Using Cluster Analysis to Quantify Systematicity in a Face Image Sorting Task

by

Alison Campbell
Bachelor of Science, University of Victoria, 2014

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Psychology

Supervisory Committee

Using Cluster Analysis to Quantify Systematicity in a Face Image Sorting Task

by

Alison Campbell
Bachelor of Science, University of Victoria, 2014

Supervisory Committee

Dr. James Tanaka, (Department of Psychology)
Supervisor

Dr. Adam Krawitz, (Department of Psychology)
Departmental Member

Abstract

Supervisory Committee
Dr. James Tanaka, (Department of Psychology)
Supervisor
Dr. Adam Krawitz, (Department of Psychology)
Departmental Member

Open sorting tasks that include multiple face images of the same person require participants to make identity judgments in order to group images of the same person. When participants are unfamiliar with the identity, natural variation in the images due to changes in lighting, expression, pose, and age lead participants to divide images of the same person into different "identity" piles. Although this task is being increasingly used in current research to assess unfamiliar face perception, no previous work has examined whether there is systematicity across participants in how identity groups are composed. A cluster analysis was performed using two variations of the original face sorting task. Results identify groups of images that tend to be grouped across participants and even across changes in task format. These findings suggest that participants responded to similar signals such as tolerable change and similarity across images when ascribing identity to unfamiliar faces.

Table of Contents

List of Tables

List of Figures

Acknowledgments

This project was supported by the National Sciences and Engineering Research Council of Canada. And Jim, obviously, because he's the best.

Introduction

Traditionally, the research focus in face recognition has been mainly on the problem of telling people apart. For example, investigations on the holistic nature of face processing (e.g. Carey & Diamond, 1977; Sergent, 1984; Farah, Wilson, Drain, & Tanaka, 1998) emphasized the perceptual problem of discriminating between different faces despite that all faces have the same overall template (i.e., two eyes, a nose, and a mouth in the same configuration). Because faces are recognized at the level of *identity*, research in face perception and memory has focused on measuring perceptual sensitivity to differences between individuals, or *between-person variability.* In the majority of these studies, each identity is represented by single photograph throughout the experiments or a small set of highly constrained, experimenter-created photos that differ in only one aspect, such as expression or pose. There is therefore little to no difference in the images used to represent each identity in the experiment – the *exemplars* of a given face.

By extension, much of the theory of face recognition is dedicated to explaining the processes involved in facial identity discrimination. For example, the influential face-space model of face recognition was intended to capture how "the natural variation of real faces affected face processing" (Valentine, Lewis, & Hills, 2016, p. 1998) and has been successfully implemented to account for a multitude of face recognition phenomena, including differences for faces of different ethnicities and distinctiveness (Valentine, 1991; Blanz & Vetter, 1999; Lewis, 2004; Todorov, Said, Engell, & Oosterhof, 2008; Valentine, Lewis, & Hills, 2016). To represent faces with respect to naturally occurring variability, face representations are construed as points within a multidimensional feature space and located along "dimensions that could be used to discriminate faces" (Valentine,

1991; Figure 1). The "variation" which is intended to be captured is therefore *between-person* variation, with the implication that face recognition operates on a system that primarily codes this between-person variability.

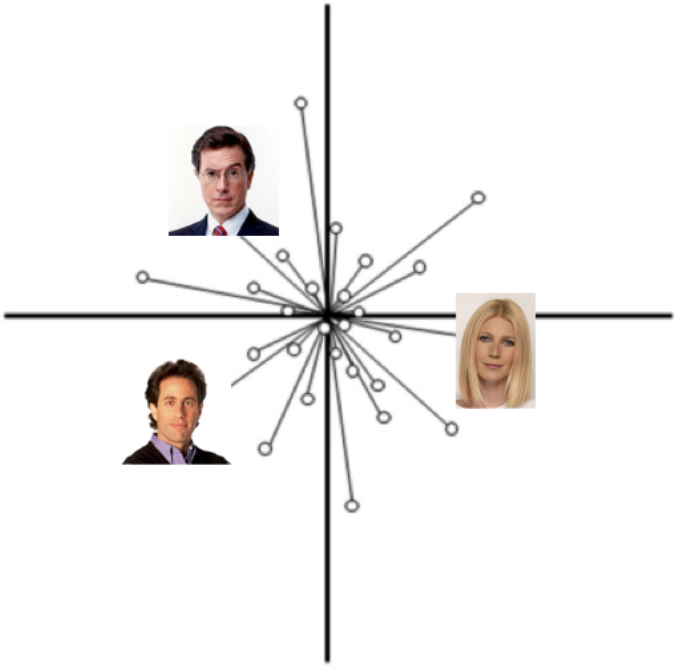


Figure 1. Face identity representations in face-space models of face memory and perception as first proposed by Valentine (1991). Each identity is represented as a point in a multidimensional psychological similarity space.

However, a separate line of research has revealed that between-person variability is only half of the equation in face recognition. In experimental studies, the use of a single or very few image exemplars for each individual's face reflects an assumption that the appearance of a face can be adequately captured and described by a single photograph. Of course, facial appearance can vary due to changes in expression, pose (head angle), hairstyle, makeup, age, health, and weight. Environmental factors such as illumination level and lighting direction can also change low-level image properties and,

consequently, the appearance of a face. Jenkins, White, Van Monfort, and Burton (2011) empirically tested the question of how much these changes affect identity recognition using a novel face image sorting task. In this task, British participants were given 40 images and instructed to sort the images by identity, so that photos of the same face were grouped together. Unbeknownst to the participants, the set only contained two identities with 20 different photos per person (two Dutch celebrities). Although a correct sorting of the images would only have 2 piles, the average number of piles created was 7.5, and virtually all of the sorting errors were due to the participants' failure to group together images of the same face. Importantly, this effect was only obtained for participants for whom the faces were unfamiliar – that is, when participants did not know the two identities. When testing Dutch participants who were familiar with the two identities, participants sorted the images quickly and accurately.

This finding has two major implications. The first is the demonstration of the scale of *within-person variability*; far from being trivial, the extent of variability across images of the same person was enough to lead participants to perceive different images of the same person as entirely different people. In other words, the degree of *within-person* variability was large enough that variance between photos of one person was perceived as *between-person* variance. Indeed, the same researchers found that variance in attractiveness across photos of the same person can amount to differences found between persons (Jenkins et al., 2011). Moreover, effects of within-person variability on face recognition is not specific to face matching across photographs. Performance on tasks requiring participants to match a live person to their photograph is also very poor (Kemp, Towell, & Pike, 1997; Megreya & Burton, 2008).

Second, the finding that within-person variance impaired the ability to recognize identity when the face was unfamiliar (for British participants) but not when the face was familiar (Dutch participants) indicates that there is important learning that occurs in the process of familiarization that confers tolerance to within-person variability. Research on the role of within-person variation in the process of face learning has shown that learning to match different exemplars of one face – a task known as *face matching* – does not lead to better face matching for other faces (Dowsett, Sandford, & Burton, 2016). Put differently, developing tolerance for the within-person variability of one identity does not transfer or generalize to the within-person variability of another identity; learning what to accept as within-person variance is identity-specific.

The trouble that participants have in recognizing an unfamiliar face across different images is likely related to the idea that every face varies a little differently. Using principal components analysis to compare the physical properties of image exemplars of different faces, Burton et al. (2016) found idiosyncrasies in the dimensions along which the exemplars for different individuals varied. These results suggest that developing tolerance to the within-person variability of a given identity would require learning the parameters that are specific to their face (Bruce, 1994; Burton, 2013).

The finding that the dimensions along which individual faces vary is idiosyncratic across identities yet can be learned through experience (i.e. familiarization) suggests that variance of a given face has a structure that might be represented as part of a face representation. Instead of being coded as a single point in face space (a single combination of space dimension values), identities may be coded as a region within

which an exemplar for that individual may fall. Such a model may depict identities as regions or "clouds" within face-space (Figure 2).

The finding that participants *fractionate* images of an unfamiliar face by dividing them into multiple group has been replicated multiple times, and the face sorting task is now commonly used for measuring unfamiliar face perception (Neil, Cappagli, Karaminis, Jenkins, & Pellicano, 2016; Laurence & Mondloch, 2016; Baker, Laurence, & Mondloch, 2017; Laurence, Zhou, & Mondloch, 2016; Short & Wagler, 2017; Short, Balas, & Wilson, in press). In discussing their results, Jenkins et al. (2011) note that unfamiliar faces tend to be fractionated because of the problem of "integrating dissimilar images. It is difficult to find commonalities among photos of the same face that justify grouping them together." (p. 315). In face-space terms, the errors in recognizing identity across different images of the same person when a face is unfamiliar may be due to participants not having learned the boundaries or structure of the face variance for the given identity. If participants perceive identity by defaulting to similarity, it would be predicted that there would be systematicity across participants in which images are grouped together as one "identity".

The primary purpose of Experiment 1 is to conduct a novel examination of sorting patterns across participants to determine whether there is systematicity in images that are grouped. By pooling participant data and applying cluster analysis techniques, this analysis not only addresses the surface question of whether the same images are grouped across participants, but whether participants perceive the same "identities" within the image set. An additional version of the face sorting task is conducted in Experiment 2 to

determine if presentation format of the face images influences sorting accuracy and, by extension, identity perception.

