhelm the prior – that is, we are guarding against a false positive conclusion. In addition, we also focus on a null *region* rather than a point-null hypothesis, which is again more conservative, as the coefficients must be outside a region considered to be null and not just a point value (Kruschke, 2018; Kruschke & Liddell, 2018).

Moreover, our focus is not on general hypothesis testing here, but on fitting a model that is generalisable. We employ cross validation for testing this – whether our restricted, fitted model is able to predict new data that is unseen.

4.1.2. Comparing cluster solutions with existing models – Linear Discriminant Analysis

To draw comparisons between the cluster profiles and two- and three-factor models, it is necessary to generate a vector that describes the differences between the clusters in terms of their perceptual profiles, and project each face's ratings on this vector. This operation allows instantiations of the two- and three-factor models to be compared in a common geometric space, where each component is represented by a vector with *n* elements, where each *n* is a face projected onto the component (through PCA or FA).

To do this with the cluster solution generated by mean shift in the previous sections, we employed linear discriminant analysis (LDA). LDA, while historically a statistical classification technique (Izenman, 2008) has the desirable property of producing axes that maximise the separation between the classes it discriminates, upon which it is possible to project data points. As such, by using the perceptual ratings of each face as predictors, and the assigned cluster label as the target, an LDA axis can be generated that sits in the common geometric space with the PCA and FA components by projecting the perceptual ratings for each face onto the LDA axis, in the same way that the perceptual ratings are projected onto PCA or FA solutions.

Analysing the relationships amongst these components - produced by several different attempts to model social face perception - can thus be achieved by computing the Euclidean distances between the resulting axes and using Pearson correlations – or equivalently cosine similarity (CS°) – which both represent the angle between vectors. These metrics allow for an intuitive comparison of the clustering solution with existing models in simple terms of how far away the cluster axis lies from two- or three-factor model components, and how closely aligned it is to those components. The power for these statistical tests is limited by the sample size (here *n* = 2,222). For these correlations, the smallest effect we can detect reliably is *r* = .07, with power at 95% and alpha of .05.

Here, we took the full set of 2,222 images and used LDA to predict the mean shift assigned cluster labels from the perceptions of all 14 traits, recovering the axis of maximal separation between the clusters. After projecting each face's trait ratings onto it, the axis was then compared with the two principal components representing the two-factor model (Oosterhof & Todorov, 2008), and the three factors from FA representing the three-factor solution (Sutherland et al., 2013), both of which were derived in Experiment 1 on the full data set.

4.2. Results

4.2.1. Estimating cluster profiles – Bayesian logistic regression

The coefficients from the Bayesian logistic regression estimated on the training data were exponentiated and converted to odds, and are displayed in Figure 2. Several predictors – whether the face was male or female, and ratings of the traits *boring*, *confident,* and *attractive* overlapped with the null region, indicating insufficient evidence of their contribution to assignment to cluster one or zero. However, other predictors showed clearer effects. A one standard deviation increase in ratings of *caring, sociable,* and *trustworthy* were associated with lower odds of a face being assigned to cluster one by around 20%, while the same increase in in ratings of *aggressive, mean,* and *unhappy* were associated with an increase in odds of around 26% of belonging to cluster one.

When predicting the cluster labels of the test data, the model had good accuracy (labelling cluster one faces as one, and cluster zero faces as zero), *M* = 0.88, 95% HDI [0.85, 0.89]. Recall (the number of cluster one faces identified correctly out of all cluster one faces) was somewhat less, with the model identifying around 71% of all cluster one faces correctly, *M* = 0.71, [0.62, 0.78]. The receiver operating characteristic area under curve (ROC-AUC) showed very clear separation between clusters, *M* = 0.999, [0.998, 1].

