Within-person variability promotes learning of internal facial features and facilitates perceptual discrimination and memory

Elliott Robins[1], Tirta Susilo[1], Kay Ritchie[2], Christel Devue[1]*

[1]School of Psychology, Victoria University of Wellington, New Zealand

[2]Universtiy of Lincoln, United Kingdom

*Corresponding author: Christel Devue, School of Psychology, Victoria University of Wellington, PO Box 600, Wellington 6040, New Zealand

Phone: +64(0)4 463 5898

Email: christel.devue@vuw.ac.nz

Abstract

Recent research indicates that exposure to within-person variability is essential for developing robust representations of new faces. For example, people perform better on a face matching task after exposure to highly variable photos, compared to less variable photos. However, the specific aspects of face processing that benefit from variability remain unclear. We investigated whether within-person variability improves the ability to match and recognise individual faces, and whether it promotes learning of internal facial features. In one exploratory and one confirmatory experiment, we tested matching and recognition performance of participants after they learned 4 individual faces in a high variability condition and another 4 in a low variability condition. Further, to assess if variability promotes robust learning of invariant facial features (e.g., eyes, nose), we compared performance with and without external facial features (full headshots vs. cropped images showing only internal features). We found a large benefit of variability in the recognition task, and a smaller effect on the matching task, but the size of the benefit was comparable with and without the presence of external features. Therefore, within-person variability improves a variety of face recognition skills, and it encourages the encoding of stable internal facial features.

Keywords: face learning, within-person variation, perception, memory, ambient images

Within-person variability promotes learning of internal facial features and improves perceptual discrimination and memory

We are very good at recognising familiar faces, and yet unfamiliar face recognition is error-prone. We are able to recognise a familiar face across a wide range of variability in lighting, head angle, facial expression, and image quality (Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999; Johnston & Edmonds, 2009; Megreya & Burton, 2007). Such variability, however, gives rise to poor recognition rates with unfamiliar faces (Bruce et al., 1999; Megreya & Burton, 2008), even when two unfamiliar faces are presented simultaneously in a simple matching task (Burton, White, & McNeill, 2010; Kramer & Ritchie, 2016; Megreya & Burton, 2006; Ritchie et al., 2015). This difference in performance with familiar versus unfamiliar faces raises the key question of how a face becomes familiar, and what face processing skills benefit from exposure to variability.

Earlier research manipulated exposure to viewpoint and illumination, and showed that learning of new faces was viewpoint- and illumination-specific (Longmore, Liu & Young, 2008; Liu, Bhuiyan, Ward & Sui, 2009). Recent research has focused on the use of naturally-occurring images termed “ambient images”. Ambient images allow for a high degree of within-person variability (e.g., in terms of viewpoint and lighting, but also expression, hairstyle, facial hair, make-up, adiposity or age) to be shown during learning, which has been argued to be important for developing familiarity (Jenkins, White, van Montfort & Burton, 2011; Burton, Kramer, Ritchie, & Jenkins, 2016). A growing body of evidence suggests that exposure to within-person variability helps us to build robust representations of new people (Andrews, Jenkins, Cursiter, & Burton, 2015; Baker, Laurence, & Mondloch, 2017; Dowsett, Sandford, & Burton, 2016;

Laurence & Mondloch, 2016; Murphy, Ipser, Gaigg, & Cook, 2015), but the specific aspects of face processing that benefit from variability are not well understood.

A study by Ritchie and Burton (2017) showed that exposure to multiple high variability images led to more accurate learning than exposure to the same number of images showing less variability. High variability images were taken from a Google Image search. Each identity was photographed on different occasions with different cameras, in different lighting conditions, presenting different facial expressions, head angles and appearance. Low variability images were taken from one interview video per identity, and so could only vary in expression and head angle. Identities learned from high variability images were processed more efficiently in subsequent tests than those learned from low variability images, both across a speeded name verification task and a face matching task. In another study using an unsupervised and incidental learning procedure, Murphy and colleagues (2015) showed that variability improves face memory. Participants were exposed to 48-item arrays consisting of 6 images of 8 identities, and they had to determine the number of identities present. Some participants saw the same 6 images of each identity throughout learning, others saw 6 unique images every trial. Estimates of the number of identities improved similarly in the two conditions, but participants who saw unique images recognised more faces in a subsequent recognition test. Taken together, both studies suggest that within-person variability can facilitate a wide range of face recognition skills, but because they used different experimental procedures, a direct comparison of the benefit across the different skills is not possible.

It is also unclear whether exposure to variability results in the encoding of a large amount of variation in someone's appearance, or whether it promotes learning of internal facial

features (e.g., eyes, nose, or mouth). Focusing learning on internal facial features might be advantageous because these features are stable over time and across images, and so they constitute more reliable identity cues. Consistently, seminal studies have shown that while unfamiliar faces are mostly processed based on external features, familiar faces are processed based on both internal and external features (Ellis, Shepherd, & Davies, 1979; Young, Hay, McWeeny, Flude, & Ellis, 1985). Moreover, recent data from a card sorting task (i.e., in which participants sort a deck of images of faces into separate identities) suggest that, regardless of familiarity, external features alone convey less identity information than internal features alone, or internal and external features combined (Kramer, Manesi, Towler, Reynolds, & Burton, 2018). Crucially, images used in Murphy and colleagues' (2015) recognition test showed internal features only. Thus it seems that exposure to variability has promoted the encoding of these stable diagnostic features. However, performance with headshots was not tested and it is uncertain whether and how the removal of external features might have affected performance.

