Unfamiliar face matching, within-person variability, and multiple-image arrays

Sandford, A.1, & Ritchie, K. L.2

1Psychology Department, University of Guelph-Humber, Canada

2School of Psychology, University of Lincoln, UK

Corresponding author:

Dr. Adam Sandford

Psychology department, University of Guelph-Humber

207 Humber College Blvd., Toronto, ON M9W 5L7, Canada

[adam.sandford@guelphhumber.ca](mailto:adam.sandford@guelphhumber.ca)

**Abstract**

Human unfamiliar face matching is error-prone, but some research suggests matching to multiple-image arrays instead of single images may yield improvements. Here, high or low variability arrays containing one, two, and three images, and a target image from the high and low variability image sets were displayed. Arrays were presented simultaneously or sequentially, and the target image was presented simultaneously with arrays or sequentially after arrays, in three experiments. Benefits from exposure to multiple images of the same person required simultaneous viewing of images and improvements were observed in match trials only. Only sequential viewing of a multiple-image array followed by a high variability target image enhanced overall accuracy across trial types, particularly for high variability arrays. Accuracy was highest when the target image and array items were visually similar. Results show the importance of image similarity, and suggest variability is most helpful when array and target are presented sequentially.

**Keywords:** unfamiliar, face, matching, variability

**Introduction**

Photo-ID is a common method for proving our identity, yet the person employed to check photo-ID is almost always unfamiliar with the person pictured. This is problematic as research has shown most people are poor at unfamiliar face matching, that is deciding whether two different photographs present the same person or not (e.g. Bruce, Henderson, Greenwood, Hancock, Burton, & Miller, 1999; Bruce, Henderson, Newman, & Burton, 2001; Megreya & Burton, 2006; Ritchie, Smith, Jenkins, Bindemann, White, & Burton, 2015). This is even true when the comparison is made between a person standing in front of you and a photograph (e.g. Davis & Valentine, 2009; Kemp, Towell, & Pike, 1997; Megreya & Burton, 2008; Ritchie, Mireku & Kramer, 2020). It is therefore important both practically and theoretically to understand different factors affecting face matching accuracy. A key factor is likely to be the types of images shown.

Optimal conditions in unfamiliar face matching display the ‘best case’ scenario where faces were photographed in frontal pose, in neutral expression, and in very good lighting, resembling closely conditions we expect in passport photographs under most government regulations. Indeed this is the case in the standardized Glasgow Face Matching Test (GFMT). Images were high quality and taken only minutes apart, yet average performance levels are around 80% (Burton, White, & McNeill, 2010). However, studies have found reduced average accuracies when matching photographs of unfamiliar faces taken months apart compared with images taken minutes apart (Fysh & Bindemann, 2018; Megreya, Sandford, & Burton, 2013; see also Kramer, Mulgrew, & Reynolds, 2018 for results with infant faces).

Recently, there has been an increased interest in using naturally-occurring (or “ambient”) images in studies of face perception (Balas & Pearson, 2017; Balas, Gable, & Pearson, 2019;Burton, 2013; Jenkins, White, Van Montfort, & Burton, 2011; Murphy, Ipser, Gaigg, & Cook, 2015; Ritchie & Burton, 2017; Sutherland, Oldmeadow, Santos, Towler, Burt, & Young, 2013;Zhou & Mondloch, 2016). Non-rigid deformations (e.g., muscular movements, aging), surface reflectance changes from cardiovascular activity, general health, and lighting, and camera characteristics, such as resolution and depth of contrast, guarantee that no two photos of a face are identical. This within-person variability is thought to account for poor unfamiliar face matching ability. For example, different photo-ID images of the same person result in different matching accuracies (Bindemann & Sandford, 2011). In fact, even small changes in images such as the addition of glasses or sunglasses in one image results in poorer face matching performance (Graham & Ritchie, 2019; Kramer & Ritchie, 2016).

Previous studies have shown that exposure to within-person variability is crucial for learning a new identity (Dowsett, Sandford, & Burton, 2016; Matthews & Mondloch, 2018; Murphy, Isper, Gaigg, & Cook, 2015; Ritchie & Burton, 2017), and this has also been applied to face matching. One study found that presenting participants with arrays containing multiple images improved face matching accuracy (White, Burton, Jenkins, & Kemp, 2014). Two-, three-, and four-image arrays produced higher accuracy than one-to-one matching, with no increase in performance between two and four images. Another study showed that comparing a target image to a two-image array comprising two high variability images produced increased face matching performance compared to a low variability two-image array (Menon, White, & Kemp, 2015a). A further study used a different type of matching task to progressively introduce array items, therefore exposing the participant to variability over time (Dowsett, Sandford, & Burton, 2016). Observers were presented with six physical photographs each of three male models and a pile of 30 photo-cards in which to find a different image of the same male model. On a trial-by-trial basis, participants were presented with one photograph of a model and were asked to find the correct image in the photo-card pile. However, the comparison photographs remained on the table, such that on the second trial, participants were able to compare the two photographs of the model with images in the photo-card pile. This continued until participants had all six images from which to make their decision. Accuracy significantly improved between seeing one photograph and six photographs of target models. Two recent studies, however, found no advantage for the presentation of multiple-image arrays over single images in both a lab-based face matching task (Kramer & Reynolds, 2018), and a live face matching task (Ritchie et al., 2020).

One possible explanation for the difference in results of the utility of multiple images between different face matching studies could be that different studies have used more or less variable stimuli. None of the studies mentioned above present a real definition of variability beyond ‘different-looking images of the same person’. The study by Menon and colleagues that compared matching for high and low variability image pairs defined variability by the human perception of similarity between images. The experimenter selected pairs of images in which they considered the face to look very similar, or dissimilar, to form the low and high variability pairs respectively. These image selections were then confirmed by 10 other people providing similarity ratings between the image pairs (Menon et al., 2015a). Two studies of face learning used faces showing neutral expressions as less variable images, and images expressing emotions as more variable images. Here learning was slower and more error-prone with the high variability emotion set of images than the neutral faces (Redfern & Benton, 2017), but testing on novel images showed higher accuracy when participants had learned identities from the expressive compared to neutral set (Redfern & Benton, 2017, 2019). Another technique used in a face learning study was to define variability based on the source of the images. Ritchie and Burton (2017) gathered high variability images of minor celebrities from Google Image searches. These differed in lighting, camera angle, hair style, facial expression, head angle etc. Their low variability images were stills from one video per identity in which the facial expression and head angle could vary, but the images were taken seconds apart in the same lighting conditions, with the same haircut etc. It is possible that different ways of defining variability have led to different amounts of variability being presented in each study. Therefore it is possible that matching studies which do find a multiple image advantage, have used more or less variable image arrays than studies which have not found a multiple image advantage.

Furthermore, it is unclear what role is played by the similarity between the target image and any one image in the array. It is possible that multiple image benefits are found when the target image is more visually similar to one image in the array than the others. Here, we aim to systematically compare face matching accuracy for both high and low variability image arrays and target images. The definition of variability is consistent between experiments, and is based on that used in a face learning study (Ritchie & Burton, 2017). The inclusion of high and low variability targets also allows us to vary the similarity between the target and the array, as low variability images (both array items and target images) are more visually similar to each other than high variability images.

