

The Effect of Variability in Appearance and Context on Face Learning

Kevin Nguy, Tom Lapadula and Christel Devue

Psychology and Neuroscience of Cognition Research Unit, University of Liège,

Liège, Belgium

Author Note

Kevin Nguy  <https://orcid.org/0000-0003-4745-1231>

Tom Lapadula  <https://orcid.org/0009-0005-7723-2787>

Christel Devue  <https://orcid.org/0000-0001-7349-226X>

We have no known conflict of interest to disclose.

Correspondence should be addressed to Christel Devue, Université de Liège, Place des Orateurs 2, 4000 Liège, Belgique. Email: cdevue@uliege.be

Acknowledgments

The authors thank Vincent Didone for his helpful input on the power and statistical analyses.

This work was supported by ULiège with a grant from the Wallonia-Brussels Federation —

Concerted Research Actions (ARC 23/27-05 – COMPRESS).

Abstract

Exposure to multiple unique images of the same face enhances recognition. Using ambient pictures and videos, face-learning studies have shown improved generalization, suggesting more refined memory representations for faces exhibiting greater variability. Although ambient images provide more ecologically valid learning conditions, the relative contribution of different sources of variability remain largely unexplored. In two pre-registered experiments, we investigated the effect of variability in appearance (stable, variable hairstyle) and context (unique, multiple background scenes) during face learning, as these factors may influence the coarse-to-fine encoding of diagnostic facial features. In Experiment 1, eight identities were learned in a blocked manner, whereas they were randomised in Experiment 2. Recognition was assessed using images displaying only inner facial features. We hypothesized that conditions promoting finer encoding of inner facial features (variable appearance and multiple contexts in Experiment 1) or providing stable learning anchors (stable appearance and unique context in Experiment 2) would improve recognition. In Experiment 1, variable appearances enhanced recognition performance relative to stable appearances, despite diverting gaze from inner features during learning. However, Experiment 2, along with supplementary cross-experiment analyses, revealed that randomizing identities eliminated the recognition disadvantage observed for stable faces, possibly by encouraging participants to re-engage attention with each learning image. Contextual variability did not affect recognition performance in either experiment, suggesting that appearance variability and learning procedure are stronger determinants of efficient learning. These findings highlight the critical role of appearance variability and of learning procedure in shaping recognition outcomes.

Keywords: Face recognition, Within-person variability, Context variability, Learning procedure

The Effect of Variability in Appearance and Context on Face Learning

In daily life, successful social interaction requires recognizing other individuals regardless of context and despite unpredictable changes that may occur between encounters. Indeed, besides stable and variable facial characteristics idiosyncratic to an individual (Burton et al., 2016), such as constrained by bone and muscular structure, faces may vary on multiple dimensions due to changes in appearance (hairstyle, makeup, accessories), age, weight or health. The aspect of a face may also differ more or less drastically due to external conditions such as lighting or the observer's viewpoint. These changes have been shown to impair performance in recognition tasks (Bruce, 1982; Davies & Milne, 1982; O'Toole et al., 1998; Righi et al., 2012), even among professionals trained daily in face recognition, such as passport or police officers (Burton et al., 1999; White, Kemp, et al., 2014). However, this effect applies mainly to unfamiliar faces, which are suggested to be processed primarily at an image-based level (Hancock et al., 2000; Megreya & Burton, 2006). In contrast, recognition of familiar faces is relatively resistant to such variations, as familiarity supports the formation of a stable internal representation of the face (Hancock et al., 2000; Jenkins et al., 2011; Jenkins & Burton, 2011).

Although unexpected variations hinder facial recognition, a substantial body of research has demonstrated that exposure to these variations during learning contributes to the development of robust facial representations (Andrews et al., 2015; Baker et al., 2017; Bruce, 1994; Corpuz & Oriet, 2022; Dowsett et al., 2016; Kramer, Young, et al., 2018; Matthews et al., 2024; Menon et al., 2015; Reedy & Devue, 2019; Ritchie & Burton, 2017). Importantly, this effect appears to be driven specifically by variability rather than mere exposure to faces (Murphy et al., 2015). Indeed, Murphy et al. (2015) found that recognition of internal facial features was enhanced when learning occurred through a large set of diverse ambient images,

compared to repeated exposure to a limited number of identical images, even when the total exposure time was the same.