Here, we sought to characterise how variability benefits face processing in two ways. First, we measure the extent to which exposure to variability during learning improves face processing across simultaneous matching and recognition tasks. Second, we assess whether exposure to variability promotes learning of internal features of the face. In two experiments, participants learned 8 faces from headshots with high or low within-person variability. We tested face perception and face memory performance with novel images not shown during learning. To assess whether exposure to variability focuses learning on internal facial features, test images showed internal features only (Experiments 1 and 2) and performance with headshots, that conserve external features, was compared (Experiment 2). If exposure to

variation promotes a focus on stable facial features, participants should show comparable benefit of variability when tested with full headshots or cropped pictures showing internal features only. But if the benefits of variability do not rely on developing robust representations based on invariant features of the face, then they should be restricted to headshots that also include external features.

Experiment 1

Method

Participants. Our experimental design was adapted from Ritchie and Burton (2017). Effect size calculation (performed with GPower) suggests that 35 participants are required to replicate the original effect found in the matching task (i.e., based on accuracy in match trials of Experiment 2, η^2 = .28, in Ritchie and Burton, 2017) with .95 power. To anticipate for data loss, we recruited 56 MTurk workers, located in the US (26 women, 30 men; mean age = 36.28 years ± 12.27). The study was approved by the School of Psychology Ethics Committee at Victoria University of Wellington.

Stimuli. We used stimuli from Ritchie and Burton (2017, Experiment 2). They were faces of 10 Australian celebrities, unfamiliar to our participants. To avoid that some uncontrolled variations present in the low variability (LV) condition sets (e.g., facial expressions resulting from speech during interviews)[1] were not represented in the high variability (HV) sets (i.e.,

[1]Because they were screenshots from video interviews, some images in the LV set were of lesser quality than some images in the HV set for some celebrities (N = 6), as reflected by image size in KB, $Mean_{HV}$ = 33.76 KB ± 3.11, $Mean_{LV}$ = 26.08 ± 4.56. However, data for headshots and inner features (Experiment 2) reported below showed the

posed expression), we selected three images per identity with the most neutral expressions in the LV set and included these three images in the HV set[2]. In the HV condition, we also selected 9 images from the original HV set, giving 12 different images in total. As a result of this selection, variability would be higher in the HV set than in the LV set on every possible dimension. In order to counterbalance gender across learning conditions, we used only 4 men and 4 women from the original set. We created two reviews slides measuring 1280 x 720 pixels for each identity. Slides showed 12 unique images in the HV condition, and 3 unique images repeated 4 times in the LV condition so that images cover the same surface in both conditions. Stimuli are illustrated on **Figure 1**.

The images used at test were all novel and edited to show only internal features within a circle; the rest was concealed with a grey mask. We used three unique images per identity in the recognition task, and 6 unique images in the matching task (4 for match trials and 2 for mismatch trials). In addition, 24 new images of different individuals were used as foils (12 male, 12 female) during the recognition task, and 32 others as unlearned faces or foils for the matching test. All images used across the experiment were resized to 380 x 570 pixels.

Procedure. Participants completed the experiment online via Testable.org.

---

same patterns with the two celebrities whose LV images were of similar quality as in the HV condition, although recognition accuracy was numerically higher in *both* the HV and the LV conditions. Higher accuracy across the two learning conditions suggests that the faces of these two specific persons (i.e., Dave Hughes and Hamish Blake) may be easier to remember than that of other people in the set. This rules out the possibility that lower quality of some images in the LV set alone accounts for the patterns of results.

[2]Note that this is also why we opted for repeating the three LV images four times in the LV condition (see Procedure below) instead of using a different image on each trial like in the original study by Ritchie and Burton (2017). Using a different image on each trial would make it impossible to include the LV set in the HV set without changing the timing of presentation or the total duration of exposure to each identity.

