We know from previous research that unfamiliar face matching (determining whether two simultaneously presented images show the same person or not) is very error-prone. A small number of studies in laboratory settings have shown that the use of multiple images or a face average, rather than a single image, can improve face matching performance. Here, we tested 1,999 participants using four-image arrays and face averages in two separate live matching tasks. Matching a single image to a live person resulted in numerous errors (79.9% accuracy across both experiments), and neither multiple images (82.4% accuracy) nor face averages (76.9% accuracy) improved performance. These results are important when considering possible alterations which could be made to photo-ID. Although multiple images and face averages have produced measurable improvements in performance in recent laboratory studies, they do not produce benefits in a real-world live face matching context.

We rely on photographic identification on a daily basis, yet it is well established that face matching performance, that is, telling whether two images show the same person or not, is poor for unfamiliar faces (Bruce, Henderson, Newman, & Burton, 2001; Ritchie et al., 2015; White, Burton, Jenkins, & Kemp, 2014). So why do we rely so heavily on our face to prove our identity? Recognizing familiar faces is easy (Ritchie et al., 2015; White, Burton, et al., 2014), even from very low-quality images (Bruce et al., 2001; Hole, George, Eaves, & Rasek, 2002), and so it may be that we tend to over-estimate our ability at unfamiliar face matching.

The reason behind poor unfamiliar face matching performance could be the fact that images of the same person can vary significantly (Burton, 2013; Burton, Kramer, Ritchie, & Jenkins, 2016). People can look very different from one image to the next; even simply putting on a pair of glasses creates enough variability between two images to impair face matching performance (Kramer & Ritchie, 2016). One study found different matching rates for different photo-ID images of the same person (Bindemann & Sandford, 2011), suggesting that some images simply provide a better ‘likeness’ than others (Ritchie, Kramer, & Burton, 2018). Another study demonstrated that images picked out by the models themselves produced lower accuracy on a face matching task than images picked out by people unfamiliar with the models (White, Burton, & Kemp, 2016), highlighting that our impressions of how we look do not align with the impressions of strangers. With familiar people, however, it seems that we can cope with a large range of variability, and increasing familiarity correlates with increasing likeness ratings for variable images of the
same person (Ritchie et al., 2018). Put simply, once familiar with someone’s face, any image of them is judged to be a better likeness, compared with unfamiliar raters.

While variability can impair face matching, there is converging evidence that it helps with face learning (Dowsett, Sandford, & Burton, 2016; Longmore, Liu, & Young, 2008; Longmore et al., 2017; Murphy, Ipser, Gaigg, & Cook, 2015). A recent study showed that seeing a high variability set of photographs of the same person at encoding (collected from Google Images) gave rise to higher subsequent recognition and face matching accuracy than seeing several similar images (low variability still frames taken from a single interview video) at encoding (Ritchie & Burton, 2017). This result suggests that exposure to within-person variability helps with later recognition, which is supported by recent computational modelling work (Kramer, Young, & Burton, 2018).

Two studies have suggested that variability may also be used to improve face matching (Menon, White, & Kemp, 2015; White, Burton, et al., 2014). Participants are shown a number of images of a person and are told that these images all show the same person. They are then asked to compare this array of images to a new image, and are asked if the new image also shows the same person. Groups of two, three, and four images produced better performance than single images, with no increase from two to four images (White, Burton, et al., 2014), and a single image was more easily matched to a high variability pair of images compared with two similar images (Menon et al., 2015). Additionally, a recent study showed that when we are informed that multiple exposures portray the same person, as opposed to different people, variability improves performance on a matching task (Menon, Kemp, & White, 2018). However, other recent work found no advantage of presenting two pairs of images of a person (specifically, two different ‘frontal and profile view’ pairs) on face matching accuracy (Kramer & Reynolds, 2018). Therefore, there is some conflicting evidence regarding the utility of multiple images in face matching.