In the present paper, we combine prior research on variability (e.g. Menon et al., 2015a), multiple-image arrays (e.g. White et al., 2014), and the progressive introduction of increasing numbers of images (e.g. Dowsett et al., 2016) in unfamiliar face matching tasks. We implemented a matching task that required participants to compare a target image to a high or low variability array. Across the three experiments presented here, we examined the effect of progressively adding items to the array, presenting array items individually, and presenting the array and the target image sequentially. Based on previous evidence, we expected performance to be better with an increasing number of images in the array (White et al., 2014), and high variability arrays to confer additional benefits in task accuracy (Menon et al., 2015a; cf Kramer & Reynolds 2018; Ritchie et al., 2020).

**Experiment 1a**

In Experiment 1a, we examined face matching with high and low variability arrays of increasing numbers of images in a simultaneous matching task. Participants saw one, two, or three images presented simultaneously with a target image. The target image belonged to either the high or low variability image set, but never showed an identical image shown in the array. We hypothesized that increasing the number of images in the array would lead to an increase in face matching accuracy, and that this effect would be greater for high compared to low variability arrays.

**Methods**

*Participants*

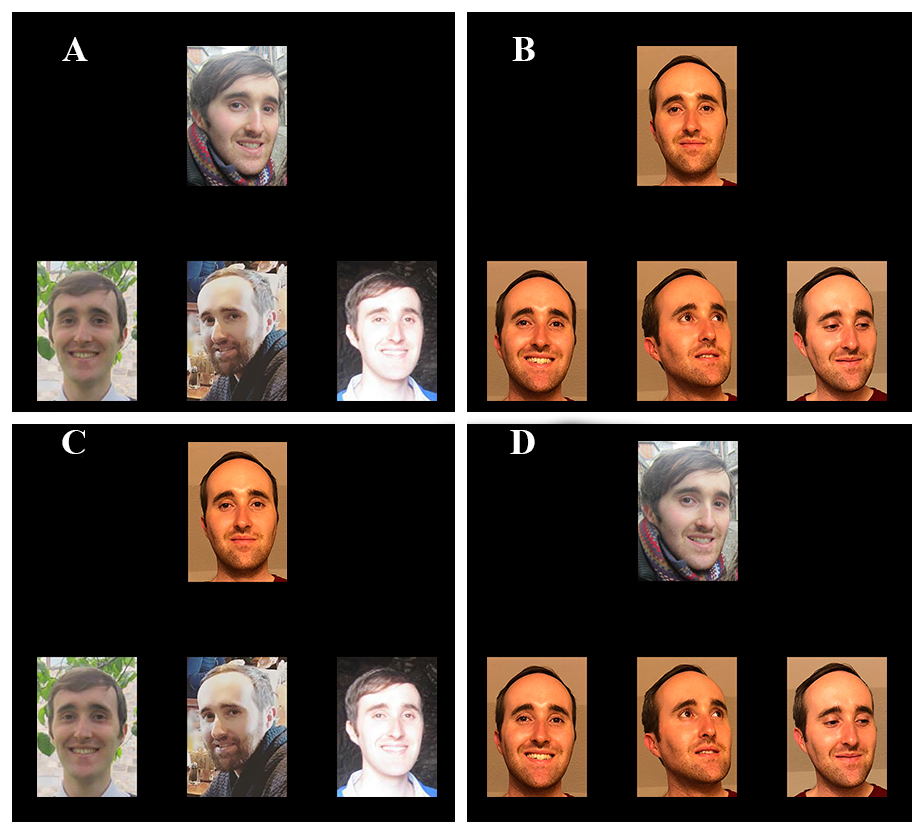
Forty students or staff were recruited from the University of Guelph-Humber and Humber College Institute of Technology and Advanced Learning (average age 20 years, SD 2.39 years, 33 female participants). Participants were compensated with a gift card for their participation. All participants in experiments 1a, 1b, and 2 gave informed consent, and all experiments received approval from the Humber Research Ethics Board.

*Stimuli*

The stimuli showed 32 identities (UK and Australian celebrities) chosen to be unfamiliar to our Canadian participants. We confirmed this with an independent group of participants, but as familiarity is often an idiosyncratic experience, we removed some trials from data analysis following a post-experiment familiarity check (see results section for details). For each identity we found four high variability images and four low variability images. Following Ritchie and Burton (2017), the high variability images were downloaded from Google Images and showed variability in facial expression, head angle, and environmental conditions (e.g., lighting, camera characteristics). Low variability images were stills from interview videos of each identity. These varied in head angle and facial expression, but not in other person-specific dimensions (such as hairstyle or age), or world dimensions (such as lighting or camera characteristics). Background information was kept for the high-variability images (because backgrounds varied in this condition and so were uninformative), but not the low-variability images (to encourage matching based on the face rather than details in the background). We did not remove the background from the high variability images as it was sufficiently different across images so as not to cue identity, and since the high and low variability images were systematically different from each other in terms of the face, we did not consider a systematic difference in background cropping to be important for these studies.

For each identity, one high variability image and one low variability image was selected as target images for match (same identity) trials. An additional 32 foil images (1 for each unique identity) were downloaded from Google Image searches based on a verbal description of each identity. These served as the foils for mismatch (different identity) trials.

All face images measured 200 x 280px and were placed on a black background that measured 800 x 710px. This stimulus display was presented on a monitor measuring 1366 x 768px. Within the stimulus display, the target image always appeared centrally at the top, while the array images appeared at the bottom. One array image appeared at the bottom left, a second appeared at the bottom centre (directly underneath the target image), and a third appeared at the bottom right of the display. The gaps between the left and right contours of the comparison images measured 100px, and the gap between the target and bottom-centre comparison image measured 150px (see Figure 1).

****

**Figure 1.** Example stimuli as seen when all array images presented to participants. **A** high variability target with high variability array, **B** low variability target with low variability array, **C** low variability target with high variability array, **D** high variability target with low variability array. A participant would see the same image in positions one and two in the array. NB the assignment of images to array positions was counterbalanced across participants such that another participant would have seen image two, then images two and three, then images two, three, and one, etc.

*Procedure*

Participants saw six unique blocks. Trials were blocked by the number of images in the array such that all participants saw a block of one-to-one trials, followed by a block of two-to-one trials, and finally a block of three-to-one trials. Each block was then repeated but with the assignment of identities to matching conditions reversed such that each identity which had been seen as a match trial in the first three blocks was now seen as a mismatch trial in the last three blocks. For half of the participants, the first three blocks displayed high variability arrays and the last three blocks displayed low variability arrays. Assignment of array type was reversed for the other half of participants. This counterbalancing meant that each participant did not observe the same identity in match and mismatch trials or in the same array type between the first three and last three blocks. Though the same foil was used for high and low variability arrays, the counterbalancing also meant the same foil was only seen in high *or* low variability array conditions (in other words, the same foil was seen only three times rather than six times by any participant).

We are confident that the repetition of identities does not present a problem for the design, as the image-bound nature of unfamiliar face processing makes it highly unlikely that participants would be able to identify that the same person was shown in two different blocks. In addition, the standardized Glasgow Face Matching Test (Burton, White, & McNeill, 2010) repeats identities, and so this is not unprecedented.