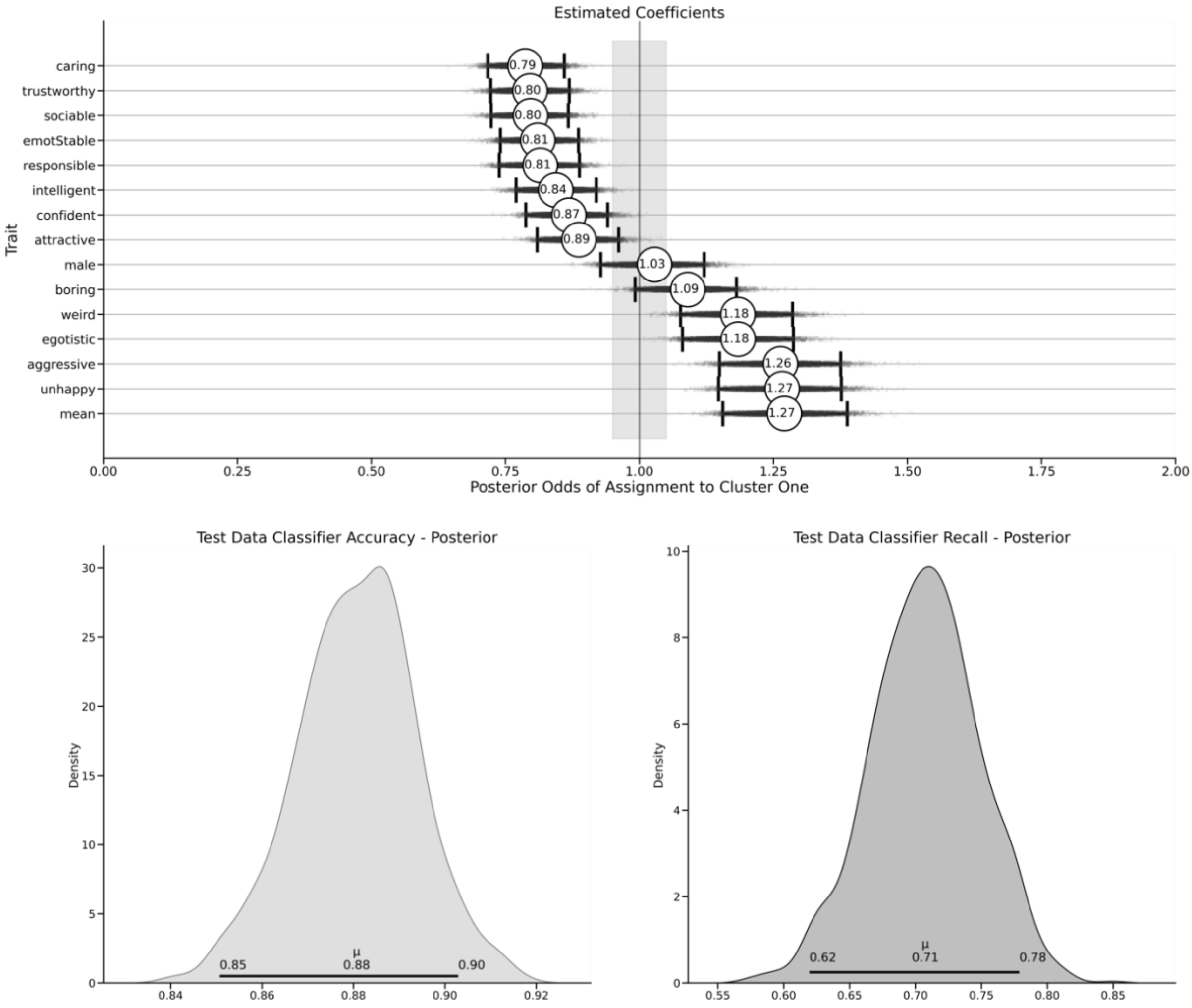

Figure 2. Top – posterior distributions of the coefficients of the Bayesian logistic regression. White circles represent the mean of the posterior in odds, while the error bars depict the limits of the 95% highest posterior density intervals. The shaded central region represents odds of 0.95 to 1.05, which defines our null effect region – posteriors overlapping this have inconclusive evidence of an effect. Bottom – the distributions of accuracy and recall of the fitted model on the test data, the clusters of which were generated independently of the training clusters.

4.2.2. Comparing cluster solutions with existing models – Linear Discriminant Analysis

LDA was used to maximally separate the mean shift assigned labels from Experiment 1 to the full dataset. This generated an axis that separated each cluster from one another, and that we were able to project the trait ratings of each face onto in order to form a 'cluster axis'. This allowed for the comparison of the cluster solution with existing two- and three-factor solutions by way of bivariate correlations, angle measures, and Euclidean distances between the vectors. The results of this analysis are shown in Figure 3.

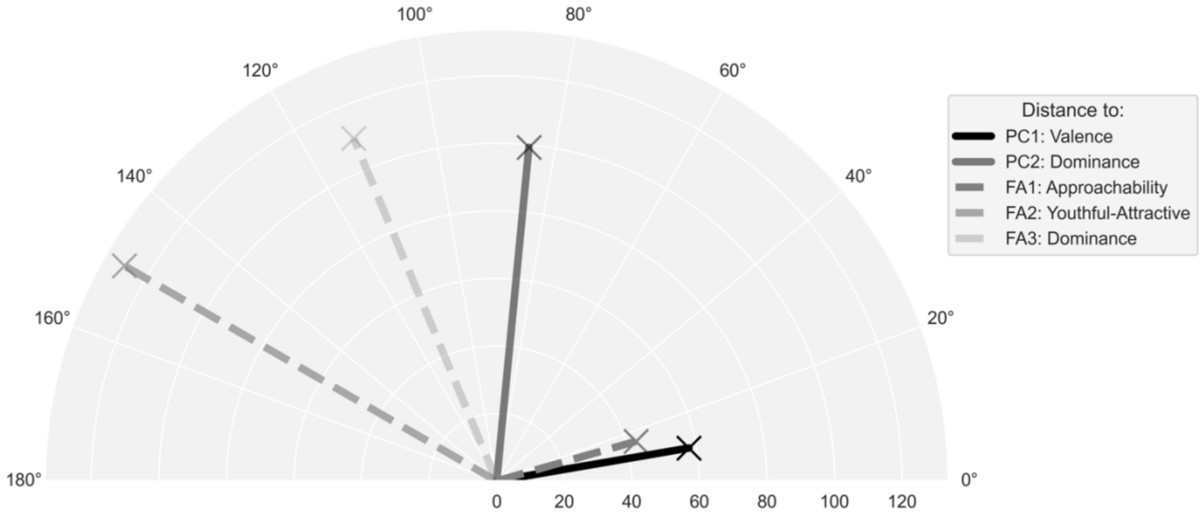

Figure 3. Polar coordinate representation of the relationship between the cluster axis derived with LDA (centre pole) and existing model factors. Angles are represented in degrees, and distance (the length of the vector between the cluster axis and another) is given in Euclidean distance.

We first compared the cluster axis to the two PCs of the two-factor model. The cluster axis was strongly correlated with PC1, *r*(2220) = 0.99, with a very small cosine similarity, CS^o = 9.74, and was close in Euclidean space, distance = 57.74. However, for PC2, the relationship was close to orthogonal, *r*(2220) = 0.10, CS^o = 84.49, and the vector was further away, distance = 99.29.

For the three factor model, the cluster axis showed a strong relationship with the first factor, *r*(2220) = 0.96, CS° = 15.90, and this factor was the closest in Euclidean space to the cluster axis, distance = 42.88. The cluster axis showed a strong negative relationship with factor two (the vectors point in opposite directions), *r*(2200) = -0.87, CS° = 149.95, and was furthest away from this vector, distance = 127.37. The cluster axis also showed a negative relationship with factor three, *r*(2220) = -0.39, CS° = 112.65, distance = 109.95.

4.3. Discussion

Taken together, the results so far suggest that multidimensional impressions of individuals give rise to essentially two stable and separable clusters. Inspection of the odds of assignment to either category from the classifier model points towards a kind of approach-avoid continuum - higher levels of impressions such as meanness and aggressiveness (e.g., avoid) are associated with greater probabilities of belonging to cluster one, and higher levels of caring and sociability impressions (approach) are associated with cluster zero. We took a stringent approach that used heavily restricted priors to avoid overstating any effects, and while some traits had insufficient evidence of contributing to the separation, those with the largest effects differed strongly in their valence. Moreover, the model did reasonably well in identifying out-of-sample clusters. While less restricted priors may have given better performance, fitting models to the same dataset that clusters were generated from should be approached with caution. However, the results indicate separation between clusters appears to be a property of a general positive to negative impression, and that some traits hold more influence than others, as indicated by the posterior distributions of the coefficients.

The difference between these clusters, as derived using an LDA and vector projection techniques, showed similarities to existing model components. The cluster axis was geometrically closest, and most strongly aligned, to the first components of the two- and three-factor models. Both of these components index a form of positive expression, and indeed are characterised as *valence* and *approachability* in the original studies. Interestingly, the cluster axis was distant from and pointed in the opposite direction to the *youthful-attractive* component of the three-factor model, indicating a negative correlation. Conceptually, this means that faces lower on the cluster axis are generally higher on the *youthful-attractiveness* component, and the reverse, reflecting the low probability of attractiveness perceptions resulting in a cluster one assignment. Finally, the cluster axis seems relatively weakly related to *dominance* components. The PCA *dominance* component was almost orthogonal, and the FA *dominance* component was similarly distant but showed a negative relationship. This is a significant theoretical advance as it indicates that perceptual profiles of a face result generally in two clusters, which are best separated by valence-like information. That the cluster axis – derived by considering how faces are grouped, and not how trait perceptions are grouped – shows weak relationships with dominance may speak to differences in how impression formation functions practically. At least in the context within which faces were rated here, valence seems a vital component of impression formation, with dominance only weakly so.