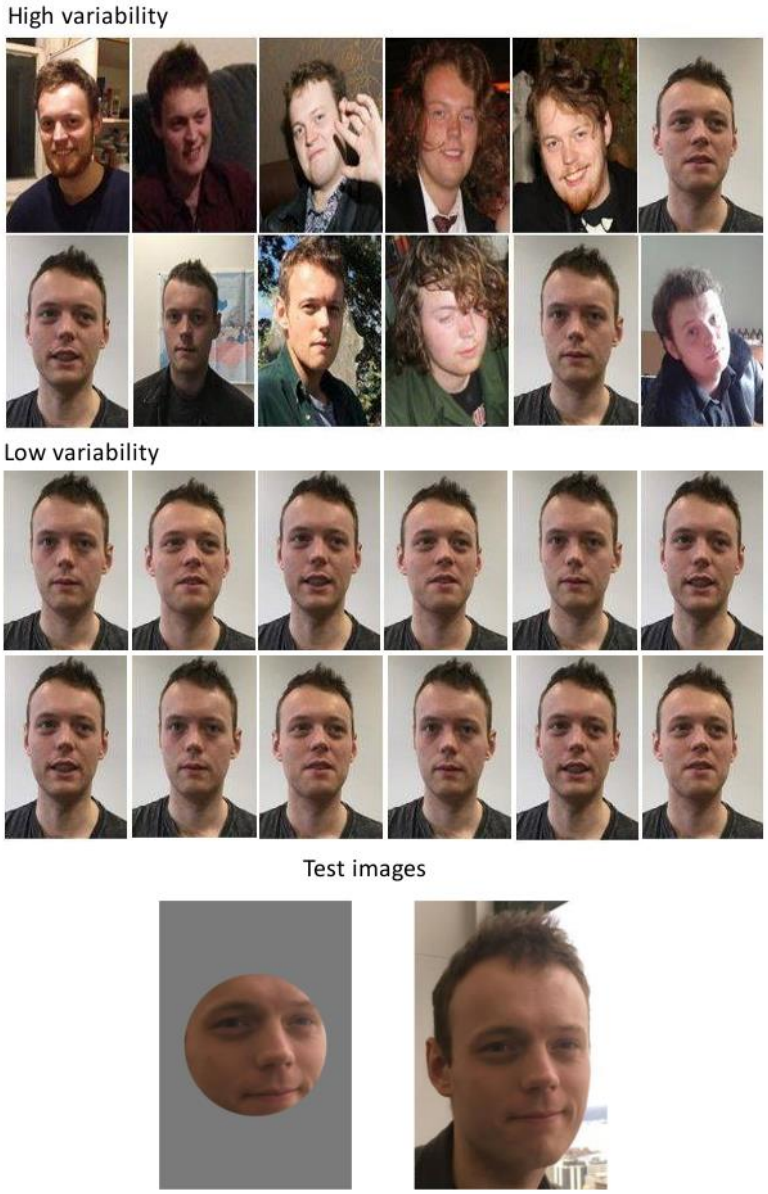

Figure 1. Illustration of stimuli used in the learning phase and tests. The low variability learning condition (middle) included three different pictures that were repeated four times each per individual. The high variability condition (top) included the same three pictures plus nine different ambient images. For each identity, 12 individual images were first presented one by one for three seconds, after which a review slide including the 12 images was presented for 12 seconds. Bottom panel shows examples of stimuli used at test. Images were cropped to only show internal features (left; Experiments 1 and 2) or showed full headshots (right; Experiment 2). For copyrights reasons, the picture set used in the experiment cannot be shown. The person depicted here has provided permission.

Learning phase. All participants studied 4 identities in the LV condition and 4 in the HV condition, in 8 blocks presented in a random order. The variability condition in which each identity was learned was counterbalanced across participants. Faces were learned via 12 unique images in the HV condition, and via 3 unique images in the LV condition, repeated 4 times each to equalise the total exposure duration. Within each identity block, images were randomized, presented for 2 seconds each, and separated by 500 ms blanks. Each block ended with a 12-second review slide showing the same set of images that had just been presented one by one.

Recognition and matching tasks. The recognition task consisted of 48 trials, 24 learned identities (3 trials for each of the 4 identities across the 2 learning conditions) and 24 new identities. Participants judged if they had seen the person during the learning phase or not.

The simultaneous matching task consisted of 48 trials, 32 with learned identities (2 match and 2 mismatch trials for each of the 4 learned identities across the 2 learning conditions), plus 16 (8 match and 8 mismatch trials) for unlearned identities different than those used in the recognition task. Participants judged if 2 faces presented side by side showed the same person or not.

Both tasks required participants to respond as quickly and as accurately as possible by means of keys 1 and 2 of their keyboards (seen/not seen in the recognition task, and same/different in the matching task). Images (always new exemplars with external features masked) were shown for maximum 3 seconds after which participants were prompted to respond, and trials were separated by 500 ms blanks. Two examples were presented before each task began. The recognition task was conducted immediately after the learning phase to

avoid additional learning that could have occurred if the matching task was administrated beforehand.

Familiarity Check. A post-experimental name recognition survey was used to check that participants were unfamiliar with the 8 celebrities. Participants were presented with the name of each identity and asked to indicate whether they were familiar with each person.

Design and analyses. In both tasks, learning condition (HV, LV, Unlearned) was manipulated within-subject. In the matching task, following previous research (White et al., 2014), we analysed the data for match and mismatch trials separately because they imply different abilities (i.e., tell people together or tell people apart). We analysed mean accuracy in the recognition task and in each trial type of the matching task with one-way Analyses of Variance (ANOVAs – run in Jasp), with learning condition as repeated measure factor. We followed up main effects with Student t-tests. We excluded participants who reported familiarity with any of the celebrities (N = 8), and participants who showed signs of no compliance with instructions (i.e., performance at chance and/or response times under 600 ms that suggest button pressing; N detailed below). The same analyses performed on median correct reaction times (RTs) are presented in Supplementary materials; overall, they show similar patterns and no speed-accuracy trade off.

Results and discussion

Recognition task. We discarded data of 4 participants who did not follow instructions, leaving 44 participants (22 men; Mean age = 37.86 years ± 13.27). There was a significant main effect of learning condition on accuracy, $F(2,86) = 126$, $p < .001$, $\eta^2 = .746$. Follow-up t-tests

showed that participants were more accurate in reporting that they had not seen new faces (Mean = .90 ± .113, that is, a false alarm rate of 10%) than in recognising previously studied faces in both the HV, $t(43) = 6.859$, $p < .001$, $d = 1.034$, and LV conditions, $t(43) = 16.466$, $p < .001$, $d = 2.482$. High accuracy in the unlearned condition can be explained by a conservative response bias, whereby participants tended to respond that a face was new. Importantly, HV faces (Mean = .64 ± .21) were better recognised than LV faces (Mean = .341 ± .17), $t(43) = 8.862$, $p < .001$, $d = 1.336$, showing that exposure to within-person variation enables accurate recognition from internal features, and may contribute to the development of robust representations based on invariant features. Results are presented on **Figure 2**.