Crucially, the specific images in the array (across blocks) were presented in order for each participant such that a participant would see image one in block one, followed by images one and two in block two and so on (see Figure 1). This was done so that we could measure the benefit of adding specific images, not simply increasing the number of images in the array with different images across conditions. The assignment of images to array positions was counterbalanced across participants. Each block contained an equal number of male and female identities, equal numbers of match and mismatch trials, and each block showed 16 high variability arrays and 16 low variability arrays. Of course, and image in itself cannot be described as being high or low in variability, but we use these terms to describe the target images and single array images in terms of the set of stimuli from which those images were taken. Half of the 16 high variability and 16 low variability arrays were shown with target images from the high-variability set, while the remaining arrays were shown with target images from the low-variability set. In this manner, participants completed trials in the four following conditions: High-variability target and high-variability array; low-variability target and low-variability target; low-variability target and high-variability array; and high-variability target and low-variability array. Identities were counterbalanced across these conditions for each participant with an equal number of male and female identities appearing in each of the four conditions. Participants were informed how many images would be in the arrays at the beginning of each block, and blocks were separated by self-paced breaks. Following the experimental blocks, we checked participants’ familiarity with the identities. Participants were shown the names of all of the identities in the experiment and were asked to indicate whether they were familiar with each person.

**Results of Experiment 1a**

Trials were removed case-by-case based on participant reports of familiarity with any of the 32 identities in the experiment. This resulted in 3% of the total number of trials being removed from data analyses which were conducted only on trials where participants were unfamiliar with faces. We analysed the results of match and mismatch trials separately as these have been shown previously to be uncorrelated processes, and so most previous studies of face matching analyse these separately (see Megreya & Burton, 2006, 2007; Ritchie et al., 2020; White et al., 2014), and analysed the high variability target and low variability target data separately. We used a Bonferroni adjustment on pairwise comparisons making alpha .0167 for three comparisons.

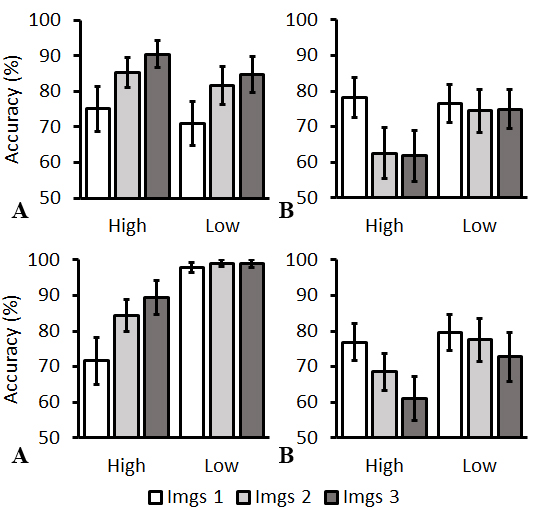
A 2 (target variability: high, low) x 2 (array variability: high, low) x 3 (number of images (1, 2, 3) within-subjects ANOVA on match trials returned a significant three-way interaction, *F*(2, 78) = 6.47, *p* = .003, ηp2 = 0.14. This was due to the high accuracy across number of images in the low target variability, low array variability condition (see figure 2a, bottom row). Therefore, for consistency, we conducted two-way ANOVAs for high variability targets and low variability targets separately.

All data for each experiment reported here have been archived on the Open Science Framework (<https://osf.io/xw7yg/?view_only=f5d1f5302e974748a3aa54ebe1b2944b>). Supplemental analyses for signal detection theory measures of sensitivity and criterion for each reported experiment here can be found at the OSF link above.

*High variability targets*

A 2 (array variability: high, low) x 3 (number of images: 1, 2, 3) within-subjects ANOVA on match trials with high variability targets showed a significant main effect of number of images *F*(2,78) = 34.45, *p* < .001, ηp2 = 0.47, a non-significant main effect of array variability *F*(1,39) = 2.85, *p* = .099, ηp2 = 0.07, and a non-significant interaction *F*(2,78) = 0.20, *p* = .819, ηp2 = 0.01. Performance was more accurate for arrays of two (*M* = 83.42%; *t*(39) = 5.358, *p* < .001, Cohen’s *d* = 0.847) and three images (*M*= 87.63%; *t*(39) = 7.156, *p* < .001, Cohen’s *d* = 1.131) compared with one image (*M* = 73.02%). Performance was also more accurate with arrays of three images compared with two images (*t*(39) = 3.043, *p* = .002, Cohen’s *d* = 0.481) (see Figure 2a, top row).

A 2x3 ANOVA on mismatch trials with high variability targets showed a significant main effect of number of images *F*(2,78) = 13.02, *p* < .001, ηp2 = 0.25, a significant main effect of array variability *F*(1,39) = 8.73, *p* = .005, ηp2 = 0.18, and a significant interaction *F*(2,78) = 9.50, *p* < .001, ηp2 = 0.20. Simple main effects showed an effect of number of images for high variability arrays, *F*(2,156) = 22.45, *p* < .001, ηp2 = 0.22, but not for low variability arrays, *F*(2,156) = 0.31, *p* = .734, ηp2 < .001. Performance was less accurate for arrays of two (*M*= 62.56%) and three images (*M*= 61.77%) compared with one image (*M*= 78.10%; *t*(39) = 5.818, *p* < .001, Cohen’s *d* = 0.920 and *t*(39) = 4.886, *p* < .001, Cohen’s *d* = 0.773). The difference between two and three images was non-significant, *t*(39) = .287, *p* = .388, Cohen’s *d* = 0.045 (see Figure 2b, top row). Simple main effects analyses also showed no difference between array variability with one image, *F*(1, 117) = 0.22, *p* = .640, ηp2 < .01, but accuracy was significantly lower for high variability arrays containing two (*M* = 62.56%) and three images (*M* = 61.77%) compared with low variability arrays containing two (*M* = 74.42%) and three images (*M* = 74.94%) (*F*(1, 117) = 12.04, *p* < .001, ηp2 = .09 and *F*(1, 117) = 14.84, *p* < .001, ηp2 = .11).

****

**Figure 2.** Top row: Accuracy for Experiment 1a with *high* variability targets. Bottom row: Accuracy for Experiment 1a with *low* variability targets. **A** match trials, **B** mismatch trials. Error bars show 95% confidence intervals.

*Low variability targets*

A 2x3 ANOVA on match trials with low variability targets showed a significant main effect of number of images, *F*(2, 78) = 28.48, *p* < .001, ηp2 = .42, a significant main effect of array variability, *F*(1, 39) = 57.71, *p* < .001, ηp2 = .60, and a significant interaction, *F*(2, 78) = 20.00, *p* < .001, ηp2 = .34. Simple main effects analyses showed significant differences between number of images for high variability arrays, *F*(2, 156) = 47.91, *p* < .001, ηp2 = .38, but not low variability arrays, *F*(2, 156) = 0.27, *p* = .764, ηp2 < .01. For high variability arrays, performance was more accurate for arrays of two (*M* = 84.28%) and three images (*M* = 89.44%) compared with one image (*M* = 71.59%; *t*(39) = 4.623, *p* < .001, Cohen’s *d* = 0.731 and *t*(39) = 6.234, *p* < 001, Cohen’s *d* = 0.986), and with arrays of three images compared with two images, *t*(39) = 2.921, *p* = .003, Cohen’s *d* = 0.462) (see Figure 2a, bottom row). Accuracies were also significantly higher for all low variability arrays compared with high variability arrays as shown in simple main effects analyses (one image arrays: *F*(1, 117) = 93.61, *p* < .001, ηp2 = .44; two image arrays: *F*(1, 117) = 29.74, *p* < .001, ηp2 = .20; three image arrays: *F*(1, 117) = 12.35, *p* < .001, ηp2 = .10).