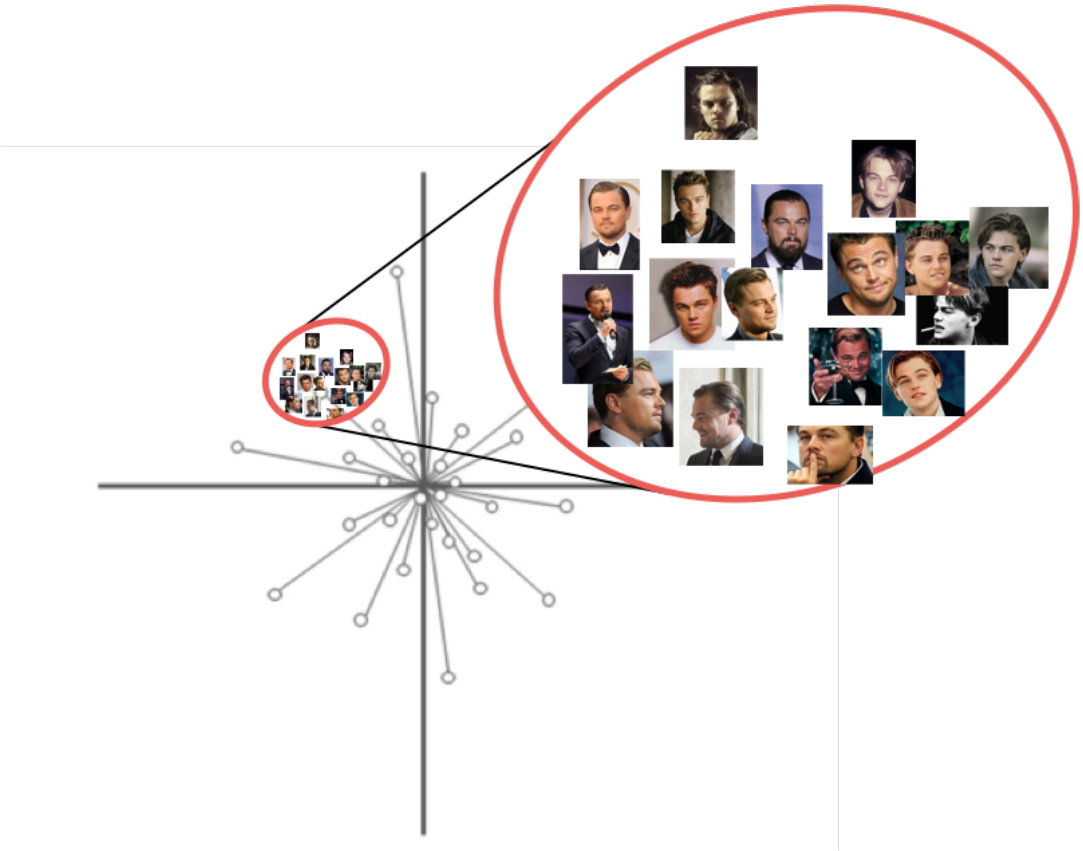


Figure 2. Leonardo DiCaprio's face "cloud" in face-space. The boundaries of the cloud represent the within-person variability unique to him.

Experiment 1

Method

Participants

Twenty undergraduate students from the University of Victoria participated for course credit (14 female, mean age = 20.4 years, SD = 3.0).

Stimuli

Unlike conventional tests of face memory and perception (e.g., Cambridge Face Memory Test, Duchaine & Nakayama, 2006) that use highly constrained stimuli to control the parameters along which the stimuli vary, face matching tasks use natural images of faces from the kind of photographs that would be found on someone's Facebook page or by Google Image search. To underline the fact that these images are sampled from the real world, these unconstrained stimuli are often referred to as "ambient images" (Jenkins et al., 2011; Sutherland et al., 2013; Murphy, Ipser, Gaigg, & Cook, 2015; Ritchie & Burton, 2017). Collectively, ambient images have the advantage of containing an individual's natural range of within-person variability. Using ambient images in face learning and face matching tasks is therefore not only about using ecologically valid images in the sense that the images reflect the environmental conditions under which faces are actually and recognized; it is perhaps even more so about sampling real world images to capture the within-person variability that is unique to each identity. The stimulus set therefore consisted of 20 images each of two Dutch celebrities (Bridget Maasland, Chantal Janzen) obtained using a search on Google Image.

The two identities were chosen on the basis that they are well known celebrities in the Netherlands and consequently images are easily obtainable online, yet they are not known to Canadian participants. Second, these identities were the same two used in the original Jenkins et al. (2011) face image sorting task, allowing our results to be more easily comparable to the results obtained in the original study. For each identity, the first 20 images retrieved using the their full name as the search terms that fulfilled image quality requirements were accepted (Figure 3). To preserve the range of within-person variability, the only requirements were no occlusions (glasses, sunglasses, hats, etc.) to the face and that the image resolution was at least 72 pixels per inch when cropped to surround the face. All images were cropped so that the face was centered and occupied roughly 70% of the image. Images were then printed in greyscale on 48 mm x 48 mm cards.

Figure 3. Images used in the face sorting task. Images were printed in greyscale on 48 mm x 48 mm cards. (Top) Bridget Maasland; (Bottom) Chantel Jansen.

Procedure

After shuffling the image cards, the experimenter spread out all images in front of the participant. The participant was instructed to sort the photos by grouping together images of the same person. Participants were told that they were allowed to re-arrange the images as much as they wanted and that there were no time constraints.

Data Analysis

Sorting measures. For each participant, the total number of piles were recorded and, for each pile, a list was made of the images that were grouped together (Figure 4). Because participants were allowed to re-arrange the images freely throughout the task, the numbering of the piles was arbitrary. From this list, all pairwise combinations of images that were grouped together were generated by taking each pile list separately and computing all possible image combinations within it (Figure 5). From the image combinations, every image pairing was scored as either a match (images depicting the same person) or mismatch (images depicting different people). In a perfect sorting solution (two piles each containing images of only one identity), two piles of 20 images would generate the maximum number of obtainable of matches (380) and no mismatches. If all images of both identities are grouped together in a single pile, the maximum number of mismatches obtained is 400. Thus, for each participant, a match rate can be computed as the ratio of the observed match count to the maximum match number, and similarly for the mismatches. Finally, as image matching reflects person recognition and image mismatching occurs when one person is misidentified as another person, these rates provide a measure of recognition and misidentification. Thus, the dependent measures for the card sorting task for each participant are number of piles, recognition rate, and misidentification rate.

subjno	pile	image_id
8002	1	E10
8002	1	E2
8002	1	E4
8002	1	E11
8002	2	F6
8002	2	F7
8002	3	F3
8002	3	F11
8002	4	F13
8002	4	F5

Figure 4. A sample of recorded data from a single participant showing that images E10, E2, E4, and E11 were grouped together (pile 1), F6, and F7 were grouped together (pile 2), etc. The numbering of the piles was arbitrary.

subjno	pile	img_1	img_2
8002	1	E10	E2
8002	1	E10	E4
8002	1	E10	E11
8002	1	E2	E4
8002	1	E2	E11
8002	1	E4	E11
8002	2	F6	F7
8002	3	F3	F11
8002	4	F13	F5

Figure 5. An example of all pairwise image combinations generated from the sample data in Figure 4. The four images grouped together in pile 1 yield six image pairings; piles 2 and 4 which contained only two images only yield the single image pairings.

Image-level similarity coefficients. To prepare the data to be submitted to cluster analysis, a measure of association between the images was computed. The sorting data from all participants were first aggregated and image similarity coefficients were computed for each pair of images. *Similarity* in the context of cluster analysis refers to a quantification of the strength of association between objects when derived from binary (or categorical) data.[1] In this case, the observed data is considered binary because the observed data represents either the presence or absence of an image within a pile. Figure 2.4 shows a sample of the data represented in binary (rather than listed combination) format. Each row represents one pile and every column represents a different image. The 94 x 40 matrix therefore contains data from 94 piles aggregated from all participants, with each cell containing either a '1' to indicate the presence of the image in the pile or a '0' to indicate the absence.

Images grouped together in the same pile may be described as having a co-occurrence; for example, the co-occurrence of images E1 and E3 in pile 3 represented by the code '1' in columns 1 and 3 along row 3. Similarity coefficients are derived from binary data by comparing these co-occurrences to co-absences and single occurrences. Table 1 shows the four possible alternatives when comparing binary data for two objects. In the cases when the objects have equivalent values (*a* and *d*), the *positive matches* that indicate co-occurrence can be distinguished from *negative matches* that represent co-absence. There are several options for deriving similarity that differ with respect to whether both kinds of matches are treated as a factor of association (a quality referred to

---

[1] By contrast, when object proximities are inferred from continuous data, the relationship is expressed as *dissimilarity*, or more specifically as a *distance* when the items project into a metric space.

as *symmetry*), and the weight or importance given to matches relative to mismatches (Table 2). The choice of a similarity coefficient depends on the nature of the material under study (Sneath & Sokal, 1973), but as different coefficients can yield different patterns of relationships between objects (for an extensive review, see Gower & Legendre, 1986), it is important to choose a coefficient that is well suited to the data and the research question.

The most simple matching coefficient treats all matches equivalently and weights matches and mismatches equally. However, as in the present case, the primary interest may be only matches related to co-occurrence. Like other asymmetrical coefficients, the Jaccard coefficient (Jaccard, 1912) developed by the ecological surveyist Paul Jaccard are useful when the question of object association focuses on *co-dependence*. The Jaccard coefficient is a ratio of how frequently objects co-occur relative to the frequency that they are observed independently. Co-absence data is not directly informative on the dependence between objects, so it is not included in the calculation of similarity. This yields a value between 0 and 1 for every possible pair of images, which can be represented in a n x n matrix, where n is the number of objects.

Because the interest in the association between images relates to their likelihood of being grouped together, the Jaccard coefficient was used to calculate similarity coefficients for every pair of images. The similarity values for each pair of images generated from all card sorting data collapsed across participants were represented in a 40 x 40 matrix (Figure 7).

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	1	0	0	0	1	1	0	0	0	0	1	1	1	0
4	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0
5	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 6. A sample of observed image sorting data represented in binary format. Each row represents one pile and every column represents a different image. A sorting data set can then be represented in a p x n matrix, where p is the number of observations and n is the number of objects in the set. The full sorting data observed in Experiment 1 is therefore represented in a 94 x 40 matrix from 94 total piles across participants and 40 different images.

Table 1

Four alternatives when comparing binary data for two objects.

		Object 1	
		1	0
Object 2	1	*a* Co-occurrence (Positive match)	*b* Single occurrence (Mismatch)
	0	*c* Single occurrence (Mismatch)	*d* Co-absence (Negative match)

Table 2

Various similarity measures for binary data.

Similarity Coefficient	Formula	Symmetry	Weighting
Simple matching	$s = (a + d) / (a + b + c + d)$	Yes	Matches and mismatches are equally weighted.
Jaccard coefficient (1908)	$s = a / (a + b + c)$	No	
Rogers & Tanimoto (1960)	$s = (a + d) / (a + 2(b + c) + d)$	Yes	Increased penalty for mismatches.
Sneath & Sokal (1973)	$s = a / (a + 2(b + c))$	No	Mismatches are weighted more heavily than matches.