One issue with the current work is that the data used to generate the clusters were also used to classify them. This may explain the very clear separation achieved by the classifier on test data, with an ROC-AUC of close to one. A second issue is that the clusters may only be meaningful in the current set of trait impressions, though they emerge in other datasets. While trait impressions are generally consistent between raters (Sutherland et al., 2019), they arise from the physical appearance of individuals (Todorov et al., 2013). In the following experiment, we test the reliability of the identified clusters by using data other than trait impressions to separate them – namely, parameters representing facial appearance itself.

5. Experiment 4: Predicting cluster membership from facial appearance

Existing two- and three-factor models of social perception (Oosterhof & Todorov, 2008; Sutherland et al., 2013) allow for the recovery of the facial appearances associated with scoring high or low on the generated axes. For example, Sutherland and colleagues (Sutherland et al., 2013) illustrated the facial appearances representing high and low scorers at the extreme ends of their three factors. Approachable (factor one) faces had prominent smiles rather than more neutral expressions, youthful-attractive (factor two) faces appeared more feminine and attractive versus more elderly, and more dominant (factor three) faces appeared more masculine and older (but not necessarily aggressive). In contrast, the two-factor model described by Oosterhof and Todorov (2008) showed that faces scoring high and low on factor one appear smiling and angry respectively, and those high and low on factor two appear masculine versus babyfaced.

Thus far, we have demonstrated that clusters of individuals exist who receive a particular combination of impressions, with faces in cluster one receiving generally negative social evaluations compared to cluster zero. This pattern also emerged when the trait space was reduced to its component measures, indicating that the realisation of these models – how this more compact representation of social evaluations translates to perceptions of individuals – can be summarised as two groups of appearances.

Next, we seek to extract the facial appearances associated with these clusters. Rather than simply averaging or transforming appearances between faces in each cluster, we take a more stringent approach. Given only parameters describing facial shape or texture, can a statistical model separate faces into their assigned clusters? This represents a stricter and more general test of the reliability of our cluster solutions, in that if clusters do not encode meaningful differences in facial appearance then it is likely they are statistical artefacts of the trait impressions used. However, if they are separable based solely on appearance, then it is possible for statistical models, given a representation of a face in a photograph, to assign cluster labels without the need for a reiteration of the initial clustering algorithms. That is, the clustering solution can be represented beyond the current dataset without the need for face ratings or the application of further clustering algorithms. Indeed, existing work demonstrates that facial first impressions can be computationally modelled from shape features (Vernon et al., 2014), suggesting that statistical regularities in impression-appearance relationships can be successfully computationally extracted from face images.

5.1. Method

5.1.1. Stimuli

We continued with the 10k US Adult Faces Database due to the sample size and direct availability of facial photographs. For each of the 2,222 faces, a set of 77 landmarks placed manually by two annotators (Bainbridge et al., 2013; Khosla et al., 2013) describing the outer face shape, eyebrows, eyes, nose, and mouth were available, and these were used as face shape parameters. These *xy*-coordinates were converted to a vector of 154 values. The facial photographs themselves (256 x 256 pixels, each with RGB values) were used as facial texture parameters, extracted by converting the colour image arrays (producing a vector of 196,608 pixels per image). Facial textures were warped to the average facial shape before being analysed, ensuring each pixel represented the same position across faces (Kramer et al., 2017).

We conducted an additional analysis alongside the use of these more traditional representations of faces by using a deep convolutional neural network (DCNN) to embed each face into a 128-dimensional vector. We used the Python implementation of the OpenFace face recognition model (Amos et al., 2016) based on the FaceNet model (Schroff et al., 2015). These models are trained using triplet loss to optimise a representation of face identity into 128 dimensions (unlike other network models used in psychological research; Wang & Kosinski, 2018). These vectors can be thought of as an identity-optimised representation of faces that is more compact than the landmark and texture representations, which have to undergo PCA before being analysed. The drawback of this DCNN approach is that, despite its efficiency in representing faces and thus generalisability to new faces, the embeddings themselves are largely uninterpretable (O'Toole et al., 2018). However, we use these embeddings as a representation of appearance in an attempt to model cluster appearance in a generalisable way. The DCNN failed to encode the appearances of 15 fifteen faces, and as such the analysis proceeded for that data on the remaining 2207 images.

As in our earlier experiments, we used the cluster labels derived from mean shift in Experiment 1 as the dependent variable.

5.1.2. Analytical Strategy - Bayesian Logistic Regression

We took a similar approach to Experiment 3, focusing on the use of Bayesian logistic regression models tested on out-of-sample data to assess how well the clusters could be separated. The raw shape coordinates, pixel values, and DCNN representations were all split into a training and test set (75% for each dataset, *n* = 1666, 1655 for the shape and texture data, and DCNN representations, respectively). The shape and texture data were subject to scaling and a PCA that retained 95% of the variance, resulting in 13 components for shape, and 364 components for texture. For the DCNN representation, the full 128 embeddings were used as predictors, and each was scaled before fitting. The test data for shape, texture, and DCNN representations were scaled by the means and standard deviation of the training data, and the shape and texture data were then projected onto the principal components estimated from the training data.

Each logistic regression had Normal priors on each of the coefficients with mean zero and standard deviation of one, introducing some regularisation to the estimates. All models were evaluated using ROC-AUC, accuracy, and recall metrics on the test data. Finally, we estimated an intercept-only model on the training data, and used it to predict the test data, to serve as a baseline comparison.

For this model, we again focused on the out-of-sample predictive ability, measured using ROC. We do not seek to directly interpret the coefficients of these models, since any one coefficient is difficult to interpret, as they correspond to a principal component or DCNN embedding. Our priors introduce some regularisation which will necessarily make the differences in appearances between clusters more difficult to discriminate. More generally, the use of three different representations of faces to test whether cluster differences can be extracted lends robustness to our conclusions.

We also sought to extract the general appearances of faces that were associated with higher or lower probability of belonging to cluster one. To test this, we took the mean posterior probability predicted by the logistic regression for each face, and regressed it against the raw shape and texture data in a multivariate regression (i.e., shape and texture parameters were the dependent outcome measure and the probability was the sole predictor), which reveals a linear function between probability of cluster one assignment and appearance.

5.2. Results

5.2.1. Cluster separation by the classifiers

For each classifier – trained on shape or texture PCs, or DCNN representations of faces, the test data ROC-AUC scores were generally high, indicating the ability of the classifier to separate the clusters based solely on facial appearance information - shape model *M* = 0.914, 95% HDI [0.907, 0.921], texture model, *M* = 0.868, [0.854, 0.884], DCNN model, *M* = 0.769, [0.743, 0.792]. The ROC curves for each model are shown in Figure 4; the ROC-AUC score was 0.5 for the intercept only model.