A 2x3 ANOVA on mismatch trials showed a significant main effect of number of images, *F*(2, 78) = 15.86, *p* < .001, ηp2 = .29, a significant main effect of array variability, *F*(1, 39) = 7.71, *p* = .008, ηp2 = .17, and a significant interaction, *F*(2, 78) = 4.02, *p* = .022, ηp2 = .09. Simple main effects analyses showed significant differences between number of images for high variability arrays, *F*(2, 156) = 18.65, *p* < .001, ηp2 = .19, and for low variability arrays, *F*(2, 156) = 3.55, *p* = .031, ηp2 = .04. For high variability arrays, performance was poorer with arrays of two (*M* = 68.54%) and three images (*M* = 60.99%) compared with one image (*M* = 76.82%; *t*(39) = 3.515, *p* < .001, Cohen’s *d* = 0.556 and *t*(39) = 4.825, *p* < 001, Cohen’s *d* = 0.763), and for three images compared with two images, *t*(39) = 2.403, *p* = .011, Cohen’s *d* = 0.380. For low variability arrays, performance was poorer for arrays of three images (*M* = 72.79%) compared with one image (*M* = 79.52%) and two images (*M* = 77.52%; *t*(39) = 2.836, *p* = .004, Cohen’s *d* = 0.438 and *t*(39) = 2.661, *p* = .006, Cohen’s *d* = 0.464). The difference between one image and two images was non-significant, *t*(39) = .864, *p* = .196, Cohen’s *d* = 0.137 (see Figure 2b, bottom row). Simple main effects analyses also showed accuracies were higher for low variability arrays containing two and three images compared with high variability arrays (*F*(1, 117) = 6.98, *p* = .009, ηp2 = .06 and *F*(1, 117) = 12.06, *p* < .001, ηp2 = .09), and no difference between one image arrays, *F*(1, 117) = 0.63, *p* = .429, ηp2 = .01.

**Discussion of Experiment 1a**

The results show that for high variability arrays (with targets from both the high and low variability sets), multiple-image arrays lead to higher accuracy compared to single images on match trials but with an accompanying decrease in accuracy on mismatch trials. When the target comes from the high variability set, low variability arrays also lead to higher accuracy than single images on match trials, but without the accompanying deficit on mismatch trials. Signal detection analyses confirmed no overall benefit of multiple images for high variability arrays for targets coming from either the high or low variability set (see supplemental analyses at the OSF link above).

When the target is from the low variability set, low variability arrays yield higher accuracy than high variability arrays, with no effect of increasing the number of images. Signal detection analyses confirmed an overall benefit of low variability compared to high variability arrays, with no effect of number of images (see supplemental analyses at the OSF link above). This non-significant effect may be due to statistical power (see White et al., 2014).

These results indicate that both the amount of variability in an array, and the similarity between the array and the target image are both important when considering any potential benefit of arrays for face matching. When the array is highly variable, participants tend to respond ‘same’, leading to higher accuracy on match trials but lower accuracy on mismatch trials. However when the array is low in variability, accuracy on match trials increases with increasing numbers of images in the array, but without a corresponding drop in performance for mismatch trials. Matching accuracy is, unsurprisingly, highest when the array and target images are all visually similar (i.e. both from the low variability image set), but crucially here increasing the number of images does not improve performance, suggesting that the effects seen with high variability arrays may be due to increasing similarity between the array items and the target image.

It is possible that the repeating of identities, and indeed images, across blocks in this experiment could have led to participants learning the identities across the blocks, or simply that completing three blocks led to practice effects. It is also possible that the repetition of images (as in image one, then images one and two in the next block etc) could have produced anchoring effects whereby participants were less likely to change their decision each time they were presented with the same identity. Responses did change across blocks with the addition of new images, and so anchoring effects likely cannot explain this effect.

Similar to the findings reported by Dowsett et al. (2016) but with a different methodology, our results suggest exposure to multiple images is beneficial for verifying identity, but only when the arrays and target show the same person (see also White et al., 2014). It is unclear from our results, and the results of prior studies which have reported this effect, whether the multiple image benefit for match trials is due to one of the images in the array (particularly in the high variability conditions) simply looking more visually similar to the target image than the single comparison image. If participants used a strategy to compare each image in the array individually with the target, then one image being more similar to the target could give rise to the observed effect. Alternatively, participants may be gaining a specific benefit from the simultaneous viewing of the array images, such as extracting commonalities between the images in the array and comparing this overall abstraction of the identity with the target image. To investigate these two competing explanations of the multiple-image match benefit, we conducted a second experiment in which each of the images in the array were presented sequentially across blocks to explore whether simultaneous exposure to comparison images is necessary for a multiple-image match advantage. This design also addresses the issue of potential anchoring effects as images are not repeated in the following experiment, and the issue of participants learning identities across blocks, as again the same identities are presented across the blocks.

**Experiment 1b**

In Experiment 1b, a new group of participants completed trials that each contained only a target image and one comparison image as in a one-to-one simultaneous matching task (e.g., Burton et al., 2010). All other aspects of this experiment remained the same as in Experiment 1a. This design utilizes certain features of Dowsett et al.’s (2016) study. Specifically, participants of Experiment 1b could not inspect the variability displayed in an array on any trial and could only rely on one image in the array for determining whether the target showed the same or different person. If simultaneous viewing of comparison images were necessary for a multiple-image advantage, then we would expect to observe reduced benefits of exposure to multiple images across our experiment.

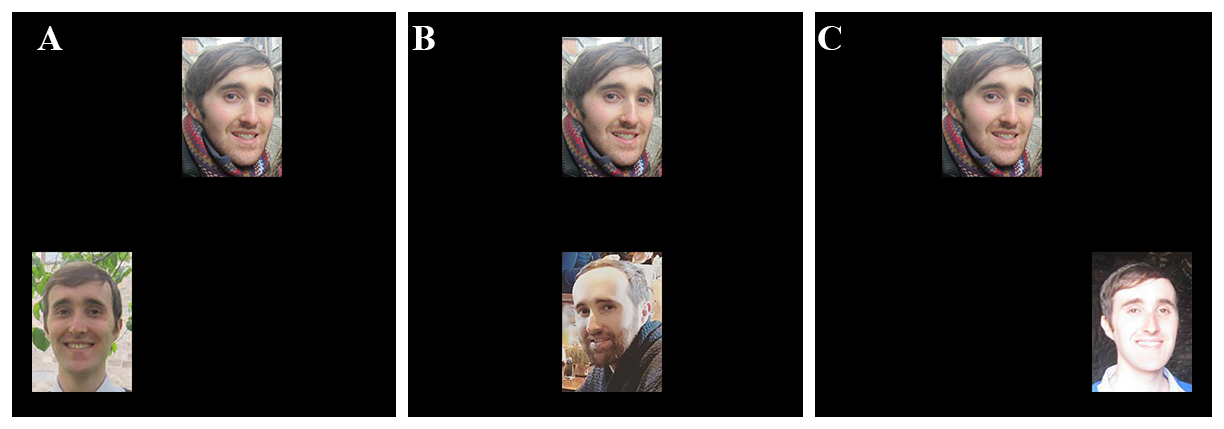
**Methods**

*Participants*

Forty students or staff were recruited from the University of Guelph-Humber and Humber College Institute of Technology and Advanced Learning (average age 30 years, 35 female participants). Participants were compensated with a gift card for their participation.