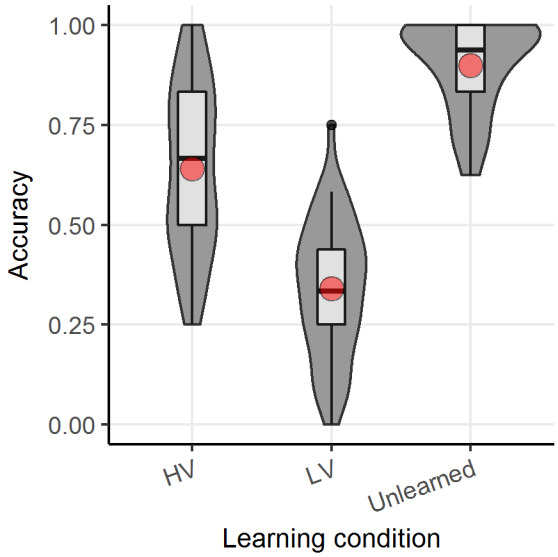

Figure 2. Accuracy in the recognition task as a function of learning condition. Red circles represent mean values, boxplots show distributions in quartiles, and violins' width is proportional to performance distribution across participants.

Matching task. We discarded data of 3 participants who failed to follow instructions, leaving 45 participants (23 men; Mean age = 37.64 years ± 13.2). For match trials, see **Figure 3**

(left panel), there was only a marginal effect of learning condition, $F(2,88) = 2.49$, $p = .089$, $\eta^2 = .054$. We explored differences between pairs of conditions with paired sample t-tests. Consistent with Ritchie and Burton's (2015) findings, matching accuracy was higher for HV faces (Mean = .76 ± .186) than LV faces (Mean = .633 ± .191), $t(44) = 2.666$, $p = .011$, $d = 0.397$. Accuracy for unlearned faces (Mean = .731 ± .177) was not significantly different from either HV faces or LV faces, $t(44) = 0.724$, $p = .473$, $d = 0.108$, and $t(44) = -1.302$, $p = .2$, $d = -0.194$, respectively. For mismatch trials, learning condition had no significant effect on accuracy, $F(2,88) = 1.367$, $p = .26$, $\eta^2 = .03$, see **Figure 3** (right panel).

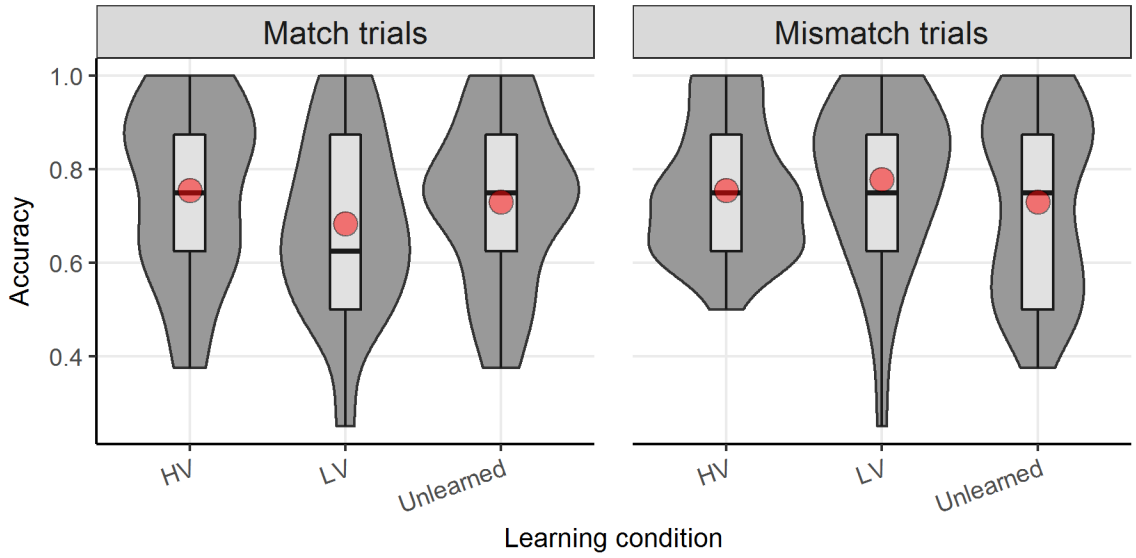

Figure 3. Accuracy in the matching task as a function of learning condition, in each trial type (match, left; mismatch, right). Red circles represent mean values, boxplots show distributions in quartiles, and violins' width is proportional to performance distribution across participants.

Data from our recognition task suggest that exposure to variability during learning helped the encoding of internal features. Our matching task results, however, do not replicate the significant main effect of learning condition on face matching accuracy for match trials (see

Ritchie & Burton, 2017) and only find a marginal effect. Nevertheless, a direct comparison of accuracy in HV and LV conditions shows the expected advantage of variability. It could be that the benefits of variability are less obvious when we process a face based only on internal features, and are stronger with headshots, perhaps because they incorporates external features useful to perform the task. The smaller effect size obtained here might also be due to other changes in the learning phase compared to the original paradigm. Experiment 2 addresses these questions.

Experiment 2

We replicate and expand Experiment 1 by testing more directly whether variability promotes a focus on stable internal features. To that end, we compared the extent to which variability benefits performance with headshots (ambient images showing internal and external features), as in Ritchie and Burton (2017), and with cropped photos showing internal features only, like in Experiment 1 and in Murphy et al. (2015). Similar benefits of variability regardless of testing images would be consistent with the notion that variability results in learning based on internal features. Larger benefits of variability with headshots than with cropped images would suggest that exposure to within-person variability also promotes learning based on external features.

Method

We preregistered our analyses plan on Open Science Framework before data collection [https://osf.io/pqv9f/]. We used the same stimuli and procedure as in Experiment 1, with an additional testing condition with headshot images. To simplify the exclusion procedure and

detect button pressing, we included two attention checks in each task. They consisted of an image of the same size as the other stimuli, showing an flesh tone coloured circle with text instructing participants to press a different key (i.e., 5, 6, 7, or 8) than the usual response keys (i.e., 1 and 2).