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
E1	1.00	0.43	0.54	0.43	0.54	0.43	0.67	0.67	0.60	0.43	0.43	0.60
E2	0.43	1.00	0.38	0.74	0.43	0.67	0.43	0.43	0.38	0.82	0.67	0.38
E3	0.54	0.38	1.00	0.48	0.60	0.54	0.60	0.74	0.60	0.38	0.43	0.48
E4	0.43	0.74	0.48	1.00	0.43	0.74	0.48	0.48	0.43	0.74	0.74	0.48
E5	0.54	0.43	0.60	0.43	1.00	0.48	0.60	0.54	0.82	0.38	0.48	0.60
E6	0.43	0.67	0.54	0.74	0.48	1.00	0.48	0.54	0.43	0.67	0.74	0.43
E7	0.67	0.43	0.60	0.48	0.60	0.48	1.00	0.60	0.60	0.43	0.48	0.54
E8	0.67	0.43	0.74	0.48	0.54	0.54	0.60	1.00	0.54	0.43	0.48	0.60
E9	0.60	0.38	0.60	0.43	0.82	0.43	0.60	0.54	1.00	0.38	0.43	0.60
E10	0.43	0.82	0.38	0.74	0.38	0.67	0.43	0.43	0.38	1.00	0.74	0.43
E11	0.43	0.67	0.43	0.74	0.48	0.74	0.48	0.48	0.43	0.74	1.00	0.43
E12	0.60	0.38	0.48	0.48	0.60	0.43	0.54	0.60	0.60	0.43	0.43	1.00

Figure 7. A sample of the 40 x 40 Jaccard similarity matrix containing association data for each pair of images based on all card sorting data, collapsed across participants.

Clustering techniques. The choice of a clustering algorithm depends on both the type of data being analysed and expectations about the kind of latent structures that may be present in the data. Classic cluster analysis techniques broadly fall into two categories: hierarchical and non-hierarchical[2]. Hierarchical cluster algorithms analyze the data for nested structure and generate tree-shaped solutions with small clusters that merge together successively until all data items are integrated into the hierarchy. It is almost certainly the form most commonly used and known, owing to its development and ongoing use in biology (Sneath & Sokal, 1973).

By contrast, partitioning methods create "flat" solutions with distinct clusters. This approach is aptly described as "partitioning" because the process of clustering requires that each cluster contain at least one item, and each item must be assigned to one cluster. This approach is also commonly referred to as *optimization* due to the fact that the algorithm assigns objects to clusters by searching for the optimal division which minimizes the distance between items of a cluster and maximizes the distance between items of different clusters. However, this method requires a pre-specified number of clusters to be generated by the algorithm and which is represented by *k*. The *k*-means methods – the most popular of these algorithms – describes each cluster in terms of the average values of all its members. The optimal solution is generated by iterating through a process of item assignment, computing the resulting *k*-mean (the "centroid"), re-computing all within-cluster and between-cluster distances, and then either accepting or

---

[2] A third class, model-based clustering, consist of more recently developed techniques. Unlike hierarchical and optimization techniques (which cluster heuristically using rules for object assignment) model-based techniques introduce the use of log-likelihood and probability density functions to assess the underlying data distribution, but are better suited for large datasets.

rejecting the new object. This process iterates over all possible item combinations (into the pre-specified number of partitions) until it settles on the best solution.

Importantly, this approach defines clusters as groups of "homogenous" data points with low dispersion across space and, additionally, which are separable from other points in the dataset. "Good" clusters (i.e., the groups that cluster algorithms are designed to seek) are therefore described as both internally cohesive and separable (Everitt, Landau, Leese, & Stahl, 2011).

Applying the *k*-medoids algorithm. In sorting tasks, participants are instructed to group images by identity. Images assigned to different piles therefore suggests that participants perceive the face images as categorically (i.e. identity-level) distinct. A non-hierarchical partitional technique was therefore used to determine if and how the dataset could be optimally broken up. If the similarity scores – a measure of image co-dependence – from the sorting data support a multi-cluster solution, it suggests that images were systematically grouped together in the sorting task.

First, in order to represent the data spatially, the Jaccard similarity scores for the images were converted to dissimilarities by subtracting each similarity value from 1(i.e., high similarity values yield low dissimilarity values to reflect closer proximity in the variable space). The Duda-Hart test (1973) was then used to test for absence of structure before proceeding with the cluster analysis. This tests whether the dataset should be split into clusters by estimating the *p*-value of the null hypothesis (no structure, i.e. $k = 1$).

A *k*-medoids algorithm was then used as an alternative to the conventional *k*-means algorithm due to its robustness to outliers and the use of cluster exemplars (actual data points) instead of cluster means (pam in R; Kaufmann & Rousseeuw, 1990). The

process of partitioning around medoids is the same as that of *k*-means except that, on each iteration of the partitioning, the exemplar that is closest to the cluster mean in the variable space is designated as the cluster centre. The within-cluster error can therefore be computed simply by summing the pairwise dissimilarities, rather than having to first square the error before summing. Hence, the effect of outliers and noise on within-cluster error is reduced.

Silhouette coefficients. Silhouette values were used to assess cluster quality and determine the optimal number of clusters (partitions) supported by the data (Rousseeuw, 1987; Kaufmann & Rousseeuw, 1990). Unlike other clustering criteria that operate on the sum of squared error, silhouette values can be generated directly from dissimilarity values and can be computed at levels of the item, cluster, and global solution. The silhouette of each item or object (in this case, for each image) is a measure of how closely it is associated to its assigned cluster relative to the closest alternative (i.e., neighbouring) cluster. For each object *i*, a silhouette value *s*(*i*) is obtained that is specific to a particular clustering solution and that expresses a ratio of the within-cluster dissimilarity for *i* (i.e., the average dissimilarity of *i* to all other objects within the same cluster) and the between-cluster dissimilarity for the closest neighbour cluster to *i* (i.e., the average dissimilarity of *i* to all other objects in the competing cluster closest to *i*) (Figure 8).

$$s(i) := \frac{b(i) - a(i)}{max(a(i), b(i))}.$$

Figure 8. Computation of the silhouette value for any given object (*i*) in a given cluster solution. *b*(*i*) denotes the average between-cluster dissimilarity between *i* and each object in the closest neighbouring cluster to *i*. *a*(*i*) denotes the average within-cluster

dissimilarity between *i* and each object in the same cluster. The difference between $b(i)$ and $a(i)$ is divided by the largest of the two values to constrain the range to [-1,1].

The closest neighbour cluster is like the second-best choice for object *i*, so this ratio can be considered as a measure of cluster fit of an object to its assigned cluster relative to its fit to the second-best alternative. Clusters are conceptualized as data points that pool together to form homogenous groups that are distinct and separable from other data points/data groups; Good cluster fit is therefore defined as small within-cluster dissimilarity (i.e. highly similar, homogenous) and large between-cluster dissimilarity (i.e. separation). Lastly, silhouette values are normalized to range from 1 to -1, where values approaching 1 indicate very good cluster fit and values approaching -1 indicate very poor classification (smaller between-cluster than within-cluster dissimilarity). A value near zero indicates that the object fits as well to its assigned cluster as the nearest competing cluster, which may indicate either an outlier or a suboptimal partitioning solution.

The quality of each cluster or for a particular clustering solution can be summarized by averaging silhouette values for all objects within each cluster or across all clusters, respectively. The average $s(i)$ for each cluster or full solution is known as *average silhouette width*. Average silhouette widths obtained from different choices for the number of clusters (i.e., different values for *k*) can be used to compare the fit of the data to different clustering solutions and these were used as a formal criterion for choosing the optimal number of clusters.

Evaluating within-identity clusters. Strong between-identity structures were expected given the very low rate of image identity mismatches. This was confirmed in a

preliminary clustering of the full dataset where a two-cluster partition of the images by identity yielded the best fit (Figure 9). The Duda-Hart test for k=1 was significant ($p <$ .001), indicating greater hetergeneity in the data than would be expected in an absence of structure. This is due to the high degree of separability between images of the two identities relative to the separability within the images of each identity. Because the research question relates to the structure of the image sorting within each identity (i.e., how images of a single identity are fractionated), the full sorting data was split by identity and analysed in separate cluster analyses. However, the two identities can be considered to be 'natural' clusters within the data because there is a known separation between the two types of items given that the low misidentification rate indicates that the two identities are rarely grouped. The performance of clustering techniques is often assessed in recovery studies that apply clustering algorithms to simulated data where the natural clusters are known. The results obtained from the application of the partitioning around medoids (PAM) algorithm on the recovery of the two identity clusters therefore serves as a useful benchmark for assessing the quality of other cluster solutions. The average silhouette width obtained from a solution that clustered images of each identity separately is 0.51. In other words, the best case scenario for cluster separability with the current dataset yields a silhouette width of 0.51. Although Kaufmann & Rousseau (1990) have proposed a subjective interpretation for evaluating cluster quality from silhouette values, these guidelines are based on experience with their own datasets that may have data structures and distributions that differ from those of face image dissimilarities. Whereas they propose a minimum silhouette width of 0.25 and strong cluster structure at widths over 0.71, the value obtained when partitioning along the natural (meaning, in his

context, data-driven) identity clusters indicates that a value of 0.51 approaches the upper limit for silhouette width for this dataset given the distribution of data within each of the two identities. At this value, there is twice as much dissimilarity between clusters as there is within clusters. When adjusting Kaufmann & Rousseau's guidelines to account for this upper limit, we may expect strong structure at a width of 0.51, good structure for values between 0.31-0.50, adequate structure at values between 0.10-0.30, and no substantial structure for silhouette widths less than 0.10. Adopting a minimum acceptable silhouette width of 0.10 means that cluster solutions are accepted only if there is at least 10% more dissimilarity between clusters than within.

Finally, visualizations of the optimal number of clusters in a scatterplot along the first two principal components were then used to analyse the data. Plot preparation and the k-medoid analysis was done using the *cluster* package in R (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2017).