*Stimuli and Procedure*

The stimuli were the same as those used above, and all aspects of the experiment remained the same, except for the following change. The arrays from Experiment 1a were now presented sequentially to participants, such that a target image and one comparison image appeared in a given stimulus display. The location of the comparison images remained the same as before, meaning that the first array image was seen at the bottom left, the second at the bottom middle, and the third at the bottom right (see Figure 3). Again, all blocks were then repeated with identities seen as a match now seen as a mismatch and vice versa. As in Experiment 1a, the first three blocks showed high or low variability comparison images and the same identities were then shown again in the last three blocks but in the unused variability type of comparison images (e.g., high variability comparison images in blocks 1 to 3, then low variability comparison images in blocks 4 to 6, or vice versa).



**Figure 3.** Presentation of stimuli in Experiment 1b. **A** First comparison image, **B** second comparison image, **C** third comparison image. Example shows high variability target and high variability comparison image. NB stimulus displays were presented between blocks such that one block showed **A**, the next block showed **B**, and the next block showed **C**.

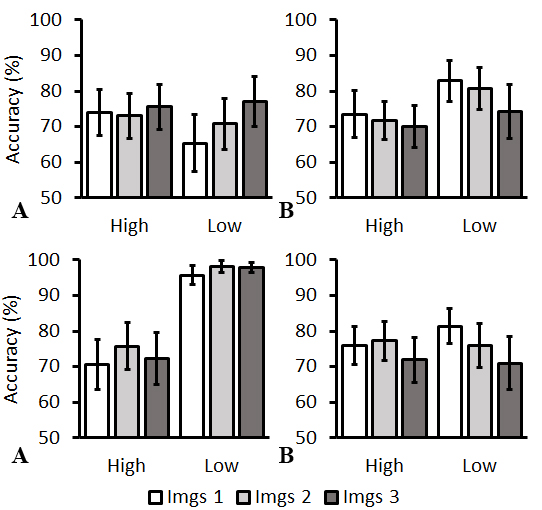
**Results of Experiment 1b**

Again, we checked participants’ familiarity with the identities and removed any trials where participants were familiar with the person pictured on a participant-by-participant basis. This resulted in the removal of 1.7% of the total number of trials across participants from the final data analysis. As before, we used a Bonferroni adjustment for each reported pairwise comparison (α = .0167 for three comparisons).

*High variability targets*

A 2 (array variability: high, low) x 3 (comparison image: 1st, 2nd, 3rd) ANOVA on match trials with high variability targets showed a significant main effect of comparison image, *F*(2,78) = 5.55, *p* = .006, ηp2 = 0.12, a non-significant effect of array variability *F*(1,39) = 1.60, *p* = .213, ηp2 = 0.04, and a non-significant interaction *F*(2,78) = 2.32, *p* = .105, ηp2 = 0.06. Irrespective of array variability, performance was more accurate when the third comparison image was presented with the target (*M* = 76.30%) compared with displays showing the first comparison image (*M* = 69.65%; *t*(39) = 3.282, *p* = .001, Cohen’s *d* = 0.479). The differences in accuracy between displays showing the first comparison image and second comparison image (*M* = 71.9%; *t*(39) = 1.136, *p* = .131, Cohen’s *d* = 0.180), and between displays showing the second comparison image and third comparison image were non-significant, *t*(39) = 2.109, *p* = .021, Cohen’s *d* = 0.333 (see Figure 4a, top row).

A 2x3 ANOVA on mismatch trials showed a main effect of array variability *F*(1,39) = 7.62, *p* = .009, ηp2 = 0.16 with better performance with low (*M* = 79.30%) than high variability array images (*M* = 71.76%). There was also significant main effect of comparison image, *F*(2,78) = 3.65, *p* = .031, ηp2 = 0.09 with a significant decrease in accuracy with the third comparison image (*M* = 72.25%) compared with the first comparison image (*M* = 78.16%, *t*(39) = 2.431, *p* = .010, Cohen’s *d* = 0.384). The differences in accuracy between displays showing the first comparison image and second comparison image (*M* = 76.18%; *t*(39) = 0.891, *p* = .189, Cohen’s *d* = 0.141), and between displays showing the second comparison image and third comparison image were non-significant, *t*(39) = 1.963, *p* = .028, Cohen’s *d* = 0.310 (see Figure 4b, top row). The interaction was also non-significant, *F*(2,78) = 1.20, *p* = .307, ηp2 = 0.03.



**Figure 4.** Top row: Accuracy for Experiment 1b with *high* variability targets. Bottom row: Accuracy for Experiment 1b with *low* variability targets. **A** match trials, **B** mismatch trials, **C** overall accuracy. Error bars show 95% confidence intervals.

*Low variability targets*

A 2x3 ANOVA on match trials with low variability targets showed a significant main effect of array variability, *F*(1, 39) = 103.74, *p* < .001, ηp2 = .73, due to significantly higher match accuracy with low variability array images (*M* = 97.16%) compared with high variability arrays (*M* = 72.88%). The main effect of comparison image and interaction were non-significant, *F*(2, 78) = 1.58, *p* = .213, ηp2 = .04 and *F*(2, 78) = 0.31, *p* = .734, ηp2 = .01 (see Figure 4a, bottom row).

A 2x3 ANOVA on mismatch trials showed a main effect of comparison image, *F*(2, 78) = 6.19, *p* = .003, ηp2 = .14. Mismatch trial accuracy was lower when the third comparison image was displayed with a foil (*M* = 71.46%) compared with displays showing the first (*M* = 78.63%; *t*(39) = 2.932, *p* = .003, Cohen’s *d* = 0.464) and second comparison image (*M* = 76.64%; *t*(39) = 2.856, *p* = .003, Cohen’s *d* = 0.452). The difference in accuracy between displays showing the first comparison image and second comparison image was not significant, *t*(39) = .995, *p* = .162, Cohen’s *d* = 0.157 (see Figure 4b, bottom row). The main effect of array variability and interaction were non-significant, *F*(1, 39) = 0.15, *p* = .701, ηp2 < .01 and *F*(2, 78) = 2.13, *p* = .126, ηp2 = .05.

**Discussion of Experiment 1b**

In this experiment, the only consistent effect was that low variability arrays (irrespective of the number of images in the array) improved performance on match trials when the comparison image was also from the low variability set. This further supports the result of Experiment 1a that the similarity between the array and the target image is important for face matching. All other effects of improvements or decrements in performance between one and three image arrays were small and inconsistent. Specifically, the results did not mirror the results of Experiment 1a, suggesting that the simultaneous viewing of the array images is important for the multiple image effect in face matching. In addition, the results of this experiment go some way to suggesting that the effects in Experiment 1a were not simply due to learning of identities across blocks, or practice effects. Again, here identities were repeated across three blocks but the results differ from the results of Experiment 1a.

While we so far have evidence of a multiple-image advantage in certain conditions, the results of first two experiments have not supported our hypothesis that high variability arrays would confer benefits on performance compared with low variability arrays. High variability arrays in Experiment 1a led to an increase in match but a decrease in mismatch performance. In contrast, low variability arrays comprised of multiple comparison images resulted in greater match trial accuracy compared with one-to-one matching when displayed with a high variability target without cost to mismatch accuracy with increasing number of comparison images. Experiment 1b showed that these effects were not driven by any one image in the array, and therefore resulted from the simultaneous viewing of the array items.