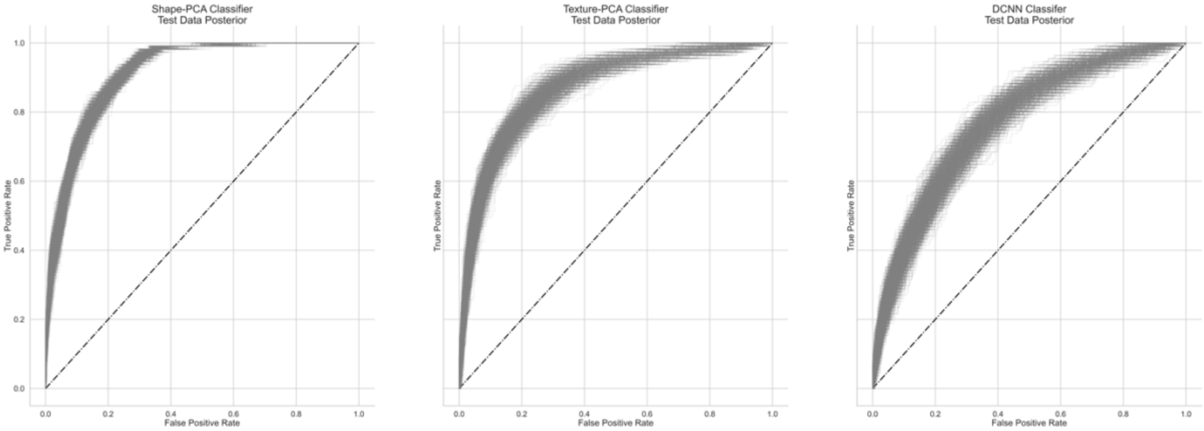

Figure 4. Receiver operating characteristic (ROC) curve posterior distributions for each of the three models, on out-of-sample data.

For accuracy, the results were also high for each model – shape model *M* = 0.80, [0.78, 0.83], texture model *M* = 0.84, [0.83, 0.86], DCNN model, *M* = 0.75, [0.72, 0.78], indicating the classifiers placed out-of-sample data into the correct cluster generally well. In comparison, the intercept only model had an accuracy score of *M* = 0.67, [0.64, 0.71].

However, for recall (the number of cluster one faces identified as cluster one correctly), there were differences – shape model *M* = 0.57, [0.48, 0.65], texture model *M* = 0.63, [0.58, 0.69], DCNN model, *M* = 0.37, [0.29, 0.45]. The intercept only classifier had a lower score, *M* = 0.20 [0.12, 0.27].

These posterior predictions and their distributions are easily interpreted when compared to the intercept only model. Where the highest density intervals do not overlap, there is evidence the classifier is performing better than the base-rate. In some instances, there is close overlap, such as for the DCNN accuracy and recall scores. It is possible to compute the probability that the DCNN classifier performs better than the intercept only model *at all* by computing the differences between the DCNN accuracy and recall posteriors and the intercept only model, and calculating the probability that the difference is greater than zero. For both accuracy and recall, this probability was 99%.

5.2.2. Extracting facial features

Statistical representations of face data, either principal components or neural network embeddings, are enough to allow a linear separation between clusters. Deriving the appearances associated with higher or lower probabilities is useful as a way to further understand the clusters. Here, we passed the mean of the predicted posterior probabilities of the shape-based classifier's test data to a multivariate regression, as the sole predictor variable (a regression with multiple dependent variables, here, shape coordinates and pixel values of images), modelling facial shape and texture simultaneously. That is, appearance was directly predicted from the probability of belonging to cluster one We used the shape classifier for computational simplicity and showed the highest ROC-AUC score. We used this model to predict appearance across the range of the probability of belonging to cluster one, allowing us to statistically recover the appearances associated with greater probability of classification to cluster one. The results are shown in Figure 5.

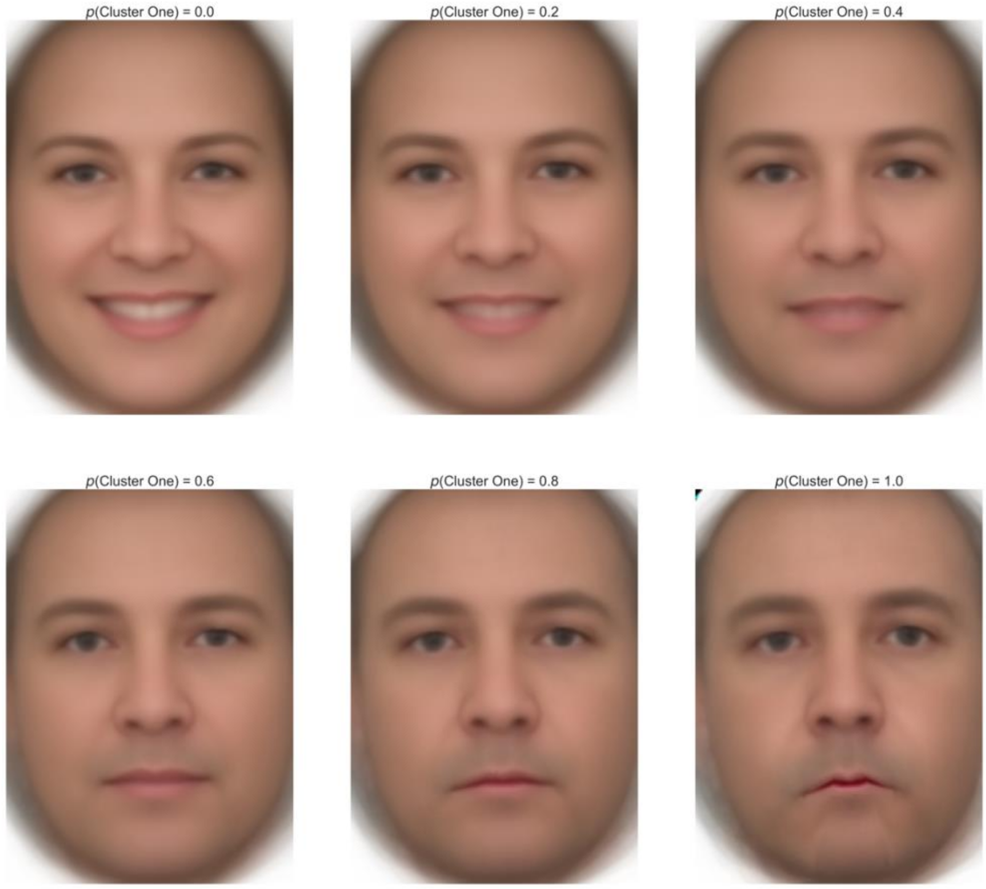

Figure 5. The predicted appearance of faces recovered from a multivariate regression. The sole predictor was the probability that a face belonged to cluster one, taken from the mean of the posterior predictions of test data for the shape classifier. The different levels of probabilities show the different appearances that are associated with higher probability of being in cluster one. These appearances were fed to the shape classifier as principal components of shape coordinates.