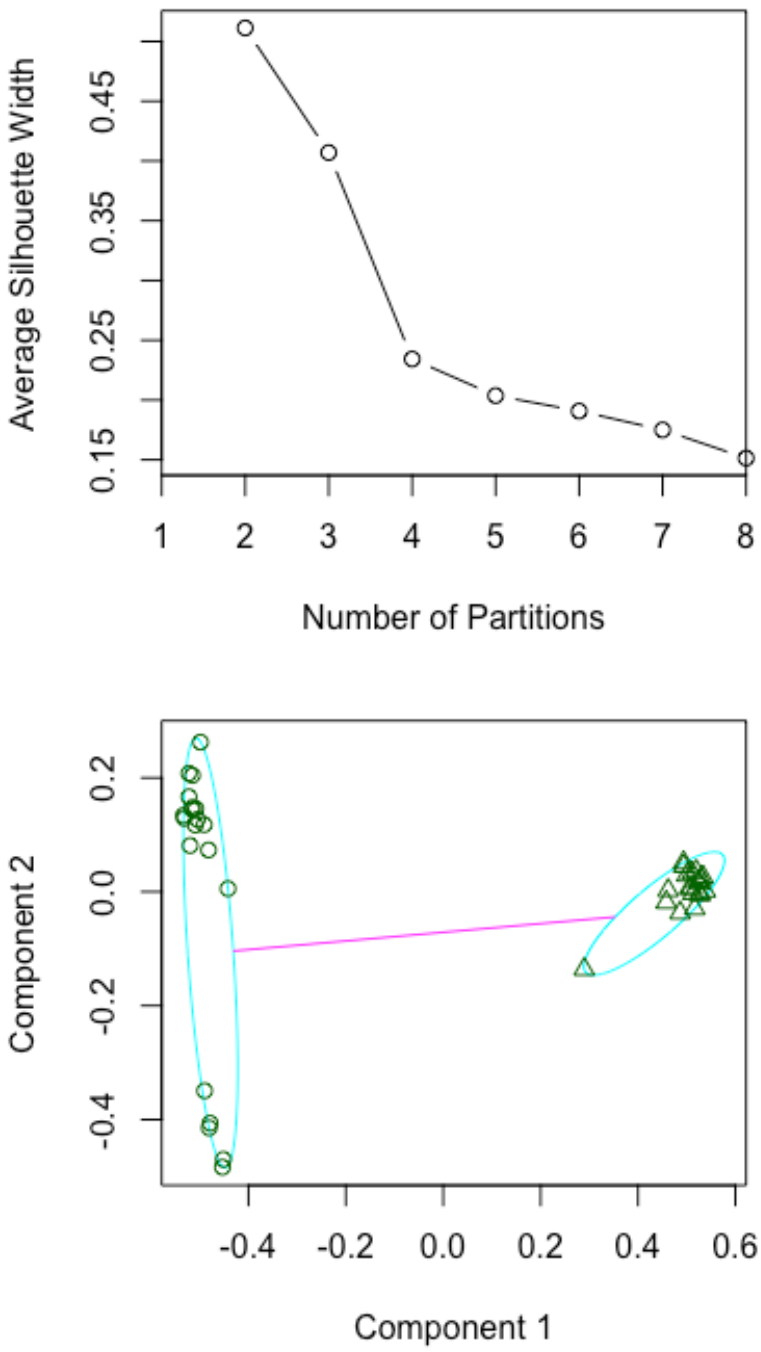


Figure 9. (Top) Average silhouette widths for various solutions from a k-medoid analysis of the image dissimilarity data for all images (Bridget and Chantel). (Bottom) A scatterplot of the two-cluster solution for images of both identities plotted along the first two principal components. Each partition contained only one identity.

Results

Sorting Measures

Sorting results are summarized in Table 3. The mean number of piles created was 4.7 (SD = 3.5, range = 2-12). Mean recognition rate was 0.67 (SD = 0.30), and mean misidentification rate was 0.008 (SD = 0.02).

Table 3

Mean number of piles, recognition rates, misidentification rates, and number of clusters obtained per identity in Experiments 1 & 2.

	Experiment 1 (simultaneous)	Experiment 2 (sequential)
Piles	4.7 *(3.5)**	8.0 *(7.0)**
Recognition rate	0.67 *(0.30)* **	0.45 *(0.30)***
Misidentification rate	< 0.01 *(0.02)*	0.07 *(0.22)*
Number of clusters		
Bridget	2 *[0.32]*	6 *[0.14]*
Chantel	2 *[0.15]*	3 *[0.11]*

Note: Standard deviation in parentheses. Average silhouette width in square brackets. Symbols indicate significance of Wilcoxen rank-sum pairwise comparisons: *p = .04 for mean number of piles; **p = .03 for recognition rate.

Cluster Analysis

First, using only the dissimilarity values between images of the first identity (Bridget), the Duda-Hart test for k=1 was significant ($p < .001$), indicating greater hetergeneity in the data than would be expected for an absence of structure. Proceeding with the analysis, the k-medoids algorithm was then run seven times to produce solutions with two to eight partitions. Average silhouette widths for each solution is shown in Figure 10. The highest average width (best fit) was achieved by a two-cluster solution with good fit (average silhouette width = 0.32). Silhouette values for each partition are

summarized in Table 4 and Figure 10 displays the two partitions in the space of the first two principal components. The optimal solution of $k = 2$ separated the data into a small, strongly structured cluster (cluster 1, $n_c = 5$, average silhouette width = 0.51) and a larger cluster with adequate structure (cluster 2, $n_c = 15$, average silhouette width = 0.25). Cluster medoids and members are displayed in Figure 11. Notably, all the images showing Bridget with her hair in a retro updo dissociate from the remaining images and form the small, highly structured cluster.

The plot of average silhouette widths (Figure 10) for different clusterings of the Bridget images shows another peak in the silhouette width for a six-cluster solution. Figure 12 shows that the $k = 6$ solution keeps the "retro updo" cluster intact but further divides the remaining images into five smaller clusters. Two of these new clusters have silhouette widths below the acceptable value of 0.10, but three other adequately structured clusters are identified (Table 5). Like cluster 1, the images in these clusters appear to share superficial qualities such as closed mouth smiles (cluster 2a) or open mouth smiles (cluster 2b) (Figure 13). Together, these analyses indicate a structure within the data of at least two internally cohesive and separable groups, plus some evidence of three other distinct image groups.

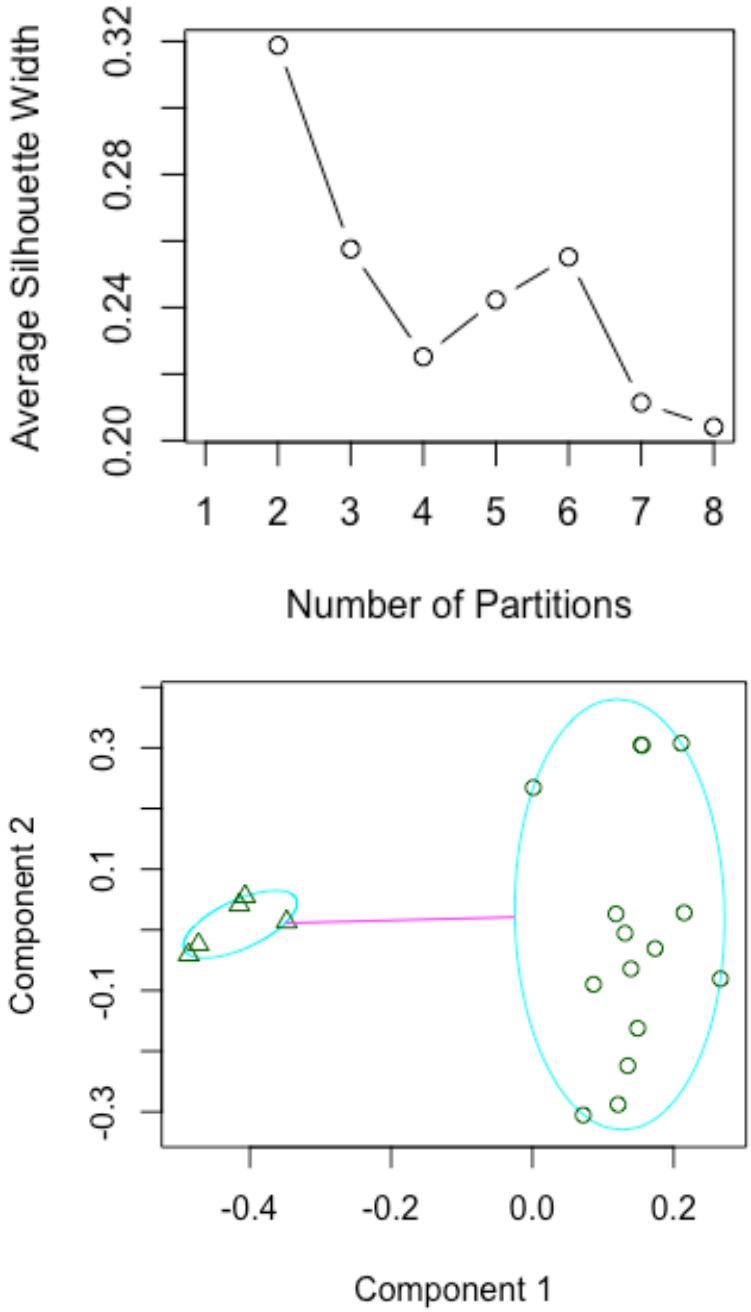


Figure 10. (Top) Average silhouette widths for various cluster solutions of Bridget image dissimilarities. (Bottom) The two-cluster solution for Bridget image dissimilarity values in the space of the first two principal components.

Table 4

Cluster sizes and silhouette widths of a two-cluster solution of Bridget images.

Cluster, k_c	n_c	average $s(i)$
1	5	0.51
2	15	0.25
Average silhouette width		0.32

Cluster 1:

Cluster 2:

Figure 11. Cluster medoids and members of the two-cluster solution of the Bridget images. Members are shown in descending order of silhouette value.