From these observations, some authors have suggested that learning a face not only involves learning its stable diagnostic features but also its idiosyncratic variability (Burton et al., 2016; Mileva et al., 2020). Two main theoretical accounts have been proposed to characterize the manner in which this idiosyncratic variability is accounted for in memory. First, the different instances of a face could be stored as separate representations (Menon et al., 2015). In this case, recognition would rely on the closest match between a face currently presented and one of the stored instances. Alternatively, the averaging hypothesis (Burton et al., 2005, 2016; Hancock et al., 2000; Jenkins et al., 2011; Koca & Oriet, 2023; Kramer, Manesi, et al., 2018) suggests that these instances are integrated into a unified, abstract representation through statistical learning (de Fockert & Wolfenstein, 2009). Across encounters, diagnostic invariant features are extracted and progressively consolidated in memory, whereas more variable ones are averaged out, describing how exposure to variations may facilitate face learning and yield robust representations. Importantly, these accounts are not mutually exclusive. Both average-based and instance-based representations may coexist within the same psychological space (Kramer et al., 2015; Menon et al., 2015; Neumann et al., 2013). Although cognitively demanding, such a system may maximize recognition chances by allowing faces to be matched either to the average representation or to the specific stored instances.

While the benefits of exposure to variability is well established, the respective contribution of each source of variability remains to be clarified. Indeed, most previous studies have relied on ambient images or videos, which favour ecological validity but limit the ability to isolate the relative effect of the different sources of variability (Baker et al., 2017; Corpuz & Oriet, 2022; Devue et al., 2019; Menon et al., 2015; Reedy & Devue, 2019;

Ritchie & Burton, 2017; White, Burton, et al., 2014; Zhou et al., 2022). Amongst the multiple parameters that may vary in ambient images, a specific positive impact of exposure to variations in lighting and viewpoint has already been suggested (Hill et al., 1997; Liu et al., 2009; Longmore et al., 2008; Yamanashi Leib et al., 2014). Both factors would facilitate the formation of three-dimensional facial representations, more robust than initial image-based representations, reducing the performance gap between familiar and unfamiliar faces. Exposure to other types of variation, such as facial expressions, has more ambiguous effects (e.g. Liu et al., 2015, 2016; Xu & Armony, 2024). In the current study, we examined the specific role of variations in physical appearance and context on face learning. Indeed, unexpected changes in appearance (e.g., in hairstyle, hair colour, makeup or paraphernalia) are known to impair face recognition (e.g. Patterson & Baddeley, 1977; Toseeb, Keeble & Bryant, 2012), across the whole familiarity range (Devue, Wride & Grimshaw, 2019). Likewise, encountering faces out of context leads to recognition errors (Davies & Milne, 1982; Gruppuso, 2007; Smith & Handy, 2014). Therefore, it is plausible that exposure to such variations during learning would specifically improve recognition performance.

Recently, the cost-efficient encoding hypothesis proposed a theoretical account that focuses on the role of variations in physical appearance (or the lack thereof) in shaping the formation of a face's internal representation. The hypothesis proposes a coarse-to-fine processing strategy, in which a face's relative variability determines the way it is encoded. By default, processing would assume stability and favour large-scale diagnostic information (e.g., hair colour, head shape, or gross facial configuration) due to its low memory cost, with a gradual encoding of finer details when variability in appearance is observed or discrimination needs increase (Devue & de Sena, 2023). These finer representations become more resilient to changes but incur a higher cognitive cost. For instance, when a person is always seen with a consistent hairstyle, making their overall appearance very stable, the

hairstyle becomes a reliable feature for recognition, providing it is sufficiently discriminant, and the encoding of more detailed features ends up less valuable. In contrast, noticing and anticipating variations in hairstyles across encounters prompts the refinement of inner features, leading to better adaptation to subsequent changes. In other words, the distinct representation of both external and internal facial details would be weighted according to their relative variability, with diagnostic features defined as those that demonstrate the strongest stability over time (see also Reedy & Devue, 2019).