5.3. Discussion

The analyses here provide further evidence of the cluster labels derived from trait ratings. Using the full range of information from facial appearance in the form of shape and texture parameters, and neural network facial identity embeddings, a Bayesian classifier was able to discriminate reasonably well between clusters from just facial appearance. Using multivariate regression with the predicted cluster probabilities effectively visualised the appearances associated with cluster placement – faces appearing relatively more unhappy, aggressive, masculine, and somewhat older, compared with more feminine faces displaying a clear positive facial expression.

The differences between the classifier metrics were generally small but showed distinct patterns. Overall ROC-AUC scores were highest for the shape-data classifier and lowest for the DCNN classifier. For accuracy and recall, the DCNN classifier had poorer performance than the shape and texture data classifiers (which had overlaps in their performance). The explanation for this lower classification performance may lie in the optimisation of network embeddings for representing identity, as opposed to other facial appearance parameters. Conversely, the use of PCA on shape and texture inputs is able to recover important statistics that vary between images that are associated with the perceptions that lead to different cluster assignments, such as smiling, which is arguably less important for identity recognition – indeed, accurate recognition should be invariant to expression. The trade-off between a reduced classifier accuracy using network embeddings versus traditional shape and texture information with PCA is its generalisability – the latter approaches require standardisation, landmarking and image warping, and costly computation of matrix inversions for dimension reduction. Conversely, pre-trained network models are becoming more widely available and can embed images rapidly without recourse to image alignment or landmarking.

These findings also support those of the classifier from Experiment 3, whereby higher perceived ratings of traits such as *aggressive* and *meanness* were associated with greater odds of belonging to cluster one. These results also sit clearly alongside existing two- and three-factor models of social perception (Oosterhof & Todorov, 2008; Sutherland et al., 2013). In those models, one factor represents valence (the first component in PCA-based models, and factor one in FA models) and another represents dominance or aggressiveness (component/factor two). By considering how individuals are judged along a combination of traits, as opposed to reducing the number of traits to as few as possible, cluster analysis extracts a kind of aggregate of the axes in factor models as realised for individual faces – cluster zero faces appear more feminine, approachable, and somewhat younger than the faces in cluster one, merging features associated with separate axes in existing models.

6. Experiment 5: Behavioural responses to faces in different clusters

The analyses thus far indicate the presence of two clusters of appearances in social perception which reliably emerge from full sets of ratings, as well as reduced theoretically relevant representations, and can be extracted solely from facial appearance. The clusters indicate a kind of basic approach-avoidance continuum, and indeed are similar to the valence component of existing models. In this section, we test an important further issue, aiming to understand whether independent observers react to faces from the different clusters in the ways indicated by the data-driven methods used above. Here, we employ two different behavioural paradigms to investigate observer decisions of approach and avoidance for faces in cluster zero and one, with the prediction that observers will be more likely to avoid faces in cluster one, as well as the converse.

6.1. Method

6.1.1. Experimental tasks

We conducted two experimental tasks concurrently to assess responses to the faces. The first was a slightly modified version of the rapid display task used by Willis and Todorov (2006). A face was presented for 100 ms, and participants were required to make a decision regarding whether to approach or avoid the person, which we label onwards as the ‘brief display paradigm’. The second task was previously used by Kramer et al. (2020). Pairs of faces were presented on screen and participants were simply instructed to select one of the faces. No further instructions were provided. When participants selected one of the faces, that face increased in size, while the other decreased. We label this as the ‘preference paradigm’.

6.1.2. Participants

Fifty-nine Amazon MTurk workers (20 females, age *M* = 40.27 years, *SD* = 13.34 years) completed the brief display paradigm. Seventy-three percent of the sample self-reported as White, 19% Black, 5% Hispanic, and 3% Asian. Fifty-seven different MTurk workers (27 females, age *M* = 37.61 years, *SD* = 11.04 years) completed the preference paradigm, with 68% self-reporting as White, 12% Black, 11% Hispanic, 5% Asian, 2% American Indian, and 2% identifying as “other”. Worker location was restricted to Australia, Canada, New Zealand, the UK, and the USA, in order to target English speakers. All workers were paid $0.60 for participation. An initial sample of 163 workers were recruited, but 34 participants were dropped from the brief display paradigm and 13 from the preference paradigm due to failing attention checks during the tasks (see below). Participants provided informed consent prior to taking part, and received an online debriefing upon completion, in line with the university’s ethics protocol. The university’s ethics committee approved this work, which was carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

6.1.3. Stimuli and paradigm details

We selected 50 faces from each cluster of the 10k US Adult Faces Database to serve as stimuli for these experiments. By taking the centroids computed via mean shift in Experiment 1, derived from the full ratings, we computed the Euclidean distance of each face from its respective cluster centroid, and took the 50 faces with the smallest distance (i.e., those closest to the centroid for that cluster) as stimuli.

For the brief display paradigm, two versions of the task were created, where the response options (e.g., press ‘Z’ to approach, press ‘M’ to avoid) were switched, to avoid order effects. For the preference paradigm, four versions were created in which pairings were randomly designated, such that faces from cluster one and zero were paired at random and were displayed to the left and right of the centre of the screen.

We included attention checks to test whether participants were continuing to attend throughout the task (Hauser & Schwarz, 2016), replacing any participants who failed one or more of these. In the brief display paradigm, two additional faces were included in which the internal features had been replaced with the words ‘attention check’. In all other aspects, these faces/trials mirrored the rest of the test. However, at the time of responding, the typical instructions were replaced with ‘Attention check: Press Z’ (or ‘M’ for the second check) on both sides of the screen. For the preference paradigm, two attention checks were also included. In each, a pair of faces appeared in which their internal features had been replaced with the words ‘Attention check: pick this face’ and ‘Attention check: pick the other face’. The left/right position of these two faces was reversed for the second attention check. For each experiment, participants had to pass both attention checks in order for their data to be retained (and to receive payment).

Data were collected for both tasks using the Gorilla online testing platform (Anwyl-Irvine et al., 2020).

6.1.4. Procedure

Participants followed a link to the studies and were allocated to either the brief display or preference paradigms in a counterbalanced order. Within those studies, participants were counterbalanced across the different versions, as discussed earlier.

Brief display paradigm. Participants were instructed that they would view a series of 100 faces and would be required to make a decision about whether they would approach or avoid that person (by pressing the Z and M keys to indicate a response). Faces were presented on screen for 100 ms and participants were shown their response options (‘Z/M’ to avoid or approach), with these labels remaining on screen until they made a response. Responses were recorded so that an ‘avoid’ decision was coded as the positive class.