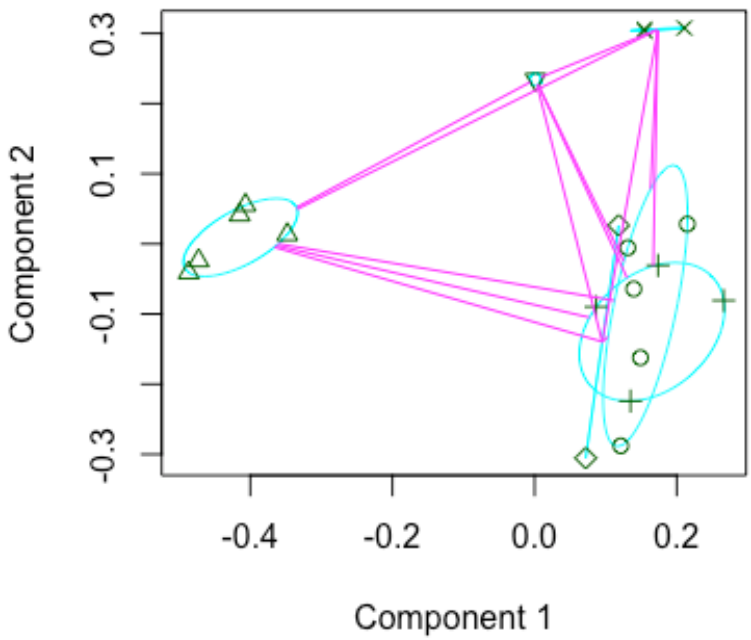


Figure 12. The six-cluster solution for Bridget image dissimilarity values in the space of the first two principal components.

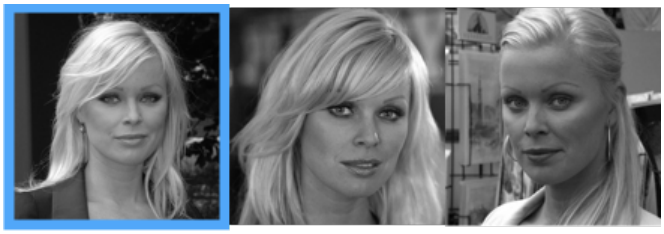
Table 5

Cluster sizes and silhouette widths of a six-cluster solution of Bridget images.

Cluster, k_c	n_c	average s_i
1	5	0.49
2 (2a)	3	0.44
3 (2b)	4	0.24
4 (2c)	2	0.18
5	5	0.003
6	1	0.00
Average silhouette width		0.26

Note: Cluster numbers in parentheses indicate the partitions of Cluster 2 from the $k = 2$ cluster solution.

Cluster 2a:



Cluster 2b:

Cluster 2c:

Figure 13. Cluster medoids and members of the three additional clusters obtained from a six-cluster solution of the Bridget images (average silhouette > 0.10). Members are shown in descending order of silhouette value.

The clustering procedure was repeated using the dissimilarity data of the Chantel images. The Duda-Hart test for k=1 was significant ($p < .001$), and the null hypothesis that k=1 was rejected. Average silhouette widths for each solution are shown in Figure 14. The highest average width (best fit) was achieved by a two-cluster solution with loose yet adequate structure (average silhouette width = 0.15)[3]. Silhouette values for each

[3] The silhouette plot (Figure 14) also showed a second peak in silhouette widths for a seven-cluster solution. However, although larger relative to other solutions, a silhouette width of 0.10 suggests little substantial structure and so the solution was disregarded.

partition are summarized in Table 6. The optimal solution of $k = 2$ contained a medium sized cluster with good structure ($n_c = 8$, average silhouette width = 0.26), but the silhouette of the second cluster shows no substantial structure in the remaining image data ($n_c = 12$, average silhouette width = 0.08). Being close to zero, the silhouette value of the second cluster indicates that the association of these images to each other is not very strong relative to their separability to items in the competing cluster. In other words, this cluster has high heterogeneity. Examination of the objects in the space of the first two principal components shows the stronger cluster on the right; by comparison, there is much more dispersion of the images in the second weakly structured cluster on the left (Figure 14). Cluster medoid and member images of the two partitions are shown in Figure 15. Of images from the one acceptable cluster, many show Chantel with loosely curled/centre-parted hair. As might be expected from the dispersion of the images in the cluster solution, the remaining images in the second partition are varied with no obviously common features among them. In summary, there is evidence for a distinct subset of Chantel images that are highly associated, the association of the remaining images is highly heterogenous.

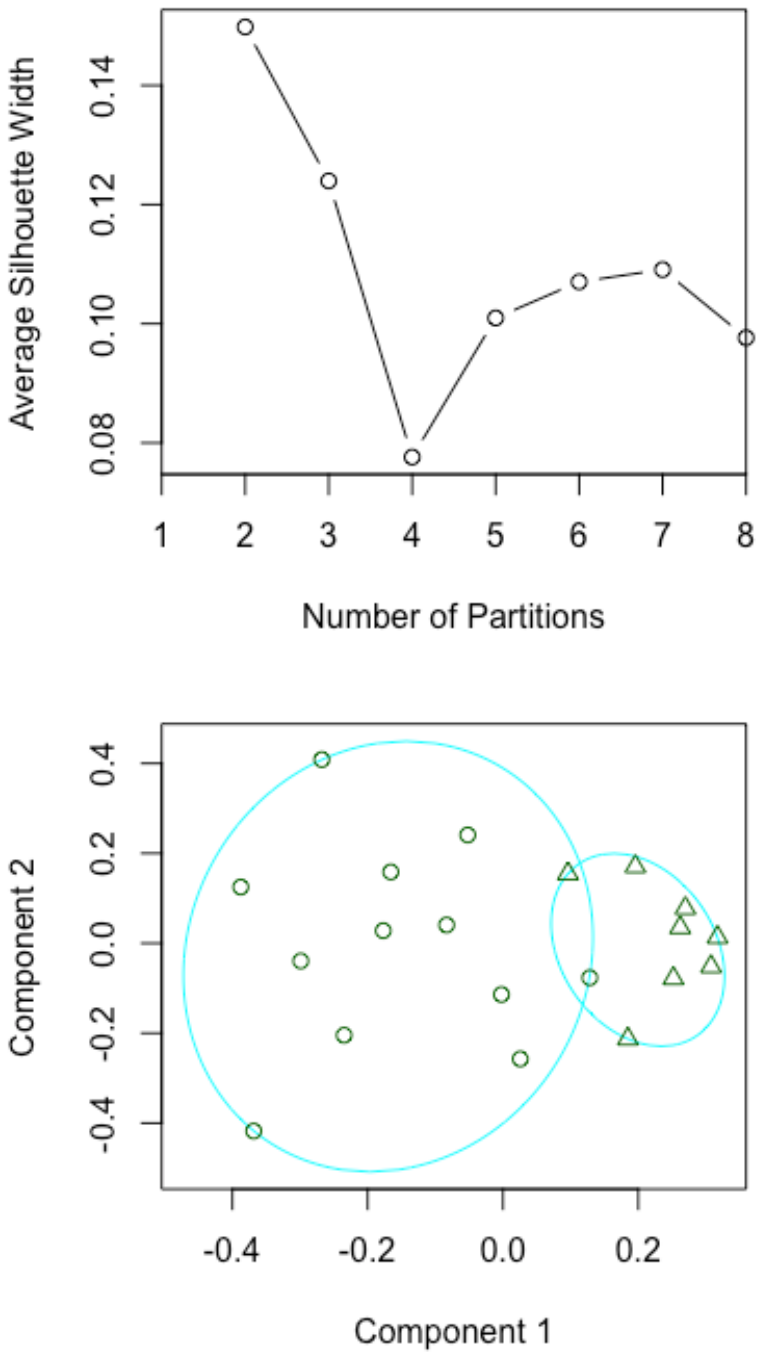


Figure 14. (Top) Average silhouette widths for various cluster solutions of Chantel images. (Bottom) The two-cluster solution for Chantel images in the space of the first two principal components.

Table 6

Cluster sizes and silhouette widths of a two-cluster solution of Chantel images.

Cluster, k_c	n_c	average s_i
1	8	0.26
2	12	0.08
Average silhouette width		0.15

Cluster 1:

Cluster 2:

Figure 15. Cluster medoids and members of the two-cluster solution of Chantel images. Members are shown in descending order of silhouette value.

Discussion

First, the results and sorting errors observed in the face sorting task replicate the general finding that participants create more piles than the number of identities in the image set because they are unable to recognize the same person across different images (e.g., Jenkins et al., 2011; Short & Wagler, 2017). Recognition errors indicate that participants are failing to match images of the same identity: instead of putting all the Bridget (or Chantel) images into a single pile, participants create multiple piles of Bridget (or Chantel) images. By contrast, the extremely low rate of identity mismatching in the piles (misidentification) indicates that participants are readily able to differentiate when images depict two different identities.

The results of Experiment 1 also demonstrate how cluster analyses can be applied to image sorting data to evaluate for common sorting patterns across participants. When aggregating image grouping data across all participants, the cluster solution of all images successfully recovered the broad identity clusters that are expected given that participants rarely grouped images of the different identities.

A cluster solution of the Bridget images returned two main clusters that included a set of strongly clustered images (cluster 1) all showing Bridget with her hair in a retro updo. Further analysis of the remaining images provided additional evidence of three other clusters of images which also share descriptive qualities. A cluster solution of Chantel images also returned two clusters. Unlike the solution for Bridget, only one of the two clusters had substantial structure. Interestingly, images within the acceptable cluster appeared to share superficial appearance qualities, whereas the images of Chantel in the highly heterogenous partition were also highly varied in appearance. Although the

number of piles created cannot be statistically compared with the results of the cluster analyses, it is interesting to note that the average number of piles that participants created (mean = 4.7) corresponded to the total number of within-identity image clusters found by the optimal cluster solutions (Bridget = 2, Chantel = 2).

There are two major implications from the cluster results. First, the discovery of well-structured clusters within the pooled image sorting data supports the hypothesis that there is systematicity in the way that participants group and divide the images of each identity. Second, the image composition of these clusters suggests that images that are regularly grouped images tend to share common superficial appearance qualities, such as hairstyle, expression, and pose. This is consistent with theories proposing that unfamiliar face representations are mainly pictorially-coded and that unfamiliar face matching relies on process of image matching (Bruce, 1982; Burton, Jenkins, & Schweinberger, 2011; Hancock, Bruce, & Burton, 2000).