In addition to providing multiple images, another candidate for improved matching performance is an average image comprising multiple photographs of the same person. Averages have been shown to produce more accurate computer face recognition than single images (Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2011; Robertson, Kramer, & Burton, 2015), even when the averages are derived from pixelated images (Ritchie et al., 2018). Face averages have also been shown to help with human face recognition when the averages comprise several ambient images (White, Burton, et al., 2014) and when facial composites are averaged together (Bruce, Ness, Hancock, Newman, & Rarity, 2002; Hasel & Wells, 2007). Evidence suggests that people may form an average as an internal representation when shown an array of images of a new identity (Kramer, Ritchie, & Burton, 2015). However, this reported improvement in matching with face averages was not replicated in other work, where no overall benefit was found in human performance (Ritchie et al., 2018).

One study on eyewitness testimony investigated the ‘live superiority hypothesis’ (Fitzgerald, Price, & Valentine, 2018). This is the widely held belief that live lineups comprising real people standing in front of participants should produce better identification accuracy than photograph lineups using images on a computer screen. The authors, in fact, concluded that live lineups do not give rise to higher accuracy. We believe that the same live superiority hypothesis may exist with face matching, whereby most people believe it would be easier to decide whether a photograph matches a live person as opposed to comparing two photographs. Very few studies have tackled face matching using live faces, presumably due to the logistical difficulties of ensuring a model is physically present during testing. In line with the conclusions of Fitzgerald et al. (2018), one study showed that participants were poor at picking out a person they had seen live
from a photograph lineup, whether the lineup was presented immediately after, or simultaneous with, the live person (Megreya & Burton, 2008). The same study showed in a final experiment that participants were no better at matching a live person to a photograph (29.9 correct responses from a possible 36) than matching two photographs (30.4 correct responses from a possible 36). A subsequent study also found poor performance when matching a live person to CCTV video footage (Davis & Valentine, 2009). Kemp, Towell, and Pike (1997) showed that a group of supermarket cashiers performed strikingly poorly at a ‘photo-ID to live face’ matching task. Even in the ‘easiest’ condition, in which the person standing in front of them may have worn the same clothes as on the photo-ID card, there were almost 7% errors. A more recent study of passport officers showed that they made an average of 10% errors in a task comparing photographic images to live people (White, Kemp, Jenkins, Matheson, & Burton, 2014). These studies suggest that photograph-to-live face matching is error-prone.

The current study uses multiple images and face averages in a live face matching task. Students acted as models and approached strangers on campus, showing them photographs and asking the question ‘is this me?’. In Experiment 1, we considered whether a four-image array would produce an increase in accuracy in comparison with a single image. In Experiment 2, we compared an average image to this single image baseline. Following from previous research, we predict that both multiple images and average images will give rise to higher accuracy than individual images.

Experiment 1: Multiple images

Method

Participants

Models. Twenty-four students acted as models for the first experiment (seven men; 23 self-reported White; mean age: 19.5 years, range: 18–21 years).

Judges. We recruited a large sample of 959 participants for Experiment 1 (355 men; mean age: 22.2 years, range: 17–63 years; 94.6% of participants were self-reported White).

All models and judges in both experiments were members of a UK university. Models participated as part of their research skills course, while judges represented an opportunity sample of students and staff that were present on campus at the time of data collection. Judges were strangers and did not know the models prior to recruitment. This study was approved by the School of Psychology Research Ethics Committee (ethics number PSY171881). All participants gave written informed consent.

Stimuli and procedure

Each model provided four images of themselves and of a foil that was chosen to match the same verbal description as them (e.g., similar age and appearance, same sex and ethnicity). Foil identities were either friends of the models or celebrities from foreign countries, chosen to be unfamiliar to our UK judges. All models, and those foils who were friends of the models, provided written consent for their images to be used, and images were downloaded from each person’s social media account. In the cases where the foil was a celebrity, images were downloaded from the Internet using Google Images searches. Celebrity images were publicly available, and no consent was sought. All images
were broadly front-facing but sampled natural variability in facial and environmental parameters, akin to those used in previous face matching research (Ritchie et al., 2015). The images were high quality, and the final stimuli were cropped to 380 × 570 pixels, displaying the head and limited surrounding detail (see Figure 1).