This cost-efficient rationale can be extended beyond facial information to include extra-facial cues such as context and clothing. For instance, when an individual is consistently encountered in a unique setting and attire—such as a waiter at one’s regular restaurant—these cues may become diagnostic of the person’s identity, reducing reliance on fine-grained facial representations. The flipside, however, is that recognition may fail when the person is encountered outside of this usual context. This account aligns with empirical findings showing poorer recognition in the absence of contextual variability (Smith & Handy, 2014; Young et al., 1985). A classic illustration is the “butcher-on-the-bus” effect described by Mandler (1980), in which a familiar individual is not recognized when seen in an unexpected context. Taken together, these observations suggest that, under conditions of limited contextual variability, recognition may depend on relatively coarse encoding mechanisms.

The goal of the current study was to assess the specific contribution of variations in facial appearance and in context on face learning. More specifically, we sought to replicate the beneficial effects of variability reported in previous studies while disentangling the underlying sources of this effect. To manipulate variability of facial appearance, identities were shown with either one single hairstyle (“stable”) or multiple different hairstyles (“variable”) across images during the learning phase. Likewise, to manipulate variability of

context, identities were presented either within a single (“unique”) background or across different (“multiple”) background scenes. To preserve the ecological quality of ambient images and avoid artificially impoverishing the learning material, identical changes in lighting and viewpoint were applied to faces in all conditions. To assess the quality of the resulting facial representations, recognition was tested using images restricted to internal facial features, with the idea that if they have been encoded coarsely, they would be less likely to be recognised than if they have been encoded more finely. In Experiment 1, based on the cost-efficient encoding hypothesis (Devue & de Sena, 2023), we expected both variability in appearance and in context to reduce reliance on pictorial features, thereby promoting more robust identity representations based on fine-grained facial information. Accordingly, we predicted better recognition performance for faces learned under variable compared to stable appearance conditions, as well as for faces associated with multiple compared to unique identities.

Experiment 1

Method

Participants and sample size calculation

Sample size was computed using G*Power for four paired t-test (tails: two; effect size: 0.25) with an alpha probability of 0.0125 (Bonferroni-Holm correction) aiming at 0.80 power. This gave us an estimation of 182 participants. This number was multiplied by an inflation factor of 1.1703 which resulted in a total of 213 participants in relation with a Pocock correction applied for sequential analyses.

We planned to perform sequential analyses with 2 interim steps: one at 84/180¹ and the other planned at 2/3 of data collection. Pocock alpha-spending function was used to

¹The initial planned sample size for the first step of the sequential analysis was a third of a total target of 252 participants, that is 84 participants. However, we realised after collecting data for this first step that an error in the order of operations used to compute the required sample size had led to an overestimation of the final sample size. Namely, we had applied a Bonferroni correction to the alpha

determine corrected alpha probabilities using the rpact library in R (see Lakens et al. (2021)). Corrected alphas are 0.0294 for the first interim analysis, 0.0189 for the second and 0.0219 for the final analysis. Data collection was stopped for futility at 84 participants following an equivalence test that resulted in the rejection of the null hypothesis for an interaction effect of variability in appearance and variability of context on AUC ($F(1,332) = 0.0883$, $p < 0.001$).

The final sample comprised 84 subjects (58 women, 26 men; mean age = 23.38, SD = 4.76) recruited using the participant pool of the Psychology department of the University of Liege (N = 50), social media or word of mouth (N = 34). They all had normal or corrected-to-normal vision. Students recruited via the participant pool were granted academic credits after completion of the experiment. Data from participants with a mean response time above or below 2.5 SD around the global mean of the participants were excluded and replaced by new ones (N = 16).

Stimuli

Stimuli of faces consisted of pictures of sixteen young Caucasian women with neutral expression originating from the Stirling/ESRC 3D Face Database (“pics.stir.ac.uk.”) (see Supplementary Materials for selection criteria). Half of them were selected as targets whereas the remaining were used as foils. For each target woman, we picked a batch of twelve photos varying in lighting and viewpoint in the database. Six photos (2 left angled view, 2 right angled view and 2 frontal view) were used for the learning phase whereas the remaining six were used for the recognition task. Variability in appearance was introduced by changing the hairstyle of the identity across the set of six to-be-learned images. For each actress, six different hairstyles were produced in addition to the original one by pasting six distinct wigs photographed in a matching orientation on the faces, using the GIMP software (The GIMP

threshold determined by the Pocock correction, rather than the opposite. After correction, the required total sample size reached 213 participants (i.e. 71 per step).