It's also possible that the format of the task contributes to some of the identity-splitting sorting behaviour that is reported here and elsewhere. The most widely used form of the face image sorting task uses a 40 image set composed of 20 images of two identities and the full set is presented to participants at the outset of the task (e.g., Jenkins et al., 2011; Neil et al., 2015; Laurence, Zhou, & Mondloch, 2016; Short & Wagler, 2017). The number of piles created by participants may be partially due to participant expectations about how many identities are likely to be present given the number of images. Additionally, if participants 'identify' (i.e. ascribe identity to) unfamiliar faces through a process of image comparison, having the full set of images presented simultaneously may increase the amount of attention required to make pairwise

comparisons of each image against all others. Indeed, research on the effects of simultaneous versus sequential image presentation on face identity judgments in forensic line-up research shows that sequential face presentation increases perceptual sensitivity (Steblay, Dysart, Fulero, & Lindsay, 2001; Steblay, Dysart, & Wells, 2011) and sequential presentation has been promoted as the optimal form for actual eyewitness line-up identification purposes (Technical Working Group for Eyewitness Evidence, 1999; Wells, Malpass, Lindsay, Turtle, & Fulero, 2000; Wells, 2006).

In order to minimize these potential task format effects, a sequential version of the face matching task was used in Experiment 2. In the sequential version, the experimenter provided the participant with only one image at a time and participants were required to make an identity judgment (i.e., to place the image in a pile) before another image was presented. In this format, effects due to expectations about how many identities are contained in the set may be lessened because participants do not know at the beginning of the experiment how many images will be provided to them. However, more importantly, it was intended that this format would promote a pairwise image comparison strategy by encouraging the participant to focus on one new image at a time when matching to other previously seen images. It was therefore predicted that participants would sort the photos with fewer recognition errors and thereby create fewer piles.

Experiment 2

Method

Participants

Twenty undergraduate students from the University of Victoria participated for course credit (17 female, mean age =21.0 years, SD = 4.0).

Stimuli

The set of face images used in Experiment 2 were the same used in Experiment 1.

Procedure

The procedure for Experiment 2 was the same as Experiment 1 except that images were provided sequentially to the participant as they sorted the images. After shuffling the deck of images, the experimenter told the participant that they would be given images and that they should sort the photos by grouping together images of the same person. The experimenter then provided the participant one image at a time from a shuffled stack of the face images. Participants were required to group together each new image before an additional image was provided. Participants were told that they were allowed to re-arrange the images as much as they wanted and that there were no time constraints.

Data Analysis

Sorting measures. Sorting measures were computed from individual sorting solutions as in Experiment 2.

Cluster analysis. As in Experiment 1, image dissimilarity values were computed from Jaccard scores obtained from collapsing sorting data across participant. The same

partitioning procedures were used as in Experiment 1, with image data clustered separately for each identity. The composition of the within-identity clusters from each experiment were also compared.

Results

Sorting Measures

Sorting measures are summarized in Table 3. The mean number of piles created was 8.0 (SD = 7.0, range = 1-32). Mean recognition rate was 0.45 (SD = 0.30), and mean misidentification rate was 0.008 (SD = 0.22).

Shapiro-Wilk normality tests indicated that the data distributions for the sorting measures deviated from normal. Two-tailed Wilcoxen-Mann-Whitney tests were used as a nonparametric alternative for independent t-tests to test differences across experiments in the number of piles, recognition rates, and misidentification rates. The number of piles was reliably greater in the sequential version of the task compared to the original task when the participant is given all images simultaneously ($W = 273.5$, $p = .04$). Recognition rate was also lower in the sequential version ($W = 119.0$, $p = .03$), although there was no significant difference in misidentification rates.

Cluster Analysis

As in Experiment 1, strong between-identity structures emerged when clustering using data from images of both identities (Figure 16), and the Duda-Hart test for k=1 was significant ($p < .001$); the null hypothesis that k=1 was rejected, and within-identity clusters were assessed by clustering the images for each identity separately.

Beginning with Bridget, the Duda-Hart test for k=1 was significant ($p < .01$). The k-medoids algorithm was run seven times to produce solutions with two to eight partitions of the Bridget images. Average silhouette widths for each solution are shown in Figure 17. The highest average width (best fit) was achieved by a six-cluster solution with weak yet acceptable structure (average silhouette width = 0.14). Figure 17 displays the six partitions in the space of the first two principal components. Silhouette values for each partition are summarized in Table 7. Four out of the six clusters from the optimal solution of k = 6 had adequate structure (silhouette values over 0.10; see Table 7). One cluster of four images showed no substantial structure (average silhouette width <0.10), and one isolated singleton cluster was found[4].

Cluster medoids and member are shown in Figure 3.3. One notable aspect of this cluster solution is that the two strongest clusters consist of image pairs. Surprisingly, there is no readily descriptive quality that links the images in these pairs. It is also notable that, as in Experiment 1, a subset of the images showing Bridget with the "retro updo" also emerge as a cluster.

[4] Note that silhouette values cannot be calculated for single item clusters because within-cluster dissimilarity does not apply.

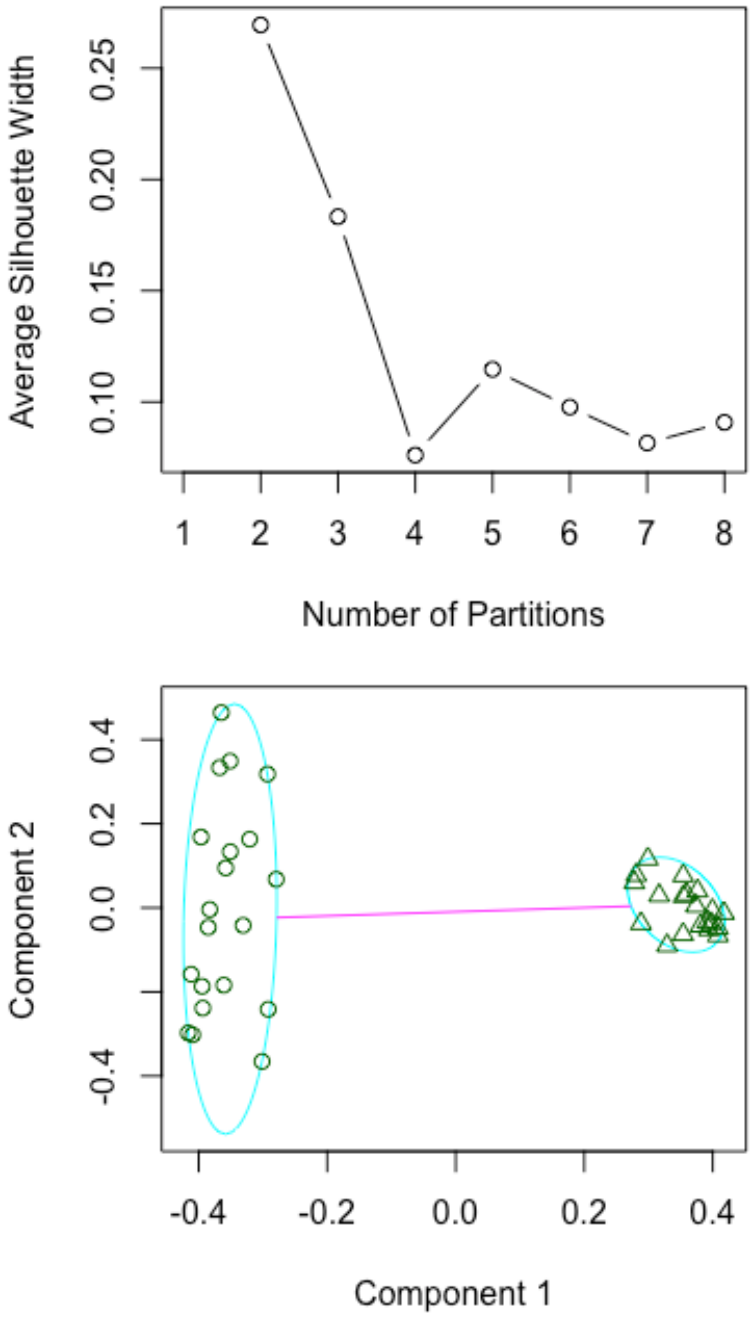


Figure 16. (Top) Average silhouette widths for various solutions from a k-medoid analysis of the image dissimilarity data of all images (Bridget and Chantel). (Bottom) A scatterplot of the two-cluster solution for images of both identities plotted along the first two principal components. The partitions each contained only one identity.