All images were presented on laminated paper, with each individual image measuring 6 cm × 4 cm. The models approached people on campus and stood at a conversational distance. Each judge was shown only one of the four conditions (single image match, single image mismatch, four images match, and four images mismatch), and asked ‘is this me?’. Judges had an unlimited amount of time to respond. For the single image condition, we chose the most neutral, front-facing image of each model and foil in order to emulate a photo-ID image.

Each judge made a single judgement of one image/array in one condition. The image they saw was determined by cycling through all four conditions in order and then repeating this process. Each model collected responses from 40 judges (10 in each condition), resulting in a total of 960 responses. Data collection spanned approximately 2 weeks during the semester. Data from one judge (i.e., one response) were excluded due to experimenter error.

**Results**

Each model collected responses in all conditions, and following the method used previously in a live face matching task (Kemp et al., 1997), we calculated mean accuracy for each condition separately (i.e., for the ten responses collected per condition) for each model.

Figure 2 shows the mean accuracies for Experiment 1. We analysed the data using a 2 (trial type: match, mismatch) × 2 (number of images: one, four) within-subjects (here, models) analysis of variance (ANOVA). There was a significant main effect of trial type, $F(1,23) = 15.28, p = .001, \eta^2_p = .40$, with higher accuracy for match ($M = 89.0\%$) than...
mismatch ($M = 72.4\%$) trials. There was a non-significant main effect of the number of images, $F(1,23) = 1.63, p = .229, \eta_p^2 = .06$, and a non-significant interaction, $F(1,23) = 0.46, p = .831, \eta_p^2 < .01$.

We can also analyse the data using signal detection theory. Here, hits correspond to correct match trials, and false alarms to incorrect mismatch trials. We analysed $d$-prime ($d'$) values using a paired samples $t$-test, which showed no significant difference between sensitivity in the single image ($M = 1.83$) and four-image array ($M = 2.03$) conditions, $t(23) = 1.07, p = .294$, Cohen’s $d = 0.22$. We also analysed criterion ($c$) values as a measure of response bias. A paired samples $t$-test on $c$ values showed no significant difference in bias in the single image ($M = 0.27$) and four-image array ($M = 0.25$) conditions, $t(23) = 0.29, p = .776$, Cohen’s $d = 0.06$.

Although the above ANOVA follows analyses presented in previous work (Kemp et al., 1997), it fails to fully utilize the large number of responses collected. By averaging across judges for each model, the resulting percentages contributed only single observations to our ANOVA despite each value being derived from ten judges’ responses. As a more powerful method of analysis, we therefore used a multilevel modelling approach, nesting our judges within models. With the dependent variable being whether the judges’ responses were correct or incorrect, we carried out a mixed-effects logistic regression, with trial type, the number of images, and their interaction, as fixed effects. Mirroring our previous analysis, we found a significant main effect of trial type, $F(1,955) = 42.09, p < .001$, with judges being 3.47 (95% CI [2.04, 5.92]) times more likely to respond correctly on match in comparison with mismatch trials. There was a non-significant main effect of the number of images, $F(1,955) = 2.07, p = .150$, and a non-significant interaction, $F(1,955) = 0.06, p = .802$.

Contrary to previous literature (e.g., White, Burton, et al., 2014), we therefore found a numerical but statistically non-significant advantage (with a small effect size) of four images compared to one in a live face matching task. In Experiment 2, we use face averages in another live matching task.

![Figure 2](https://example.com/figure2.png)

Figure 2. Results of the multiple image matching task in Experiment 1. Error bars show the standard error of the mean (SEM).

---

1 We also find the same pattern of results even after excluding our male models and non-White judges. As minorities within this experiment, these groups may represent additional noise in our data set, but fail to explain our lack of a benefit for multiple images.
Experiment 2: Face averages

Method

Participants

Models. A new group of 26 students acted as models for Experiment 2 (one man; 25 self-reported White; mean age: 19.5 years, range: 19–21 years).

Judges. A new group of 1,040 people participated in this experiment (310 men; mean age: 21.3 years, range: 18–61 years; 96.3% of participants were self-reported White).