Participants. A power analysis based on the recognition data in Experiment 1 suggests that 7 participants are required to replicate the main effect of learning condition (η^2 = .746) with a power of .95. By contrast, for accuracy in the matching task, Experiment 1 shows a smaller effect size (η^2 = .054) than in the original study by Burton and Ritchie and only yields a marginal effect. Since the effect size of the difference between headshots and pictures showing internal features is unknown, we aimed to have at least 70 participants per testing condition, namely double what the initial power calculation yielded in Experiment 1.

We first recruited 180 first year students at Victoria University of Wellington (45 per identity counterbalance/testing condition combination) who were participating for course credits. Four participants did not complete the experiment. We excluded an additional 115 who reported familiarity with any celebrity (this large number can be explained by the geographical proximity between Australia and New Zealand), and 4 who did not follow instructions (i.e., failed more than one attention check in either task), leaving 57 participants (16 men, 41 women; mean age = 19.36 years ± 5.24). We then ran 107 extra MTurk workers located in the US. We excluded 19 who were familiar with at least one celebrity, and 2 others who did not follow instructions, leaving 86 participants (45 men, 41 women; mean age = 38.7 years ± 11.87). The final combined sample comprised 143 participants, 73 in the headshot condition (36

women, 37 men; Mean age = 29.93 years ± 12.5), and 70 in the internal features condition (46 women, 24 men; Mean age = 32.31 years ± 14.52).

Design and analyses. The design was identical to that in Experiment 1 except for the addition of testing condition (headshot vs. internal features) as a between-subject factor. We conducted the same two-way mixed effect ANOVAs with testing condition as a between-subject factor and learning condition as a repeated measure factor on mean accuracy in the recognition task, and in each trial type of the matching task. Again, analyses of median correct RTs are reported in Supplementary Materials for exhaustiveness.

In addition, in each part (i.e., recognition, match trials, mismatch trials), we computed a Variability Effect Index (VEI), namely a score indexing the benefits of variability in each testing condition, to follow up interactions between learning condition and testing condition, or a main effect of learning condition in the absence of interaction. The VEI was calculated as follows in order to compare benefits of variability in each group in cases where performance differed between the learning conditions: (accuracy HV - accuracy LV)/(accuracy HV + accuracy LV)[3]. Positive scores reflect a benefit of HV on learning, and negative scores, a disadvantage of HV on learning. We then compared VEI in the two groups by means of Student t-tests.

Results and discussion

Recognition task. Like in Experiment 1, there was a significant effect of learning condition, $F(1.769, 249.463) = 299.685$, $p < .001$, $\eta_p^2 = .680$. Overall, participants were more accurate in

[3]In our pre-registration, we planned to calculate VEI with accuracy in the LV condition as baseline, i.e., (accuracy HV – accuracy LV)/accuracy LV. However, we later realised that using performance in both learning conditions as a baseline would more adequately account for possible differences in either learning condition between the two testing groups.

rejecting new faces (Mean = .911 ± .127, that is 8.9 % of false alarms) than in recognising HV, $t(142) = 8.965$, $p < .001$, $d = 0.75$, and LV faces, $t(142) = 20.577$, $p < .001$, $d = 1.721$. Importantly, HV faces (Mean = .749 ± .195) were recognised more often than LV faces (Mean = .435 ± .229), $t(142) = 17.059$, $p < .001$, $d = 1.427$. There was a significant main effect of testing condition $F(1,141) = 22.07$, $p < .001$, $\eta_p^2 = .135$, and recognition was more accurate from headshots (Mean = .744 ± .25) than from internal features only (Mean = .65 ± .29). The interaction between learning condition and testing condition was significant, $F(1.769, 249.463) = 4.208$, $p = .020$, $\eta_p^2 = .029$. The follow-up comparison of VEI in each testing condition showed that the relative benefits of variability were similar for headshots (Mean = .302 ± .249) and internal features (Mean = .310 ± .261), $t(141) = -.206$, $p = .837$, $d = -0.034$, indicating that within-person variation allows the encoding of internal features present in both image types. In addition, we explored whether the impact of testing condition differed for faces in each learning condition with independent samples t-tests. Headshots were better recognised than internal features for HV faces, $t(141) = 4.938$, $p < .001$, $d = 0.826$, and also for LV faces, $t(141) = 2.617$, $p = .01$, $d = 0.438$. Effectively, the interaction was driven by participants' similar ability to correctly reject new faces from headshots and internal features, $t(141) = 1.639$, $p = .103$, $d = 0.274$. Results are presented on **Figure 4**.

We replicate the large advantage for HV faces relative to LV faces found in Experiment 1. Further, we show that performance is improved by the presence of external features in headshots compared to images showing internal features only. Importantly however, the relative benefits of variability are comparable with the two picture formats, suggesting that

variability promotes the encoding of invariant facial features that are present in both picture types.