The results of Experiment 1a and 1b do not show a clear overall benefit of multiple images for face matching, differing from some previous studies (Menon et al., 2015a; White et al., 2014). Other more recent studies have, however, failed to find a variability benefit in face matching (Kramer & Reynolds, 2018; Ritchie et al., 2020). It is possible that the simultaneous presentation of the comparison and the target/foil images in face matching tasks affects performance specifically with high variability arrays. Papesh and Goldinger (2014) have previously suggested that the exposure to variability simultaneously may impair participants’ ability to judge between-person differences. Our results, and others on face matching (Kramer & Reynolds, 2018; Ritchie et al., 2020) contrast with studies of face learning which have shown clear benefits of exposure to high variability images compared with low variability images (e.g., Murphy et al., 2015; Ritchie & Burton, 2017). One possible explanation for the difference in results between the learning literature, which shows a clear advantage for exposure to multiple images, and the matching literature in which the multiple-image advantage is not so clear, could be the nature of the tasks. In face learning tasks, the learned images are typically presented over time, and participants are required to remember what each person looks like. This involves memory and the abstraction of a stable representation of each person. In matching tasks, however, there is no memory component as the array and the target image are presented simultaneously. In addition, each image can be compared sequentially to the target image, without the requirement to abstract a stable representation of the person from the array.

In order to test whether having a task with a memory component is important for finding a variability advantage, we conducted a third experiment using a sequential face matching design. Here the array was presented first, followed by the target/foil image.

**Experiment 2**

In Experiment 2, we tested whether presenting the array and target image sequentially would produce a benefit for high variability arrays. Our design resembled the one we used in Experiment 1a insofar as participants were first presented with one comparison image, followed by the target image in the first block, then two comparison images followed by the target image in the second block, and finally three comparison images followed by the target image in the third block. This design forces participants to abstract a representation of the person in the array and store this in short-term memory in order to compare the target image to their stored representation of the person.

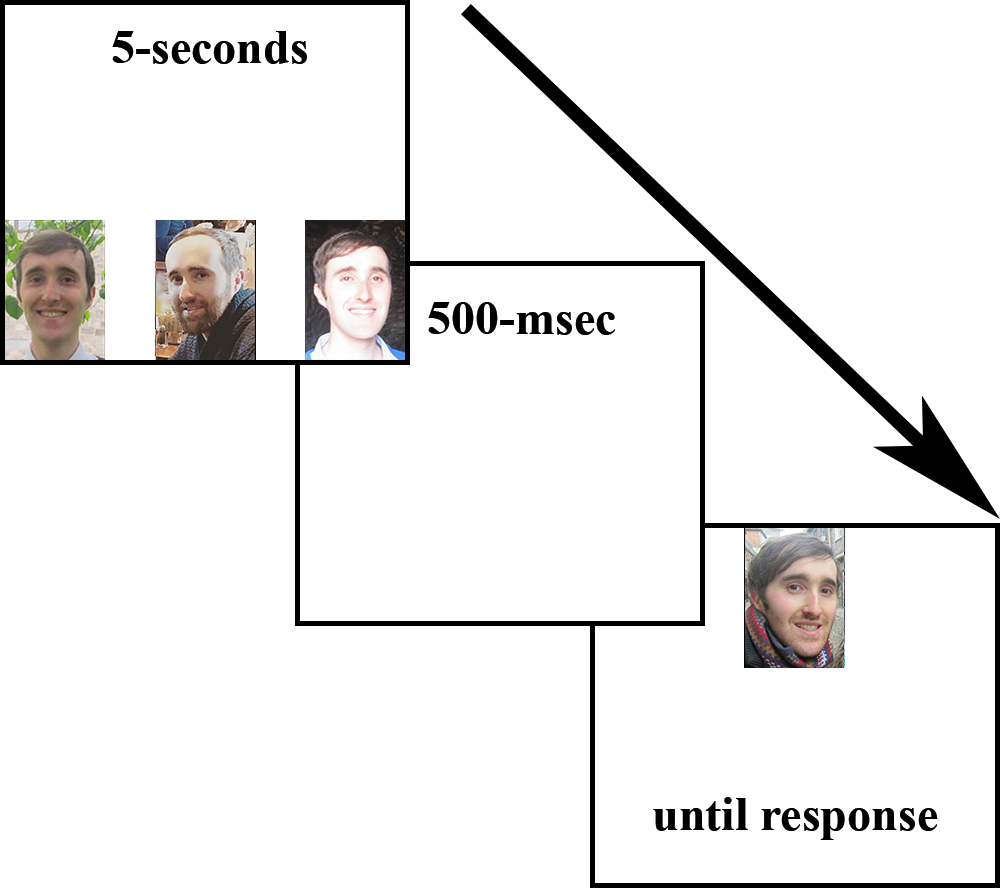
**Methods**

*Participants*

Forty students or staff were recruited from the University of Guelph-Humber and Humber College Institute of Technology and Advanced Learning (average age 23 years, SD 8.03 years, 34 female). Participants were compensated with a gift card for their participation.

*Stimuli and Procedure*

The stimuli and procedure were identical to those used in Experiment 1, but that on each trial participants viewed the array image(s) for 5s followed by a 500ms inter-stimulus interval, followed by the target image which was presented on screen until response (see Figure 3).



**Figure 3.** Example stimulus display for Experiment 2, showing three array images for five seconds, followed by a 500-msec interstimulus interval and a target image until a participant’s response. NB arrays showed one, two, and three images as reported in the method. Arrays showed high variability or low variability images. Targets showed high variability or low variability images. The display showed a black background during the experiment.

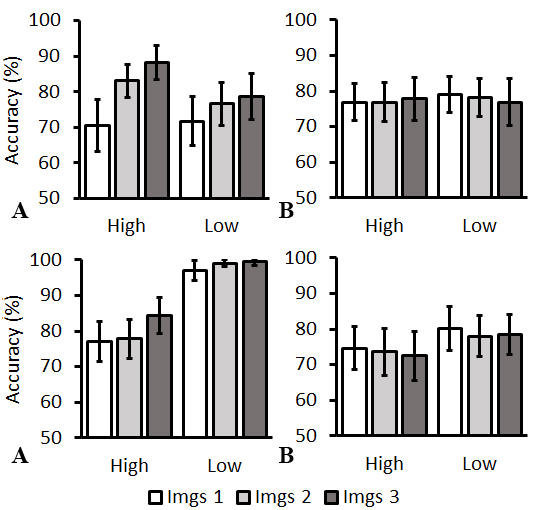
**Results and Discussion of Experiment 2**

Again, we checked participants’ familiarity with the identities and removed any trials where participants were familiar with the person pictured on a participant-by-participant basis. This resulted in the removal of 3% of the total number of trials across participants. As before, we used a Bonferroni adjustment for each reported pairwise comparison (α = .0167 for three comparisons).