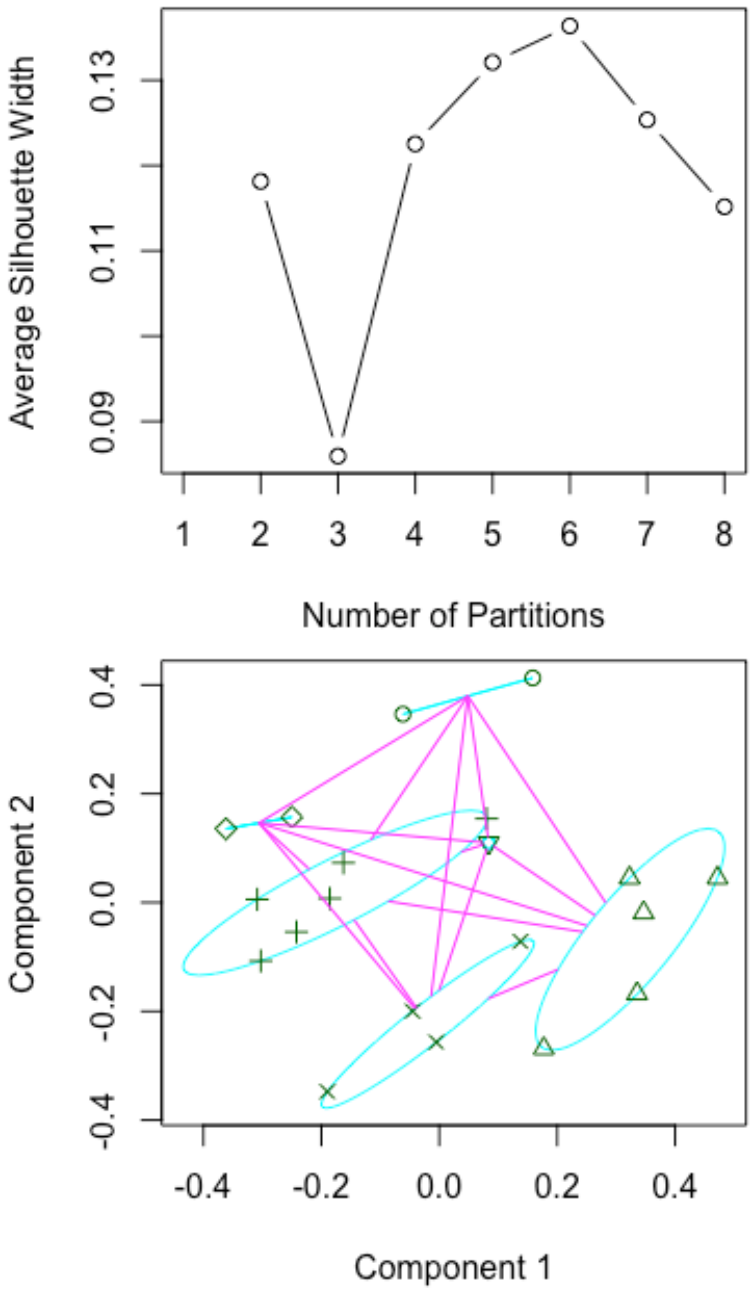


Figure 17. (Top) Average silhouette widths for various cluster solutions of Bridget images. (Bottom) The six-cluster solution for Bridget image dissimilarity values in the space of the first two principal components.

Table 7

Cluster sizes and silhouette widths of a six-cluster solution of Bridget images.

Cluster, k_c	n_c	average $s(i)$
1	2	0.23
2	2	0.21
3	5	0.17
4	6	0.10
5	4	0.09
6	1	NA
Average silhouette width		0.14

Cluster 1:

Cluster 2:

Cluster 3:

Cluster 4:

Cluster 5:

Cluster 6:

Figure 18. Cluster medoids and members of the six-cluster solution of the Bridget images. Members are shown in descending order of silhouette value.

For images of Chantel, the Duda-Hart test for k=1 was significant ($p < .05$)

average silhouette widths obtained of solutions with two- to eight-clusters is shown in Figure 19. The highest average width (best fit) was achieved by a three-cluster solution with loose but adequate structure (average silhouette width = 0.11). Table 8 shows a small two-item cluster with fairly strong structure (average silhouette width = 0.48); the two remaining clusters do not show substantial structure (average silhouette width < 0.10). Medoids and cluster member images are presented in Figure 20; images of the only acceptable cluster consists of the only two images of Chantel with bangs.

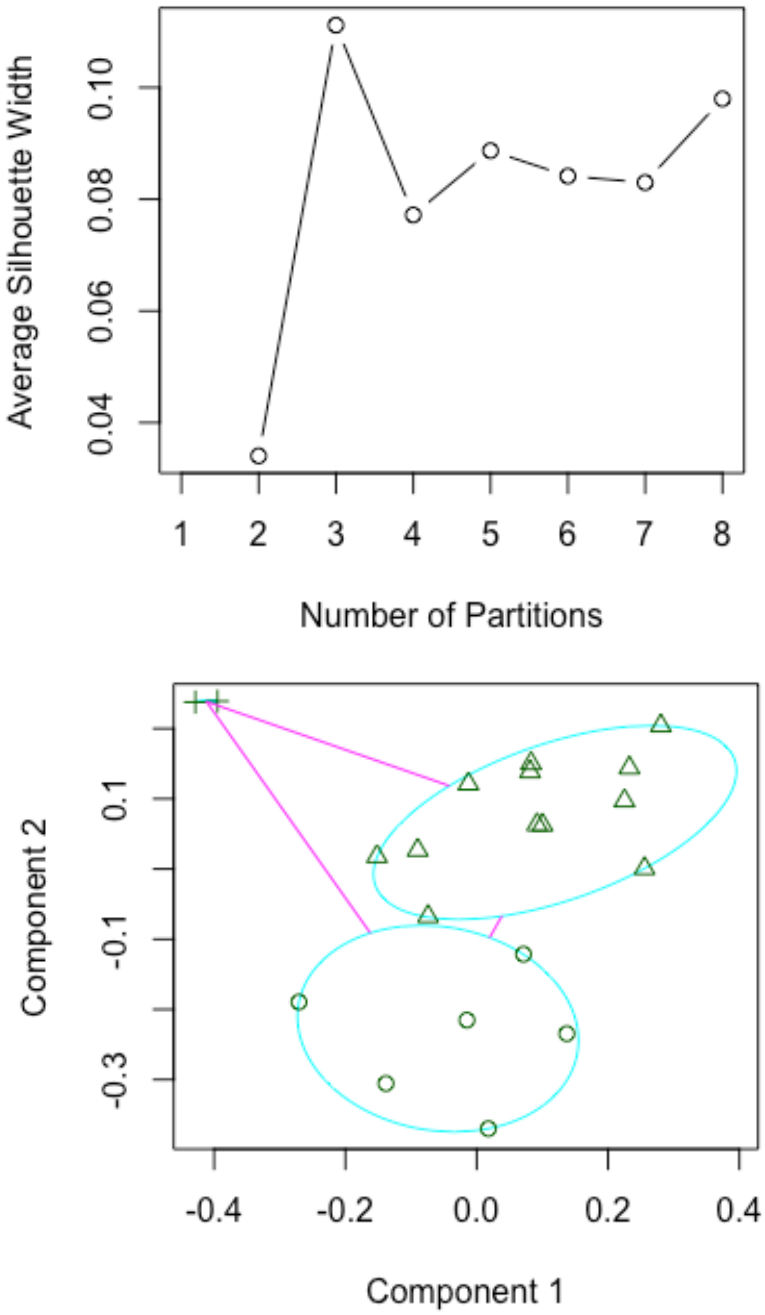


Figure 19. (Top) Average silhouette widths for various cluster solutions of Chantel image dissimilarities. (Bottom) The three-cluster solution for Chantel image dissimilarity values in the space of the first two principal components.

Table 8

Cluster sizes and silhouette widths of a three-cluster solution of Chantel images.

Cluster, k_c	n_c	average $s(i)$
1	2	0.48
2	12	0.08
3	6	0.06
Average silhouette width		0.11

Figure 20. Cluster medoids and members of the three-cluster solution of the Chantel images. Members are shown in descending order of silhouette value.

Cluster composition

Across experiments, cluster solutions can be evaluated in terms of the number of clusters obtained for each identity, the cluster quality as assessed by the silhouette values, and the image composition of the clusters. Tables 3.4 (Bridget images) and 3.5 (Chantel images) summarize the clusters and their composition from the optimal cluster solutions in Experiments 1 and 2. For both identities, the sorting data obtained when using the simultaneous versions of the task (Experiment 1) yielded an optimal cluster solution with fewer partitions compared to sorting data obtained in the sequential version (Experiment 2). Average silhouette widths of the cluster solutions was also greater for the simultaneous task data cluster solution, indicating that the images not only clustered in fewer groups, but that there was greater internal cohesion within clusters and more separability between clusters.

Despite the difference in the number and quality of the clusters obtained across experiments, certain images appear to cluster together despite changes in presentation format. For example, four of the five images that cluster together most strongly in the simultaneous sorting data (Experiment 1) also cluster together in the sequential sorting data (Experiment 2). The cluster solution of Bridget images obtained in Experiment 2 also yielded the only isolated single-item cluster consisting of image B20 (see Cluster 6, Figure 18). Interestingly, this image had the lowest silhouette value of all images in the cluster solution obtained in Experiment 1, indicating the poorest fit relative to the distribution of the other images (see Cluster 2, Figure 11, last image).

Compared to the Bridget image clusters, the clusters of Chantel images had lower silhouette values, meaning reduced within-cluster cohesion and reduced between-cluster

separability. In general, there was greater heterogeneity in the Chantel image pairings. Yet, there were still robust image groups that appeared across experiments. Table 3.5 shows that most of the images from the strongest cluster observed in Experiment 1 reappear as cluster 2 of Experiment 2. These images fall under the description of Chantel with "loosely curled/centre parted hair". The strongest cluster returned from the analysis of Experiment 2 sorting data was an image pair; these particular images are also neighbours within their assigned cluster in the analysis of Experiment 1 data. Although there is a much higher degree of heterogeneity in the image associations for Chantel, the strongest image clusters from both experiments appear robust to changes in the presentation format of the task.

Table 9

Comparison of Bridget cluster composition across Experiments 1 and 2.

Experiment 1			*Experiment 2*		
Cluster	Images	average *s*(*i*)	Cluster	Images	average *s*(*i*)
1*	**B10**, **B2**, **B4**, **B11**, B6	0.51	1	B9, B1	0.23
2	B14, B19, B9, B7, B3, B1, B5, B8, B16, B13, B17, B12, B18, B26, B20	0.25	2	B14, B2	0.21
Overall		*0.32*	3*	**B10**, **B11**, **B4**, B16, **B2**	0.17
2a	*B9, B5, B16*	*0.44*	4‡	**B15**, **B3**, **B8**, B12, B5, B6	0.10
2b‡	***B8***, ***B15***, ***B3***, *B14*	*0.24*	5	B7, B13, B17, B18	0.09
2c	*B18, B12*	*0.18*	6	B20	*N/A*

			Overall		*0.15*

Table 10

Comparison of Chantel cluster composition across Experiments 1 and 2.