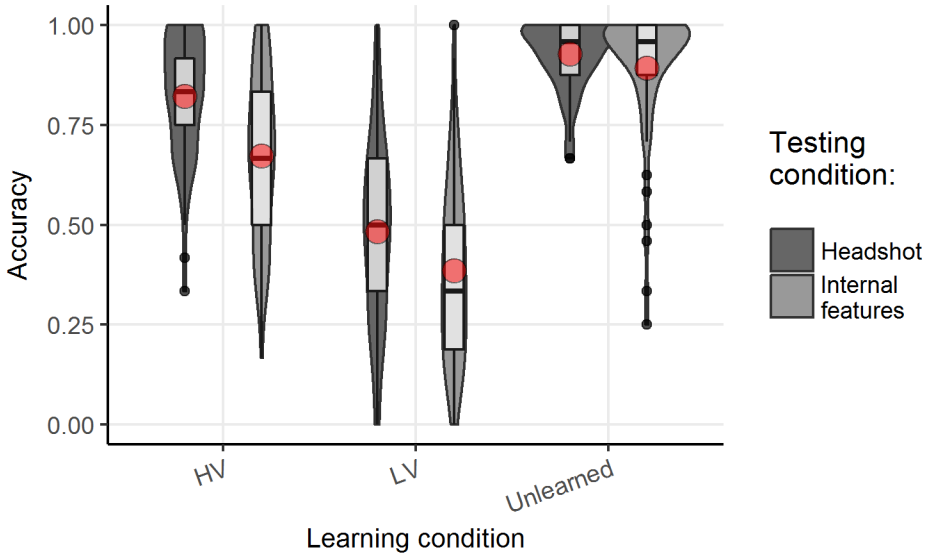

Figure 4. Accuracy in the recognition task as a function of learning condition and testing condition. Red circles represent mean values, boxplots show distributions in quartiles, and violins' width is proportional to performance distribution across participants.

Matching task. Again, we analysed match and mismatch trials separately. Analysis of accuracy on match trials, see **Figure 5** (top left panel), showed a main effect of learning condition, $F(1.927, 271.653) = 17.182$, $p < .001$, $\eta_p^2 = .109$. Follow-up t-tests showed that HV faces (Mean = .802 ± .16) were matched more accurately than both LV faces (Mean = .713 ± .204) and unlearned faces (Mean = .72 ± .166), $t(142) = 5.113$, $p < .001$, $d = 0.428$, and $t(142) = 5.424$, $p < .001$, $d = 0.454$, respectively. LV and unlearned faces did not differ from each other, $t(142) = -0.391$, $p = .696$, $d = -0.033$. There was no significant effect of testing condition, $F(1, 141) = 2.578$, $p = .111$, $\eta_p^2 = .018$, and no interaction, $F(1.927, 271.653) = .154$, $p = .849$, $\eta_p^2 = .001$. The comparison of VEI in each testing condition confirmed that the effect of variability

was comparable with headshots (Mean = .058 ± .163) and with internal features (Mean = .08 ± .16), $t(141) = -.849$, $p = .397$, $d = -0.142$.

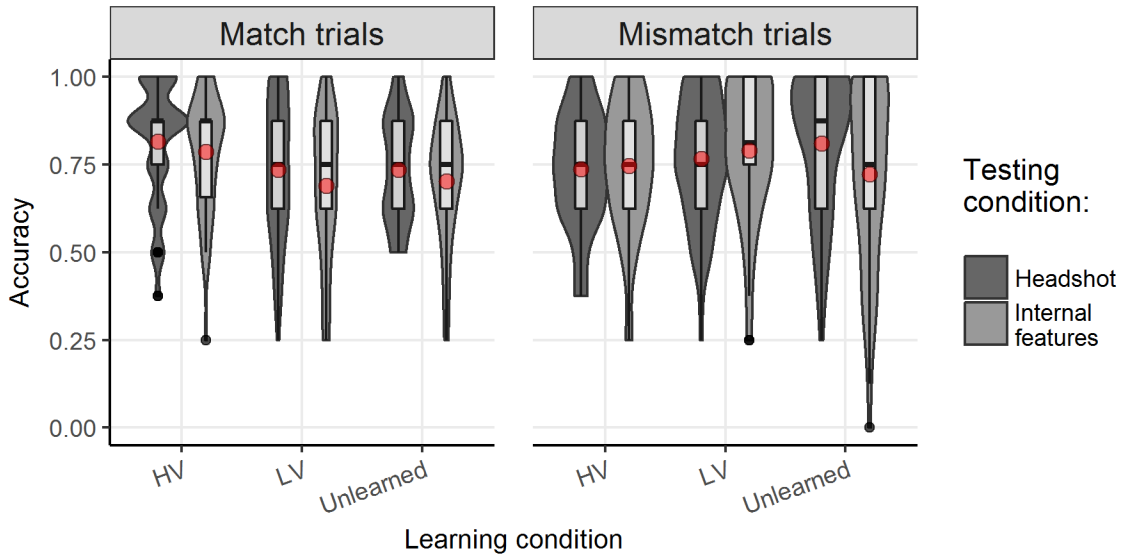

Figure 5. Accuracy in the matching task as a function of learning condition and testing condition, in each trial type (match, left; mismatch, right). Red circles represent mean values, boxplots show distributions in quartiles, and violins' width is proportional to performance distribution across participants.

Here we show an improvement in performance on match trials for identities learned from a highly variable compared to a less variable set of images. The results suggest that exposure to variability at learning helps us to subsequently determine whether two new images of a person show the same person. Mere prior exposure to a face does not seem sufficient to improve this ability, since faces learned in the low variability condition were not matched more accurately than unlearned faces.

Regarding accuracy on mismatch trials, see **Figure 5** (top right panel), there was no main effect of learning condition, $F(1.909,269.206) = 2.094$, $p = .127$, $\eta_p^2 = .015$, nor of testing condition, $F(1,141) = .518$, $p = .473$, $\eta_p^2 = .004$. The interaction was significant $F(1.909,269.206) = 5.345$, $p = .006$, $\eta_p^2 = .037$, but a follow-up t-test on VEI showed a comparable effect of variability for headshots (Mean = -.018 ± .144) and internal features (Mean = -.029 ± .15), $t(141) = 0.409$, $p = .683$, $d = 0.068$. Further, we explored the simple effect of testing condition in each learning condition. The type of picture used at test did not significantly affect performance with faces learned in the HV condition, $t(141)$, -.283, $p = .777$, $d = -0.047$, or the LV condition, $t(141) = -0.785$, $p = .434$, $d = -0.131$. However, mismatch performance for unlearned faces was more accurate when viewing headshots than when viewing internal features alone, $t(141) = 2.291$, $p = .023$, $d = 0.383$.