Stimuli and procedure

Each model provided 12 images of themselves and of a foil that matched the same verbal description as them (e.g., similar age and appearance, same sex and ethnicity). As in

Figure 3. Images of one model in Experiment 2. (a) Single image of the model (match), (b) average of 12 images of the model (match), (c) single image of the foil (mismatch), (d) average of 12 images of the foil (mismatch). The individuals pictured in these images appeared in the experiments and have given permission for their images to be reproduced here. [Colour figure can be viewed at wileyonlinelibrary.com]
Experiment 1, foil identities were either friends of the models or celebrities from foreign countries. The initial images were collected in the same way as those used in Experiment 1.

We created the average images by first deriving the shape of each image using a semi-automatic landmarking system designed to register 82 points on the face aligned to anatomical features, where only five locations are selected manually (for details, see Kramer, Young, Day, & Burton, 2017). Each average was created by warping the 12 images of the model to the average shape of those 12 images, and then calculating the mean RGB colour values for each pixel using the InterFace software package (Kramer, Jenkins, & Burton, 2017). Individual images were additionally cropped so as not to show external features. This was done to minimize the differences between these single images and the averages. See Figure 3 for example stimuli.

The data collection procedure was the same as in Experiment 1, with the images presented at the same size as used above. Each model appeared in all four conditions (single image match, single image mismatch, average match, average mismatch). As in Experiment 1, each judge made a single judgement of one image/average in one condition. The image they saw was determined by cycling through all four conditions in order and then repeating this process. Each model collected responses from 40 judges (10 in each condition), resulting in a total of 1,040 responses. Again, data collection spanned approximately 2 weeks during the semester. For the single image condition, we again chose the most neutral, front-facing image of each model and foil in order to emulate a photo-ID image. Therefore, as in Experiment 1, for each model, the same image was used in the single image condition for all of that model’s judges.

Results

Figure 4 shows the accuracy for each model in each of the four conditions for Experiment 2. A 2 (trial type: match, mismatch) × 2 (image type: single image, average) within-subjects ANOVA showed a non-significant, yet medium-sized main effect of trial type, $F(1,25) = 3.56, p = .071, \eta^2_p = .125$, a non-significant main effect of image type, $F(1,25) = .869, p = .360, \eta^2_p = .034$, and a non-significant interaction, $F(1,25) = 0.02, p = .902, \eta^2_p = .001$. Again, we analysed d-prime and criterion values using paired samples $t$-tests. We found a non-significant difference between d-prime values in the single image ($M = 1.92$) and average ($M = 1.66$) conditions, $t(25) = 0.97, p = .341$, Cohen’s $d = .19$, and a non-significant difference between criterion values in the single image ($M = 0.10$) and average ($M = 0.13$) conditions, $t(25) = .241, p = .812$, Cohen’s $d = .05$.

As in Experiment 1, we also used a multilevel modelling approach, nesting our judges within models. With the dependent variable being whether the judges’ responses were correct or incorrect, we carried out a mixed-effects logistic regression, with trial type, image type, and their interaction, as fixed effects. Here, we found a significant main effect of trial type, $F(1,1,036) = 8.48, p = .004$, with judges being 1.57 (95% CI [1.03, 2.38]) times more likely to respond correctly on match in comparison with mismatch trials. This is reflective of the medium-sized effect reported for this non-significant effect in the ANOVA above. There was a non-significant main effect of image type, $F(1,1,036) = 2.36, p = .125$, and a non-significant interaction, $F(1,1,036) = 0.00, p = .969$.\(^2\)

\(^2\)We also find the same pattern of results even after excluding our male models and non-White judges. As minorities within this experiment, these groups may represent additional noise in our data set, but fail to explain our lack of a benefit for averages.
In contrast with previous work (e.g., White, Burton, et al., 2014), we failed to find an advantage for face averages over single images in our task.