Preference paradigm. Participants were instructed that they would see 50 pairs of faces (individuals from cluster zero and one, paired together as described above) and were asked simply to select a face by clicking on it with the mouse. No further instructions were given. Faces were displayed at approximately 256 x 256 pixels initially. When a face was selected, the chosen face doubled in size, while the unselected face halved in size (512 pixels square and 128 pixels square, respectively). This configuration remained onscreen for 3 s before the next trial began. Responses were recorded so that the selection of a cluster one face was coded as the positive class.

6.1.5. Analytical strategy and sample size considerations

We utilised Bayesian logistic mixed models to model the data from both experiments. For the brief display paradigm, we modelled the probability of a participant opting to avoid a face, using two categorical predictors – most importantly, the cluster that the face belonged to (coded zero and one), and the task version they completed, to examine differences across counterbalanced versions. Random intercepts were fitted for faces, and random slopes for cluster category within participants as well intercepts, allowing for each participant to respond differently to the clusters. We set a normal prior on the difference between clusters (i.e., the cluster coefficient) with a mean of zero and standard deviation of .25. In odds, this prior assigns most of its probability density to odds of between 0.77 and 1.28, constraining the effect to be small – that is, we employ a skeptical prior on the differences in approach decisions to cluster zero and one faces.

For the preference paradigm, we estimated an intercept (baseline rate of selecting a cluster one face) with an additional predictor of task to control for differences between the counterbalanced presentations, and had random intercepts per participant. We chose very weakly informative priors for the parameters (Capretto et al., 2020), as constraining the intercept introduces unnecessary bias to the baseline odds of choosing to approach or avoid.

As stated earlier, the use of Bayesian methods is difficult to reconcile with statistical power, and we use restrictive priors to lend a conservative approach to our model. However, as we use a mixed model design, we are able to consider all stimuli and participants as random factors. Finally, we opted to conduct an analogous frequentist power analysis of the brief-display paradigm model via simulation to generate an idea of statistical power, for a range of sample sizes and effect sizes. We sampled a range of fixed effect coefficient sizes (in log-odds) from 0.5 to 1.5, in steps of .25. These effect sizes were evaluated at sample sizes of 20 through to 100 in steps of 20, and repeated 250 times per combination of sample size and effect size. The constant of 100 faces was fixed in the simulations. Even with 20 participants, our design allowed for detection of fixed effects coefficients as small as 0.5, or odds of 1.68. The inclusion of a restrictive prior again indicates that effects must be strong to overwhelm the restriction, guarding against false positive conclusions.

6.2. Results

We report all effects as odds ratios (by exponentiating the estimated coefficients) for a more direct interpretation, and report the mean of the posterior distribution, along with the 95% HDI.

Brief display paradigm. The fixed effect of cluster category showed that the odds of choosing to avoid a cluster one face were double that of a cluster zero faces, *OR* = 2.08, 95% HDI [1.43, 3.11]. This is a strong effect even with a restrictive prior, and as we can work directly with the posterior of the coefficient, it is possible to state various probabilities associated with the effect. For example, there is a 99% probability the odds are greater than one, and a 58% probability the odds are greater than two. Differences between task versions had low odds with wide uncertainty, overlapping zero, *OR* = 1.15 [0.58, 2.21].

Preference paradigm. The fixed intercept term, representing the odds of choosing to approach a cluster one face, was low, *OR* = 0.28 [0.13, 0.56]. That is, accounting for baseline selection tendencies in participants, the odds of selecting a cluster one face were around 0.28 times that of selecting a cluster zero face, indicating much lower likelihood of approaching cluster one faces. Each of the task versions showed wide intervals overlapping with null effects (version B compared to A, *OR* = 1.53 [0.55, 4.60], C compared to A, *OR* = 0.72 [0.26, 2.054], D compared to A, *OR* = 0.68 [.25, 1.76]).

6.3. Discussion

These results demonstrate that observer reaction to faces from the different clusters are in line with the results of the unsupervised clustering solutions generated from trait impressions of faces, whereby clustering techniques organised faces into an approach-avoidance continuum, which was also supported by the extraction of associated facial features. These findings show that observers can form ‘avoid’ decisions of cluster one faces quickly and reliably, and that when given the unprompted option to view faces from the clusters more closely (i.e., approach them), there is a much higher likelihood of approaching cluster zero faces.

7. General Discussion

We report evidence from both computational and behavioural investigations demonstrating a two-cluster solution to social perception using unsupervised learning techniques. First, we found that mean shift, a non-parametric clustering technique, consistently extracted two clusters of perceptual profiles from a large, multidimensional perceived social impression database. These clusters were present in both the full representation of the data as well as within current theoretical representations of impression space. Second, we revealed the consistency of these clusters in other datasets, in both the full and reduced spaces, with some caveats. Third, using Bayesian methods, we explored the likely perceptual profile differences between cluster zero and one, which correspond geometrically to valence components in existing models – cluster one faces are perceived as more mean, aggressive, and unhappy. Fourth, we demonstrated that the probability of belonging to cluster one or zero can be mapped to statistical regularities in facial appearance, captured from neural network or more straightforward (PCA) approaches. Fifth, we provided evidence that observers have a higher probability of engaging in avoidance behaviour of cluster one as compared with cluster zero faces, under both a brief display task as well as an untimed preference task.

Our findings contribute to a more nuanced understanding of the structure of social impressions. Current two- and three-dimensional models of impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013) focus on the use of dimension-reduction techniques like FA or PCA to linearly combine large numbers of correlated perceptions into a reduced number of factors, thought to represent the fundamental components on which impressions are made. These factors typically capture elements of valence/approachability, trustworthiness, or attractiveness (Sutherland et al., 2013, 2015). One motivation here was to examine not how these traits were organised, but how the individual points – the faces themselves – were organised within these theoretical structures. Here, we demonstrated the existence of two clusters of impressions that faces were grouped into, and that these were present in the full and factor representations of the traits.

We found that the separation between these clusters was easily learned – faces that were perceived as sociable, trustworthy, attractive, and caring, were much more likely to be placed in cluster zero, while faces perceived as mean, aggressive, and unhappy were placed in cluster one. An important result from this exploration was the application of LDA, which derived a vector that maximally separated the two clusters, which could be directly compared with existing two- and three-factor solutions. Using geometric techniques, we found that this vector, that maximised the difference between the cluster centres, is closely aligned to valence and approachability factors from existing models, but relatively uncorrelated with dominance factors. This outcome is interesting as it suggests person perception from facial appearance utilises cues of approachability and positive valence, and not cues to dominance. That is, while a vector may exist that scaffolds judgements around dominance specifically, it is relatively underutilised when impressions are actually formed. Our subsequent analyses provided a possible reason why this might be. Faces with a higher probability of being assigned to cluster one appeared less positive and more masculine, both of which are consistent with higher perceptions of dominance (Tipples, 2007; Todorov et al., 2013). This result indicated that, when considering the outcomes for individual faces in terms of perception rather than traits themselves, cues to dominance and approachability were essentially subsumed to give rise to overall impressions.