In sum, we replicate and expand findings of Experiment 1 and show that variability helps learning new faces, which translates in increased recognition rate, and an improved ability to tell different exemplars of the same person together (i.e., match trials). By contrast, again, mismatch trials showed no apparent benefit of variability. We also show that the presence of external features improves overall performance in the recognition task but not in the matching task. However, in both tasks, the relative benefits of variability are comparable with headshots and images showing internal features, suggesting that although external features contribute to the recognition of newly learned faces, within-person variation helps encode invariant features.

General discussion

In this study, we investigated how exposure to within-person variability facilitates learning new faces and what kinds of face processing skills it improves. First, the results of two experiments show that high within-person variability improves both face perceptual discrimination and recognition abilities. Individual faces vary idiosyncratically (Burton, Kramer, Ritchie, & Jenkins, 2015) and so exposure to high ranges of variability in images of a given person might grant insights into the dimensions on which their face varies or remains stable. Second, our data suggest that within-person variability helps build robust representations that are generalizable by promoting a focus on invariant internal features. We discuss our results for the two cognitive domains tested in this study, and then conclude on more general benefits of variability.

Recognition memory task. Our recognition results across the two experiments show a very clear memory advantage for faces learned under high variability conditions compared to those learned in low variability conditions. These results are in line with Murphy et al. (2015)'s findings, despite marked methodological differences between their study and ours: they used an unsupervised learning procedure in which participants had to determine individual identities for themselves in a face array without feedback, while our participants were explicitly exposed to distinct blocks of images for each to-be-learned identity. Such converging evidence demonstrates the robustness of the benefits of within-person variability in developing reliable memory representations for newly learned faces. The fact that the benefits of variability are similar when tested with and without external facial features indicates that exposure to within-

person variations promotes the encoding of internal features, which are stable across encounters.

Although the size of the benefit of variability was comparable for headshots and for cropped pictures, we also find that overall recognition rates are higher with headshots than with cropped images. This suggests that, when available, external facial features shown in headshots (e.g., hair, head shape) are also used to compare a new exemplar of a face to a memory representation of that face. This result is consistent with recent findings demonstrating the important contribution of external features in face recognition, even for highly familiar faces (Devue, Wride, & Grimshaw, in press). Improved performance with headshots compared to internal features alone might also be partly due to a change in picture format between learning and test phases. A previous study showed that switching from headshots at learning to internal features at test reduced discrimination sensitivity compared to consistent picture formats (i.e., headshots or internal features in both phases; Toseeb, Keeble, & Bryant, 2012). Future studies should look at this issue systematically and explore the potential interaction between variability benefits and change in image format between learning and test.

Matching task. Exposure to variability does not significantly affect people's ability to tell different people apart, in line with previous observations (Ritchie & Burton, 2017). By contrast, prior exposure to high levels of variation improves the ability to perceive that two different images show the same person. It is not unprecedented to find an effect of a manipulation on one trial type and not the other (see White et al., 2014). Match and mismatch trials may not measure the same aspect of face processing, and in some circumstances performance on one is

uncorrelated with performance on the other (Megreya & Burton, 2007). Match trials assess the ability to incorporate within-person variability across two images and correctly tell that two images show the same person, whereas mismatch trials assess the ability to determine the limits within which an individual face may vary and correctly tell people apart. Although the benefit of variability appears more modest here than in the recognition task, it replicates and expands previous findings in different learning paradigms (e.g., Andrews, Jenkins, Cursiter, & Burton, 2015; Ritchie & Burton, 2017). Also consistent with Ritchie and Burton's findings, we see that learning a face with limited variation (i.e., in the low variability condition) does not improve matching abilities compared to unlearned faces. So prior exposure to a face is not sufficient in itself to improve matching abilities, and exposure to significant levels of variation is necessary to improve performance.

In contrast to the recognition task, the format of picture used at test does not affect the ability to tell two persons together or apart overall. On match trials, where we found a benefit of variability, we see that it is comparable for headshots and for cropped images showing internal features alone. Interestingly however, for unlearned faces, performance was improved by the availability of extra-facial information in headshots compared to cropped images. This pattern of results suggests that after they have efficiently learned a face, people become better able to successfully base their perceptual discrimination on facial features that are more reliable because they are more consistent, namely the internal features. This idea is in line with findings showing that matching unfamiliar faces from internal features generate more errors than for familiar faces (Clutterbuck & Johnston, 2002), and we know that people rely mostly on external features to process unfamiliar faces (Ellis, Shepherd, & Davies, 1979; Young, Hay,

McWeeny, Flude, & Ellis, 1985). In another recent study (Kemp, Caon, Howard, & Brooks, 2016), people were also better in a matching task when external features of unfamiliar faces were available, but only on easy trials (i.e., presenting different people with dissimilar appearance on mismatch trials, or the same person with similar hairstyle on match trials). However, on difficult trials (i.e., presenting different people with similar appearance on mismatch trials, or the same person with different hairstyle on match trials), the removal of external features actually improved matching performance, probably because external features were misleading and that their unavailability forced people to focus on more informative features. Our exploratory finding on unlearned faces thus deserves more investigation and it is unclear how trial difficulty and picture format would interact with learning conditions. Nevertheless, the fact that picture format does not affect performance for learned faces might thus reflect the shift from an overreliance on external features for new faces, to a more effective reliance on internal features as people get acquainted with the face.