*High variability targets*

A 2 (array variability: high, low) x 3 (number of images: 1, 2, 3) within-subjects ANOVA on match trials with high variability targets showed a significant main effect of number of images *F*(2, 78) = 16.41, *p* < .001, ηp2 = .30, a non-significant effect of array variability *F*(1, 39) = 3.44, *p* = .071, ηp2 = .08, and a significant interaction, *F*(2, 78) = 4.57, *p* = .013, ηp2 = .09. Simple main effects analyses confirmed significant differences between number of images for high variability arrays, *F*(2, 156) = 20.17, *p* < .001, ηp2 = .21, and for low variability arrays, *F*(2, 156) = 3.20, *p* = .044, ηp2 = .04. For high variability arrays, performance was better for arrays comprised of two comparison images (*M* = 82.01%; *t*(39) = 3.908, *p* < .001, Cohen’s *d* = 0.618) and three comparison images (*M* = 88.20%; *t*(39) = 5.651, *p* < .001, Cohen’s *d* = 0.893) compared with one comparison image (*M* = 70.55%), and for three comparison images compared with two comparison images, *t*(39) = 2.376, *p* = .011, Cohen’s *d* = 0.376. None of the pairwise comparisons for low variability arrays reached statistical significance (*t*s ≤ 2.185, *p*s ≥ .018, Cohen’s *d*s ≤ 0.345; see Figure 4a, top row). Simple main effects analyses also confirmed significantly higher accuracy for high variability (*M* = 88.20%) compared with low variability arrays containing three images (*M* = 78.73%; *F*(1, 117) = 7.90, *p* = .006, ηp2 = .06). Differences between high variability and low variability arrays comprised of one comparison image and two comparison images were non-significant, *F*(1, 117) = .011, *p* = .741, ηp2 < .01 and *F*(1, 117) = 3.62, *p* = .06, ηp2 = .03.

A 2x3 ANOVA on mismatch trials showed a non-significant main effect of number of images, *F*(2, 78) = 0.04, *p* = .961, ηp2 < .01, a non-significant main effect of array variability, *F*(1, 39) = 0.10, *p* = .754, ηp2 < .01, and a non-significant interaction, *F*(2, 78) = 0.62, *p* = .541, ηp2 = .02 (see Figure 4b, top row). These results suggest no enhancement or cost in mismatch accuracy with increasing number of comparison images in high and low variability arrays. Signal detection analyses showed improved performance for multiple-image arrays compared with one-to-one matching (see supplemental analyses at the OSF link above).



**Figure 4.** Top row: Accuracy for Experiment 2 with *high* variability targets. Bottom row: Accuracy for Experiment 2 with *low* variability targets. **A** match trials, **B** mismatch trials. Error bars show 95% confidence intervals.

*Low variability targets*

A 2x3 ANOVA on match trials with low variability targets showed a significant main effect of number of images, *F*(2, 78) = 6.64, *p* = .002, ηp2 = .15. Performance was better for three arrays comprised of three comparison images (*M* = 91.85%) compared with one comparison image (*M* = 86.96%; *t*(39) = 3.111, *p* = .002, Cohen’s *d* = 0.492) and two comparison images (*M* = 88.42%; *t*(39) = 2.529, *p* = .002, Cohen’s *d* = 0.498). The difference in accuracy between arrays displaying one comparison image and two comparison image arrays was non-significant, *t*(39) = 1.023, *p* = .156, Cohen’s *d* = 0.162. There was also a significant main effect of array variability, *F*(1, 39) = 78.97, *p* < .001, ηp2 = .67, reflecting significantly higher match accuracy with low variability arrays (*M* = 98.44%) compared with high variability arrays (*M* = 79.72%). The interaction was non-significant, *F*(2, 78) = 2.03, *p* = .138, ηp2 = .05 (see Figure 4a, bottom row).

A 2x3 ANOVA on mismatch trials showed a non-significant main effect of number of images, *F*(2, 78) = 0.78, *p* = .462, ηp2 = .02, a non-significant main effect of array variability, *F*(1, 39) = 3.40, *p* = .073, ηp2 = .08, and a non-significant interaction, *F*(2, 78) = 0.18, *p* = .836, ηp2 < .01 (see Figure 4b, bottom row). As with the high variability targets, the results suggest no enhancement or cost in mismatch accuracy with increasing number of comparison images in high and low variability arrays. Signal detection analyses showed better overall performance with three-image arrays compared with two-image arrays but not one-to-one sequential face matching (see supplemental analyses at the OSF link above).

The results of Experiment 2 show a multiple-image advantage in match trial accuracy with no cost to mismatch trial accuracy for high and low variability targets. This contrasts with the results of Experiment 1a which showed enhanced match performance with a cost to mismatch performance. Here, the results support our hypothesis that exposure to multiple images would improve face-matching performance. The task in Experiment 2 introduced a memory component that was not present in Experiments 1a and 1b. The pattern of results between the three experiments suggest a memory component is potentially an important characteristic for a multiple-image advantage.

**General Discussion**

Our results show that viewing multiple images of a target identity leads to increases in match performance and decreases in mismatch performance when the comparison and target images are viewed simultaneously (Experiment 1a), a pattern which is attenuated when participants complete one-to-one matches between a comparison and target image (Experiment 1b). The decreases in mismatch performance observed in Experiment 1a, however, are eradicated when the comparison array is presented before the target in a sequential matching task (Experiment 2). We found no consistent benefit of exposure to high compared to low variability, and instead showed that the visual similarity between the target and array is more important for accurate face matching. This is shown in higher accuracy for low variability arrays with low variability target images. Together, our results suggest exposure to within-person variability requires careful consideration of how array images are presented (simultaneously vs. sequentially) and when target images are presented (simultaneously with array images vs. after presentation of array images), as well as the importance of the similarity between the array items and the target image.

Generally, our results align with previous reports that suggest exposure to multiple images enhance accuracy in match- or target-present trials (e.g., Bindemann & Sandford, 2011; Dowsett et al., 2016; Matthews & Mondloch, 2018), but our hypothesis that exposure to within-person variability in high variability image arrays would enhance (overall) face-matching accuracy was not supported in Experiments 1a and 1b. There are two related reasons for this. Firstly, our participants were exposed to multiple images of the same person in arrays, which shifts focus onto within-person differences (i.e., the similarities and differences between different *images* of the same person) and does not require participants to attend to between-person differences (i.e., the similarities and differences between different *people*), because arrays never contained images of different people. Secondly, match and mismatch accuracy has been shown to be dissociable with separable mechanisms that underlie assessments of face stimuli that show the same person or different people. Therefore, the cues and strategies that participants use to assess images could differ between trials containing the same person or different people, resulting in enhancement of match accuracy and no improvement in mismatch accuracy (e.g., Matthews & Mondloch, 2018).

In Experiment 1a, we displayed the arrays and target images simultaneously, but recent studies have suggested that the variability advantage is greater for sequential presentation of comparison and target images. For example, Menon et al. (2015b) suggested identity-level representations are developed in sequential viewing conditions during a matching task. Specifically, assessing different photographs of the same person was significantly more accurate than assessing photographs of different people when targets are sequentially presented after an array of comparison images. Manipulating whether one person or two people were to be matched with a target did not affect performance in simultaneous viewing conditions, indicating that identity-level abstraction was obtained in the sequential viewing conditions. This identity-level abstraction has been supported in a more recent study (Menon, Kemp, & White, 2018) with pairs of video clips showing the same person or different people. Our findings in Experiment 2 align with previous studies (Menon et al., 2015b; Menon et al., 2018; Ritchie & Burton, 2017) in that presenting high variability comparison images before the target images can enhance face-matching accuracy. In these studies, and in our second experiment, participants are required to remember pertinent facial identity information to make a decision on the target image. We interpret our findings in terms of identity-level abstraction, which functions optimally with high variability images as defined in our experiments (see also Ritchie & Burton, 2017).

Our results could be due to practice effects, but the differences between the results of Experiments 1a and 1b suggest that this is not the case, because participants in Experiment 1b completed three blocks as in Experiment 1a, but we did not observe the same effects. In addition, a separate study showed that simply repeating a matching task across three blocks did not improve performance (Dowsett & Burton, 2015). That study, however, did not repeat identities across blocks as our experiments did, and so we cannot rule out the possibility that participants learned identities across blocks. Again, the differing results between Experiments 1a and 1b go some way towards ruling this out.