Our findings suggest a fundamental axis that complements existing models of impression formation - whether a face should be approached or avoided. This fits closely with some of the earliest ideas in social psychological research, namely that a fundamental aspect of cognition is how to respond to stimuli in the environment with regards to moving towards or away from it (Allport, 1935; Bamford & Ward, 2008; Chen & Bargh, 1999). Earlier work on the evaluation of faces has underscored the utility of facial first impressions in making decisions about approach or avoidance behaviours (Todorov, 2008), and the results presented here are in line with this idea. The vast array of social impressions that we form of others, supported as they may be by a smaller set of latent factors, seems to serve to classify individuals into whether they can be approached or avoided. The results of our behavioural experiments also showed that these responses were made rapidly and without instruction, in line with prior work describing how quickly facial trustworthiness decisions are made (Willis & Todorov, 2006).

It is important to note our findings here speak only to aggregate level responses in facial impressions – average impression ratings were taken from a large sample of observers (Bainbridge et al., 2013). Recent work has shown that lower-dimensional representations of impressions are relatively stable within individuals (Sutherland et al., 2019), indicating that aggregate level PCA or FA are useful representations of individual perceptual spaces. An outstanding question is whether individuals exhibit the same cluster structure found at the aggregate level that indexes an approach-avoidance dimension, or whether individuals exhibit different numbers of clusters when making judgements of others.

Our analyses employed a range of different approaches to examine the influence of perceptions and facial appearance on cluster placement, with a focus on model-building and out-of-sample prediction. Where we relied upon significance tests, we generally had very high statistical power. Our use of Bayesian estimation in Experiments 3-5 also allowed us to incorporate sceptical, conservative priors into our models. While the use of priors can often be criticised as inducing subjectivity into analyses, we suggest that here, we have used them to guard against false positive conclusions (van de Schoot et al., 2021) – by using priors to regularise our estimates, we actively made it harder for our models to dissociate and find differences between clusters. This is especially appropriate as we propose a novel finding in the area of social face perception. In Experiments 3-4, we also focussed on the generalisability of our models by assessing their out-of-sample predictive capabilities, lending further robustness. In addition, we employed various approaches to converge on conclusions, such as the use of different face databases in Experiment 2, varying face image representations in Experiment 4, and two behavioural studies in Experiment 5.

Models of face perception are traditionally generated from a combination of White and/or computer-generated faces (Oosterhof & Todorov, 2008; Sutherland et al., 2013), and only recent work has made efforts to incorporate faces of different ethnicities to test for the presence of distinct factors (Jones et al., 2021). Here, the main dataset used for inferences was the 10k US Adult Faces Database (Bainbridge et al., 2013), which, while predominantly White (83%), nonetheless contains a variety of different face ethnicities under various naturalistic poses. While the other datasets we investigated also produced two clusters, they were restricted to White faces only (Lin et al., 2019; Sutherland et al., 2013), which does limit the generalisability of the emergence of two clusters. In our behavioural experiment, we obtained a mixed sample of participant ethnicities, which had a White majority, and insufficient numbers of other groups to obtain reliable estimates of differences between ethnicities. As such, while recent studies point to a general stability of social face perception factors, the generalisability of clusters across the ethnicities of both faces and perceivers has not been fully investigated here.

We conclude that social impressions result in two broad, stable clusters of individuals. Assignment to these clusters depends on a combination of correlated perceptual traits that are associated with clearly different aspects of morphology, which are themselves associated with positive and negative emotional expressions, as well as masculinity. Our findings suggest that the structure of social impression space is more straightforward than estimating the number of latent factors, and supports earlier interpretations of data-driven approaches to social impression formation in that all judgements are characterised by simple approach-avoidance decisions. That is, the practical realisation of current models is a simple binary decision about whether to approach or avoid an individual.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Allport, G. W. (1935). Attitudes. In *A Handbook of Social Psychology* (pp. 798–844). Clark University Press.

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). *Openface: A general-purpose face recognition library with mobile applications*. CMU School of Computer Science.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Balas, B., & Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior*, *77*, 240–248. https://doi.org/10.1016/j.chb.2017.08.045

Balas, B., Tupa, L., & Pacella, J. (2018). Measuring social variables in real and artificial faces. *Computers in Human Behavior*, *88*, 236–243. https://doi.org/10.1016/j.chb.2018.07.013

Bamford, S., & Ward, R. (2008). Predispositions to approach and avoid are contextually sensitive and goal dependent. *Emotion*, *8*(2), 174–183. https://doi.org/10.1037/1528-3542.8.2.174

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*(4), 703–706. https://doi.org/10.1016/j.jrp.2009.03.007

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, *115*(3), 401–423. https://doi.org/10.1037/0033-2909.115.3.401

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2020). Bambi: A simple interface for fitting Bayesian linear models in Python. *ArXiv:2012.10754 [Stat]*. http://arxiv.org/abs/2012.10754

Chen, M., & Bargh, J. A. (1999). Consequences of Automatic Evaluation: Immediate Behavioral Predispositions to Approach or Avoid the Stimulus: *Personality and Social Psychology Bulletin*, *25*(2), 215–224. https://doi.org/10.1177/0146167299025002007

Chiao, J. Y., Bowman, N. E., & Gill, H. (2008). The Political Gender Gap: Gender Bias in Facial Inferences that Predict Voting Behavior. *PLoS ONE*, *3*(10), e3666. https://doi.org/10.1371/journal.pone.0003666

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619. https://doi.org/10.1109/34.1000236

Comaniciu, D., & Meer, P. (1999). Mean shift analysis and applications. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, *2*, 1197–1203 vol.2. https://doi.org/10.1109/ICCV.1999.790416

Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *The Behavioral and Brain Sciences*, *22*(3), 491–517; discussion 518-569. https://doi.org/10.1017/s0140525x99002046

Dolnicar, S., Grün, B., Leisch, F., & Schmidt, K. (2013). Required Sample Sizes for Data-Driven Market Segmentation Analyses in Tourism: *Journal of Travel Research*.