Benefits of variability in the development of familiarity. The relative benefits of variability seem larger in a task that involves memory than in a task that requires perceptual discrimination. This is confirmed by larger Cohen's *d's* for the comparison of accuracy in HV and LV conditions in the recognition task (i.e., 1.336 in Experiment 1, and 1.427 in Experiment 2) than on match trials (i.e., 0.397 in Experiment 1, and 0.428 in Experiment 2). This might not be so surprising as matching faces relies on perceptual discrimination abilities and is in principle achievable without previous exposure to a person's face. By contrast, recognising a face necessarily requires to have successfully encoded information relative to the person's appearance first, making the benefit of variability more apparent.

Some authors have suggested that matching efficiency indexes the degree of familiarity with a face (Clutterbuck & Johnston, 2004; Clutterbuck & Johnston, 2002). In that sense, our participants did not show any sign of having developed familiarity for faces learned under low variability conditions yet, since matching performance was comparable to that for unlearned faces. However, this interpretation contrasts with increased reports of familiarity for faces learned under the low variability condition in the recognition task (i.e., 34.1% hits in Experiment 1 and 43.5% overall in Experiment 2) compared to unlearned faces (i.e., 10 % false alarms in Experiment 1 and 8.9 % in Experiment 2). This demonstrates that faces in the low variability condition were familiar beyond the level of familiarity for novel faces expressed by false alarms, and must have been encoded to some degree. Therefore, while prior exposure to a face with low levels of variations does little to improve perceptual discrimination abilities, it still allows to start developing a representation of the face. These representations are not resistant to changes in viewing conditions, and increased levels of variations are necessary to develop more robust representations.

Variations in the low variability condition included changes in viewpoint and facial expression, while variations in the high variability condition also included changes in environment (i.e., lighting, background) and physical appearance (e.g., make-up, hairstyle, facial hair, age). Future research should seek to isolate these different factors, while conserving the ecological validity afforded by ambient images, in order to identify their relative contribution to the development of robust representations.

Conclusion. Altogether, our data suggest that being briefly exposed to a face without a substantial range of variation is not sufficient to develop robust representations. Participants

showed signs that they had encoded faces they had been exposed to via 3 images showing low within-person variability. However, presenting different ambient images of the same person for the same brief duration (i.e., 48 seconds in total) produces large improvements in memory performance, and moderate improvements in perceptual discrimination abilities. We show that recognition is more accurate for headshots than images that show only internal features, but that exposure to high levels of variation while learning new identities produces comparable effects for the two types of images. This suggests that exposure to within-person variations encourages observers to focus on stable facial dimensions contained in inner features. Our results add to the growing body of evidence that exposure to variability is important for developing robust representations, and demonstrates that variability helps both recognition memory and perceptual discrimination.

Authors' notes

Contributions: ER contributed to the study design and manuscript writing, prepared material, collected and analysed data. TS contributed to the study design, data analyses, and manuscript writing. KR provided original picture stimuli, contributed to the study design and helped draft the manuscript. CD developed the study concept and design, collected and analysed data, and drafted the manuscript. A preliminary version of this manuscript was posted on the Open Science Framework website in July 2018 [https://osf.io/5b8c2/]. Aggregated de-identified data for both experiments are visible at [https://osf.io/tcvz7/]. Pre-registration for Experiment 2 is visible at [https://osf.io/pqv9f/]. Portions of the data were presented at the Experimental Psychology Conference (2018).

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology*, *68*(10), 2041–2050.

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.

Benton, C. P., Redfern, A. S. (2017). Expression Dependence in the Perception of Facial Identity. *i-Perception*, 1-15.

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.

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*(3), 207-218.

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

Burton, M. A., Kramer, R. S. S., Ritchie, K. L. & Jenkins, R (2015). Identity from variation: representations of faces derived from multiple instances. *Cognitive Science,* (40), 202–223.

Burton, A.M., White, D., McNeill, A., (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*(1), 286-291.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*(3), 243-248.

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.

Devue, C., Wride, A., & Grimshaw, G. M. (in press). New insights on real-world human face recognition. *Journal of Experimental Psychology: General.*

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.

Ellis, H.D., Shepherd, J.W. & Davies, G.M. (1979). Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition. *Perception, 8*, 431 –439.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 366*, 1671–1683.

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.

Kemp, R. I., Caon, A., Howard, M., & Brooks, K. R. (2016). Improving Unfamiliar Face Matching by Masking the External Facial Features. *Applied Cognitive Psychology*, *30*(4), 622–627.

Kramer, R. S. S., Manesi, Z., Towler, A., Reynolds, M. G., & Burton, A. M. (2018). Familiarity and Within-Person Facial Variability: The Importance of the Internal and External Features. *Perception*, *47*(1), 3–15.

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising Superman: How Glasses Affect Unfamiliar Face Matching. *Applied Cognitive Psychology, 30,* 841-845.

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.

Liu, C. H., Bhuiyan, A., Ward, J., & Sui, J. (2009). Transfer between pose and illumination training in face recognition. *Journal of Experimental Psychology: Human Perception and Performance, 35(4),* 939-947.

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34(1),* 77-100.

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. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69(7),* 1175-1184

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.

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.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*(5), 1–9.

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.

Toseeb, U., Keeble, D. R. T., & Bryant, E. J. (2012). The significance of hair for face recognition. *PLoS ONE*, *7*(3), 1–8.

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.

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M, & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14*(6), 737-746.