Our results do not show consistent effects of exposure to variability on face matching. Instead we showed that an important factor is the similarity between the target and comparison image (shown in our low variability array, low variability target conditions). This is relevant for real-world situations of verification of identity from photo-ID documents where the picture and the person may look very different. Previous studies have shown that face matching accuracy is poorer when matching photographs taken months apart compared with images taken minutes apart (Fysh & Bindemann, 2018; Megreya, Sandford, & Burton, 2013). Our low variability array and target results are align with this.

We are not the first to find that within-person variability may not enhance mismatch accuracy (e.g., White et al., 2014). In our study, we had participants only attend to variability within the same person across three images in matching of unfamiliar faces. By having participants only attend to differences and similarities between different *images* of the same person, differences and similarities between different *identities* are unattended (Papesh & Goldinger, 2014). Future research should consider whether reductions in mismatch accuracy, particularly when comparison and target images are simultaneously presented, can be reversed by varying between-person variability (e.g., high vs. low similarity in visual appearance between people in arrays). However, this might come at the cost of match accuracy, and so a combination of attending to within-person and between-person variability might produce the ideal outcome of enhancement in match *and* mismatch accuracy.

Our results suggest a possible mechanism to explain the differences between the face learning and face matching literature to date. Whereas the learning literature has shown a benefit for exposure to variability (e.g. Murphy et al., 2015; Ritchie & Burton, 2017), evidence from the matching literature has been more mixed with some studies reporting a benefit (Menon et al., 2015a; White et al., 2014) and others not (Kramer and Reynolds, 2018; Ritchie et al, 2020). Those matching studies all used simultaneous matching paradigms whereby all array images and the target were presented together on screen. Learning tasks, on the other hand, present the learned images first, followed by a test. Here we found a clear advantage for exposure to multiple images only when the array preceded the target, thus introducing a memory component seen in the learning literature but not typically in the matching literature. We suggest that including a memory component in the task forces participants to abstract a representation of the identity to compare to the target image. This benefits from multiple images in a way that simultaneous matching tasks do not, as they offer the ability to compare each image individually to the target.

The focus of our current study was on whether exposure to multiple (high variability) images in a comparison array enhances performance on matching tasks. In doing this we found evidence of enhanced match trial accuracy at a cost to mismatch trial accuracy. Our last experiment utilized an important feature of learning literature (the memory component), which showed benefits in matching from multiple image arrays without a cost to mismatch trial accuracy. This provides initial evidence that forced abstraction of identity, which occurs during our sequential presentation of array then target/foil, might explain the discrepancy between the learning and matching literatures. However, our interpretation of this result is limited by not tasking participants with simultaneous *and* sequential matching within the same experiment. Future studies should further investigate the role of memory in face matching and face learning from multiple images of the same faces.

**Acknowledgements:**

We would like to thank Diane Lamila, Ivan Wong, Malvika D’Costa, Sandra Drozdz, and Victoria Vella for their assistance with data collection.

**Funding:**

This work was supported by the University of Guelph-Humber under Grant number 004 awarded to AS and KLR, and by a University of Lincoln College of Social Science Research Fund grant awarded to KLR.

**Data availability and supplemental materials statement:**

All data for experiments and supplemental analyses have been uploaded to OSF with DOI 10.17605/OSF.IO/XW7YG (<https://osf.io/xw7yg/?view_only=f5d1f5302e974748a3aa54ebe1b2944b>)

**Disclosure statement:**

The authors declare no competing interests with respect to the publishing of this paper.

**References**

Balas, B., Gable, J., & Pearson, H. (2019). The effects of blur and inversion on the recognition of ambient face images. *Perception*, *48*(1), 58-71. DOI: 10.1177/0301006618812581

Balas, B., & Pearson, H. (2017). Intra- and extra-personal variability in person recognition. *Visual Cognition*, *25*(4-6), 456-469. DOI: 10.1080/13506285.2016.1274809

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

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339-360. DOI: 10.1037/1076-898X.5.4.339

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7,* 207–218. DOI: 10.1037//1076-898X.7.3.207

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.DOI: 10.1080/17470218.2013.800125

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286-291. DOI: 10.3758/BRM.42.1.286

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*(4), 482–505. DOI: 10.1002/acp.1490

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*(3), 433-445. DOI: 10.1111/bjop.12103

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

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology*, *109*(2), 219-231. DOI: 10.1111/bjop.12260

Graham, D. L., & Ritchie, K. L. (2019). Making a spectacle of yourself: The effect of glasses and sunglasses on face perception. *Perception, 48*(6), 461-470. DOI: 10.1177/0301006619844680

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

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11,* 211-222. DOI: 10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-O

Kramer, R. S. S., Mulgrew, J., & Reynolds, M. G. (2018). Unfamiliar face matching with photographs of infants and children. *PeerJ*, *6*, e5010. https://doi.org/10.7717/peerj.5010

Kramer, R. S. S., & Reynolds, M. G. (2018). Unfamiliar face matching with frontal and profile views. *Perception*, *47*(4), 414-431. DOI: 10.1177/0301006618756809

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising Superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology, 30*, 841-845. DOI: 10.1002/acp.3261

Matthews, C. M., & Mondloch, C. J. (2018). Finding an unfamiliar face in a line-up: Viewing multiple images of the target is beneficial on target-present trials but costly on target-absent trials. *British Journal of Psychology*. doi: 10.1111/bjop.12301

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876. DOI: 10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*(7), 1175-1184. DOI: 10.3758/BF03193954

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), 364-372. DOI: 10.1037/a0013464

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*(6), 700-706. DOI: 10.1002/acp.2965

Menon, N., Kemp, R. I., & White, D. (2018). More than a sum of parts: Robust face recognition by integrating variation. *Royal Society Open Science*, *5*(5), 172381. DOI: 10.1098/rsos.172381

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

Menon, N., White, D., & Kemp, R. I. (2015b). Identity-level representations affect unfamiliar face matching performance in sequential but not simultaneous tasks. *Quarterly Journal of Experimental Psychology, 68*(9), 1777-1793. DOI: 10.1177/0301006615599902

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

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception, & Psychophysics*, *76*(5), 1335-1349. DOI: 10.3758/s13414-014-0630-6

Redfern, A. S., & Benton, C. P. (2017). Expression dependence in the perception of facial identity. *i-Perception, 8*(3), 1-15. DOI: 10.1177/2041669517710663

Redfern, A. S., & Benton, C. P. (2019). Representation of facial identity includes expression variability. *Vision Research, 157,* 123-131. DOI: 10.1016/j.visres.2018.05.004

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

Ritchie, K. L., Mireku, M. O., & Kramer, R. S. S. (2020). Face averages and multiple images in a live face matching task. *British Journal of Psychology*, *111*(1), 92-102. DOI: 10.1111/bjop.12388.

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D. & Burton, A. M.  (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition, 141,*161-169. DOI: 10.1016/j.cognition.2015.05.002

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

White, D., Burton, A. M., Jenkins, R., & Kemp, R. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied*, *20*(2), 166-173. DOI: 10.1037/xap0000009

Zhou, X., & Mondloch, C. J. (2016). Recognizing “Bella Swan” and “Hermione Granger”: No own-race advantage in recognizing photos of famous faces. *Perception*, *45*(12), 1426-1429. DOI: 10.1177/0301006616662046