**Discussion**

In two experiments, testing a total of 1,999 participants, we found no advantage for four-image arrays or face averages over single images in live face matching. It has previously been shown that averages (White, Burton, et al., 2014) and multiple-image arrays (Menon et al., 2015; White, Burton, et al., 2014) give rise to better face matching performance. These studies presented images on the computer screen for matching, and so our results argue that this effect is not present when the array or average is presented on paper and the comparison is made with a live face. It is possible that participants gained more information from the live face than the computer images used in previous studies, overriding the beneficial effects of the multiple-image arrays and face averages. A recent study presented participants with either one or two video clips of the same person, with participants being told in the two-clip condition either (truthfully) that both clips showed the same person or (misleadingly) two different people. The results showed that people learned the identities better from two clips than a single clip, and importantly better when the two clips were presented as the same person compared with two different people (Menon et al., 2018). As such, not only do we benefit from exposure to variability, but that benefit increases when we are certain that the images (or videos in this case) show the same person. In our study, the live person was seen in varying poses/expressions as they described the study to participants and there was clearly no question that the observed variability necessarily belonged to the same person. Therefore, participants may have used this variability to their advantage, overriding the potential benefits of the four-image array or the average.

The average and multiple-image array advantages for face matching were both originally reported in the same article (White, Burton, et al., 2014). Here, we replicated their face average methods as closely as possible, for example, by creating the averages from 12 images of each model and cropping the background for both single images and averages. This is important because both computer recognition and human performance on a speeded name verification task have previously been shown to improve as the number of images comprising the average increases (Burton et al., 2005). Despite ensuring that our
face averages comprised the same number of images as the previous study in which an average advantage was found, we did not replicate this result. Another recent article in which averages from 10 images per identity were created also failed to find an average advantage in a one-to-one matching task using high-quality images. For pixelated images, however, creating averages of these did improve overall performance (Ritchie et al., 2018). Therefore, the null effect reported in the current study is not unprecedented in the literature. We note that in the experiments reported here, each model used the same image of themselves or their foil in the single image conditions. We chose the most neutral, front-facing image of each model and foil so as to emulate a photo-ID image, the idea being that multiple images could be added (either around in the case of multiple images or together in the case of averages) to the photo-ID image in a practical setting. We acknowledge that this may not be the ideal way in which to counterbalance image presentation in a computerized experiment. Future research may seek to randomize the selection of the single image so as to obtain responses from more than one image per model.

White, Burton, et al. (2014) found an advantage of averages and multiple-image arrays only for match trials (and hence hits) but not for mismatch trials (correct rejections). We have shown only an effect of more accurate performance on match compared to mismatch trials in Experiment 1, with a weaker effect in the same direction in Experiment 2. Although one interpretation of this pattern of results is that participants were biased in their responses to the ‘is this me?’ question, saying ‘yes’ more often than ‘no’, our results are consistent with other studies of live face matching. Overall accuracy across our single image conditions ($M = 79.9\%$) was within the range of unfamiliar one-to-one matching accuracy reported in previous studies of live face matching ($M = 67.4\%$ – Kemp et al., 1997; $M = 83.1\%$ – Megreya & Burton, 2008). As such, we can be confident that the images and foils used here did not result in overly easy or difficult matching.

Our results have shown that neither multiple images nor face averages significantly improve unfamiliar face matching in a live matching task. These results are contrary to previous findings (White, Burton, et al., 2014), yet are supported by more recent studies which also failed to find an overall average advantage with high-quality images (Ritchie et al., 2018) or a multiple image advantage (Kramer & Reynolds, 2018). That we demonstrate no benefits when using these two methods is perhaps surprising, given the current thinking in this field that such techniques should produce gains in performance over single image comparisons (e.g., White, Burton, et al., 2014). The evidence presented here suggests that either these improvements in performance are not sufficiently robust to be practically useful and/or they fail to generalize to live matching contexts, again limiting their use. Taken together, our results show that multiple images and face averages do not improve unfamiliar face matching in a live comparison.

**Acknowledgements**

We would like to thank our ‘Research Skills 2’ and ‘Research Skills III’ undergraduate students for acting as models and for collecting the data.

**References**


Received 12 March 2018; revised version received 12 November 2018.