Experiment 1			*Experiment 2*		
Cluster	Images	average *s*(*i*)	Cluster	Images	average *s*(*i*)
1*	**C17**, C15, **C20**, **C14**, **C19**, **C18**, **C10**, **C16**	0.26	1§	**C6**, **C7**	0.48
2‡§	**C11**, C9, **C6**, **C7**, **C5**, C2, **C3**, C12, C8, **C4**, **C1**, C13	0.08	2*	**C18**, **C20**, C12, **C10**, **C17**, **C19**, C2, **C14**, C9, C13, **C16**	0.08
			3‡	**C11**, C15, **C1**, **C3**, **C5**, **C4**	0.06
Overall		*0.14*	*Overall*		*0.11*

Discussion

In Experiment 2, participants were blind to the size of the image set at the beginning of the experiment. Sequential presentation was used to encourage more careful and systematic image comparison during the process of sorting by identity. These changes in the task were predicted to increase sorting accuracy by promoting identity recognition across images. Contrary to this prediction, the sequential task condition actually led to poorer identification as reflected in a decrease in the rate of recognition and the number of piles created compared to the original simultaneous presentation format used in Experiment 1. Although changes in the presentation format affected recognition of the same identity across different images, the ability to differentiate identity remained near perfect as indicated by very low misidentification rates.

Corresponding to the increase in the average number of piles, the optimal cluster solutions indicated more partitions in the sorting data. However, the quality of the clusters obtained in the current experiment are lower compared to those of the first experiment, indicating less regularity in the grouping patterns across participants. Despite that regularity was reduced in this experiment, it is noteworthy that the strongest clusters of Experiment 1 re-emerged as clusters in Experiment 2. The presence of these Experiment 1 “alpha clusters” despite the additional noise in the Experiment 2 data supports the idea that the clusters obtained in Experiment 1 reflect an underlying structure in within-identity variability.

Lastly, it is again interesting to note that the number of piles created by participants increased from an average of 4.7 (simultaneous) to 8.0 (sequential) in this version of the task, and at the same time the number of within-identity image clusters

found by the optimal cluster solutions also increased from 4 to 9 (Bridget = 6, Chantel = 3).

General Discussion

The primary research question of this experiment was whether there was systematicity in how these multiple identity piles were formed across participants. As participants were instructed to group the images based on identity, a participant that divides images of Chantel into three piles presumably does so because they effectively see three identities instead of one. The piles they create therefore reflect not only how *many*, but *what* identities they perceive in the images. With this in mind, systematicity in the sorting data across participants is not just about whether participants tend to group the same images; it also gets to a deeper question of whether different participants are perceiving the same 'identities'.

A growing number of studies have implemented the face image sorting task (Jenkins et al., 2011) to measure unfamiliar face recognition (i.e., face matching ability) under a variety of conditions (Short & Wagler, 2017; Short, Balas, & Wilson, in press), including in the context of other-ethnicity faces (Laurence, Zhou, & Mondloch, 2016), in typically developing and autism spectrum disorder diagnosed children (Neil et al., 2016; Laurence & Mondloch, 2016; Baker, Laurence, & Mondloch, 2017). Yet, no other studies have examined the composition of the piles created in any systematic way. Using sorting data of unfamiliar faces from healthy young adults, we demonstrate how to aggregate sorting data across participants, apply a cluster analysis to determine the optimal number of partitions with the images, and use techniques to assess the quality of various clusters.

Despite task variations that significantly impacted face matching according to basic sorting measures, it was found that certain image clusters emerged from the data in both experiments. The flat cluster algorithm selected for this analysis generates clusters

based on both the association of items as well as their separability from other items. Solutions assigning images to different clusters indicates not only that images of a cluster are repeatedly grouped together but, just as importantly, that images from different clusters are repeatedly *not* grouped together. Evidence for systematicity in image grouping across participants therefore suggests that participants applied similar boundaries around images based on shared perceptions of the 'identities' in the image set.

The results of the analysis raise questions about the nature of the signal that is driving these perceptual identity judgments. It has been proposed that the identification of unfamiliar and familiar faces is characterized by different properties, most notably by Hancock, Bruce, & Burton (2000) who have argued that unfamiliar faces are "matched" – not identified – on the basis of simple image comparison (Bruce, 1982; Johnston & Edmonds, 2009; Burton, Jenkins, & Schweinberger, 2011). One way to use the results of the current experiment to explore this hypothesis further is to compare the image cluster 'classifications' obtained by our human participants with the classifications that may be obtained using a machine learning algorithm trained to classify visual images based on low- and mid-level image properties. Along the same line, basic physical image analysis techniques could be used to obtain physical image similarity scores for each pair of images. Physical similarity scores could then be compared against perceived identity 'similarity' as defined by grouping frequency data (i.e. Jaccard scores).

Alternatively, unfamiliar face matching may involve similarity judgments that are distinctly psychological – that is, related to high-level human face perception. Future studies may address this possibility using human face similarity ratings for the stimuli collected in a separate rating task. It may be that the identification of unfamiliar faces is

linked to specialized face perception processes that use shape, colour, and texture information differently for the perception of faces.

As a secondary research question, image presentation format was altered in Experiment 2 to present images sequentially to participants rather than the standard simultaneous presentation procedure (i.e. all at once). It was predicted that this format would increase face matching ability by decreasing attention demands during the facial image comparison process. Surprisingly, an effect in the opposite direction was obtained: sequential presentation resulted in more recognition errors, despite that participants were instructed that they could change their sorting at any time and that. Thus, by the end of the experiment when participants had received all images, they are effectively in the same sorting condition as the participants who received all the images from the outset.

Additionally, much less systematicity was observed across sorting patterns in the sequential version of the task and the strongest clusters tended to have fewer items. This may be attributable to differences in the order in which participants received images as the images were presented in a randomly shuffled order for each participant. An effect related to the temporal ordering of image presentation would suggest that identity 'boundaries' created early in the experiment with fewer images persist through the remainder of the experiment. Possible effects of ordering may be assessed in future experiments where the order of the images is pre-set to reveal images in an optimal ordering. For example, images of Bridget may be revealed by 'daisy-chaining' the images according to human similarity ratings to gradually sample the range of variance for Bridget in a step-wise manner. If participants do accept images into a category based on

similarity to a recently accepted item, such a technique could be used to "teach" participants the within-person variability boundaries of a new identity.

Although this analysis of the image sorting task was inspired by its rising popularity in the face learning literature, the method and its analysis is applicable in studies of other object categories and in category learning. For faces, the increasing appreciation for the scale and effects of within-person variability encourages a view of face recognition as essentially a process of categorizing face images (retinal or photographic) into face identity categories. One reason that open-ended image sorting may be gaining increasing popularity over conventional, 2-alternative computerized face matching tasks (e.g., Clutterback & Johnston, 2002; 2004; Megraya & Burton, 2006) is because it has the additional benefit of revealing which images collectively share enough qualities to be perceived as a cohesive identity category. Because it is intuive and natural, the sorting task is ideal for the study of categorization and yield data that is rich in information about observer responses to signals in the stimuli; the application of new techniques such as cluster analysis gives even more access to this data.

References

Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition*, *161*, 19-30.

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*(3), 277.

Blanz, V., & Vetter, T. (1999, July). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 187-194). New York: ACM Press/Addison-Wesley Publishing Co.

Bonebright, T. L. (1996). An investigation of data collection methods for auditory stimuli: Paired comparisons versus a computer sorting task. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 275-278.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943-958.

Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(1), 105-116.

Bruce, V. (1994). Stability from variation: The case of face recognition the MD Vernon memorial lecture. *The Quarterly Journal of Experimental Psychology*, *47*(1), 5-28.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science*, *195*(4275), 312-314.

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985-994.

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition*, *11*(7), 857-869.

D'Argembeau, A., Van der Linden, M., Comblain, C., & Etienne, A. (2003). The effects of happy and angry expressions on identity and expression memory for unfamiliar faces. *Cognition & Emotion*, 17, 609–622.

Dowsett, A. J., Sandford, A., & Burton, A. M. (2016). Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. *The Quarterly Journal of Experimental Psychology*, *69*(1), 1-10.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576-585.

Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis: Wiley series in probability and statistics. West Sussex: John Wiley & Sons, Ltd.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. (1998). What is ‘‘special” about face perception? *Psychological Review*, 105, 482–498.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, *3*(1), 5-48.

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330-337.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37-50.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, 121, 313–323.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577-596.

Kaufman, L. & Rousseeuw, P. J. (1990). Finding Groups in Data. An Introduction to Cluster Analysis. New York: John Wiley & Sons, Inc.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*(3), 211-222.

Laurence, S., & Mondloch, C. J. (2016). That’s my teacher! Children’s ability to recognize personally familiar and unfamiliar faces improves with age. *Journal of Experimental Child Psychology*, *143*, 123-138.

Laurence, S., Zhou, X., & Mondloch, C. J. (2016). The flip side of the other-race coin: They all look different to me. *British Journal of Psychology*, *107*(2), 374-388.

Lewis, M. (2004). Face-space-R: Towards a unified account of face recognition. *Visual Cognition*, *11*(1), 29-69.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2017). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876.

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 577.

Neil, L., Cappagli, G., Karaminis, T., Jenkins, R., & Pellicano, E. (2016). Recognizing the same face in different contexts: Testing within-person face recognition in typical development and in autism. *Journal of Experimental Child Psychology*, *143*, 139-153.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*(5), 897-905.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*. **20**: 53-65.

Sergent, J. (1984). An investigation into component and configural processes underlying face perception. *British Journal of Psychology*, 75, 221–242.

Short, L. A., & Wagler, M. C. (2016). Social Categories Alone Are Insufficient to Elicit an In-Group Advantage in Perceptions of Within-Person Variability. *Perception*. DOI: 10.1177/0301006617699226.

Short, L. A., Balas, B., & Wilson, C. (in press). The Effect of Educational Environment on Identity Recognition and Perceptions of Within-Person Variability. *Visual Cognition*.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification*. San Francisco: W.H. Freeman and Company.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law And Human Behavior*, *25*(5), 459-473.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17*(1), 99-139.

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105-118.

Technical Working Group for Eyewitness Evidence. (1999). Eyewitness evidence: A guide for law enforcement. Washington, DC: U.S. Department of Justice.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455-460.

Turati, C., Bulf, H., & Simion, F. (2008). Newborns' face recognition over changes in viewpoint. *Cognition*, *106*(3), 1300-1321.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43*(2), 161-204.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, *69*(10), 1996-2019.

Wells, G. L. (2006). Eyewitness identification: Systemic reforms. *Wisconsin Law Review*, *2*, 615-643.

Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist*, *55*(6), 581.

White, D., Rivolta, D., Burton, A. M., Al-janabi, S., & Palermo, R. (2017). Face matching impairment in developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology*, *70*(2), 287–297.