Elliot, A. J., & Thrash, T. M. (2002). Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology*, *82*(5), 804–818. https://doi.org/10.1037/0022-3514.82.5.804

Fiske, S. T. (2018). Stereotype Content: Warmth and Competence Endure. *Current Directions in Psychological Science*, *27*(2), 67–73. https://doi.org/10.1177/0963721417738825

George, P. A., & Hole, G. J. (1995). Factors influencing the accuracy of age estimates of unfamiliar faces. *Perception*, *24*(9), 1059–1073. https://doi.org/10.1068/p241059

Gerlach, M., Farb, B., Revelle, W., & Nunes Amaral, L. A. (2018). A robust data-driven approach identifies four personality types across four large data sets. *Nature Human Behaviour*, *2*(10), 735–742. https://doi.org/10.1038/s41562-018-0419-z

Gray, J. A. (1970). The psychophysiological basis of introversion-extraversion. *Behaviour Research and Therapy*, *8*(3), 249–266. https://doi.org/10.1016/0005-7967(70)90069-0

Haselhuhn, M. P., Ormiston, M. E., & Wong, E. M. (2015). Men's Facial Width-to-Height Ratio Predicts Aggression: A Meta-Analysis. *PLOS ONE*, *10*(4), e0122637. https://doi.org/10.1371/journal.pone.0122637

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Henderson, A. J., Lasselin, J., Lekander, M., Olsson, M. J., Powis, S. J., Axelsson, J., & Perrett, D. I. (2017). Skin colour changes during experimentally-induced sickness. *Brain, Behavior, and Immunity*, *60*, 312–318. https://doi.org/10.1016/j.bbi.2016.11.008

Holzleitner, I. J., Lee, A. J., Hahn, A. C., Kandrik, M., Bovet, J., Renoult, J. P., Simmons, D., Garrod, O., DeBruine, L. M., & Jones, B. C. (2019). Comparing theory-driven and data-driven attractiveness models using images of real women's faces. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1589–1595. https://doi.org/10.1037/xhp0000685

Izenman, Alan Julian. (2008). Linear Discriminant Analysis. In Alan J. Izenman (Ed.), *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning* (pp. 237–280). Springer. https://doi.org/10.1007/978-0-387-78189-1_8

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, *2*(1999), 102–138.

Jones, A. L., Batres, C., Porcheron, A., Sweda, J. R., Morizot, F., & Russell, R. (2018). Positive facial affect looks healthy. *Visual Cognition*, *26*(1), 1–12. https://doi.org/10.1080/13506285.2017.1369202

Jones, A. L., Porcheron, A., Sweda, J. R., Morizot, F., & Russell, R. (2016). Coloration in different areas of facial skin is a cue to health: The role of cheek redness and periorbital luminance in health perception. *Body Image*, *17*, 57–66. https://doi.org/10.1016/j.bodyim.2016.02.001

Jones, A. L., Tree, J. J., & Ward, R. (2019). Personality in faces: Implicit associations between appearance and personality. *European Journal of Social Psychology*, *49*(3), 658–669. https://doi.org/10.1002/ejsp.2534

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*(1), 159–169. https://doi.org/10.1038/s41562-020-01007-2

Khosla, A., Bainbridge, W. A., Torralba, A., & Oliva, A. (2013). Modifying the Memorability of Face Photographs. *Proceedings of the 2013 IEEE International Conference on Computer Vision*, 3200–3207. https://doi.org/10.1109/ICCV.2013.397

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Kramer, R. S. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods*, *49*(6), 2002–2011. https://doi.org/10.3758/s13428-016-0837-7

Kramer, R. S. S., Mulgrew, J., Anderson, N. C., Vasilyev, D., Kingstone, A., Reynolds, M. G., & Ward, R. (2020). Physically attractive faces attract us physically. *Cognition*, *198*, 104193. https://doi.org/10.1016/j.cognition.2020.104193

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience – the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Lin, C., Keles, U., & Adolphs, R. (2019). *Comprehensive trait attributions show that face impressions are organized in four dimensions* [Preprint]. PsyArXiv.

Little, A. C. (2014). Facial appearance and leader choice in different contexts: Evidence for task contingent selection based on implicit and learned face-behaviour/face-ability associations. *The Leadership Quarterly*, *25*(5), 865–874. https://doi.org/10.1016/j.leaqua.2014.04.002

Marlowe, C. M., Schneider, S. L., & Nelson, C. E. (1996). Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased? *Journal of Applied Psychology*, *81*(1), 11–21. https://doi.org/10.1037/0021-9010.81.1.11

Marquardt, D. W., & Snee, R. D. (1975). Ridge Regression in Practice. *The American Statistician*, *29*(1), 3–20. https://doi.org/10.1080/00031305.1975.10479105

Morrison, D., Wang, H., Hahn, A. C., Jones, B. C., & DeBruine, L. M. (2017). Predicting the reward value of faces and bodies from social perception. *PLoS ONE*, *12*(9). https://doi.org/10.1371/journal.pone.0185093

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, *22*(9), 794–809. https://doi.org/10.1016/j.tics.2018.06.006

Porcheron, A., Mauger, E., & Russell, R. (2013). Aspects of facial contrast decrease with age and are cues for age perception. *PLoS ONE*, *8*(3), e57985. https://doi.org/10.1371/journal.pone.0057985

Ritchie, K. L., Palermo, R., & Rhodes, G. (2017). Forming impressions of facial attractiveness is mandatory. *Scientific Reports*, *7*(1), 469. https://doi.org/10.1038/s41598-017-00526-9

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Russell, R. (2009). A sex difference in facial contrast and its exaggeration by cosmetics. *Perception*, *38*(8), 1211–1219. https://doi.org/10.1068/p6331

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Satchell, L. P. (2019). From photograph to face-to-face: Brief interactions change person and personality judgments. *Journal of Experimental Social Psychology*, *82*, 266–276. https://doi.org/10.1016/j.jesp.2019.02.010

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. https://doi.org/10.1016/j.cognition.2012.12.001

Sutherland, C. A. M., Rhodes, G., Burton, N. S., & Young, A. W. (2019). Do facial first impressions reflect a shared social reality? *British Journal of Psychology*. https://onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12390

Sutherland, C. A. M., Rhodes, G., & Young, A. W. (2017). Facial image manipulation: A tool for investigating social perception. *Social Psychological and Personality Science*, *8*(5), 538–551. https://doi.org/10.1177/1948550617697176

Sutherland, C. A. M., Rowley, L. E., Amoaku, U. T., Daguzan, E., Kidd-Rossiter, K. A., Maceviciute, U., & Young, A. W. (2015). Personality judgments from everyday images of faces. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01616

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In *Anxiety and the anxiety disorders* (pp. 681–706). Lawrence Erlbaum Associates, Inc.

Tipples, J. (2007). Wide eyes and an open mouth enhance facial threat. *Cognition and Emotion*, *21*(3), 535–557. https://doi.org/10.1080/02699930600780886

Todorov, A. (2008). Evaluating Faces on Trustworthiness. *Annals of the New York Academy of Sciences*, *1124*(1), 208–224. https://doi.org/10.1196/annals.1440.012

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738. https://doi.org/10.1037/a0032335

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1–26. https://doi.org/10.1038/s43586-020-00001-2

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353–E3361. https://doi.org/10.1073/pnas.1409860111

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257. https://doi.org/10.1037/pspa0000098

Watson, D., & Clark, L. A. (1997). Extraversion and its positive emotional core. In *Handbook of personality psychology* (pp. 767–793). Academic Press. https://doi.org/10.1016/B978-012134645-4/50030-5

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x