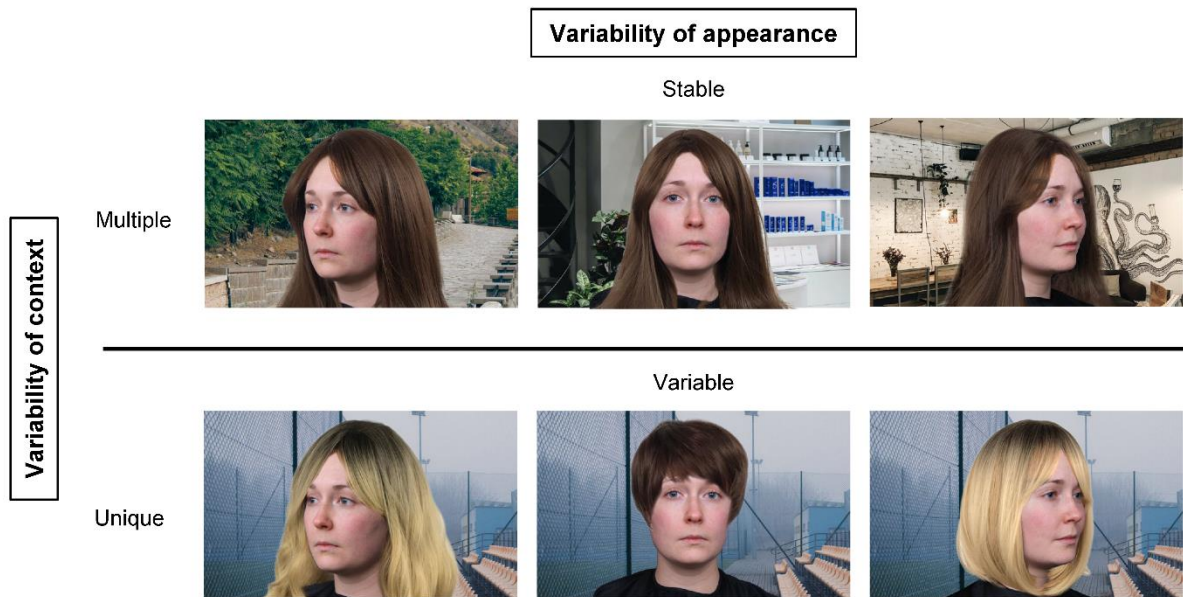
Development Team. (2019)), resulting in seven images with different hairstyle per identity. Six images were used in the variable condition (including the natural hairstyle) and the remaining one was used in the stable condition. There were four distinct sets of six wigs. Each one was used on two actresses paired based on their visual similarity (i.e. overall head shape, eye and eyebrow colour), so that one actress in a pair was learned as variable and the other one as stable, which was counterbalanced across participants. In the stable condition, we used images with a wig rather than natural hair to minimise potential confounds related to differences in hair characteristics between the stable and variable conditions. In addition, distinctive features (e.g. moles, pimples, piercings) were edited out and images were resized to approximately 500 x 450 px (13.0 x 11.7 degrees of visual angle) so that internal features would be within parafoveal vision (approximately 300 x 300 px; 8° x 8°). The original background was replaced by a transparent field, allowing faces to be superimposed onto background scenes (see **Fig. 1**). Face-background pairings were assigned pseudo-randomly (see Procedure).

The scene stimuli consisted of 24 indoor and 24 outdoor scenes, selected from the “Pexels” website (<https://www.pexels.com>). Scenes were resized to 750 X 450 px (19.5° x 11.7°).

The test set consisted of six pictures of each target identity, all of which were different from those used during the learning phase (i.e. different combinations of lighting and view, with a slightly different angle either way compared to learning images), along with six pictures of a foil identity. Combinations of viewpoints and lightings were identical among the pictures of targets and foils. Test faces were cropped around the chin and the forehead and displayed on a grey background, with any visible hair concealed with the same grey (see **Fig. 2**). Image size was standardised to 600 x 600 px (15.7° x 15.8°).

Fig. 1

Example of stimuli used in the learning phase of Experiment 1 and Experiment 2



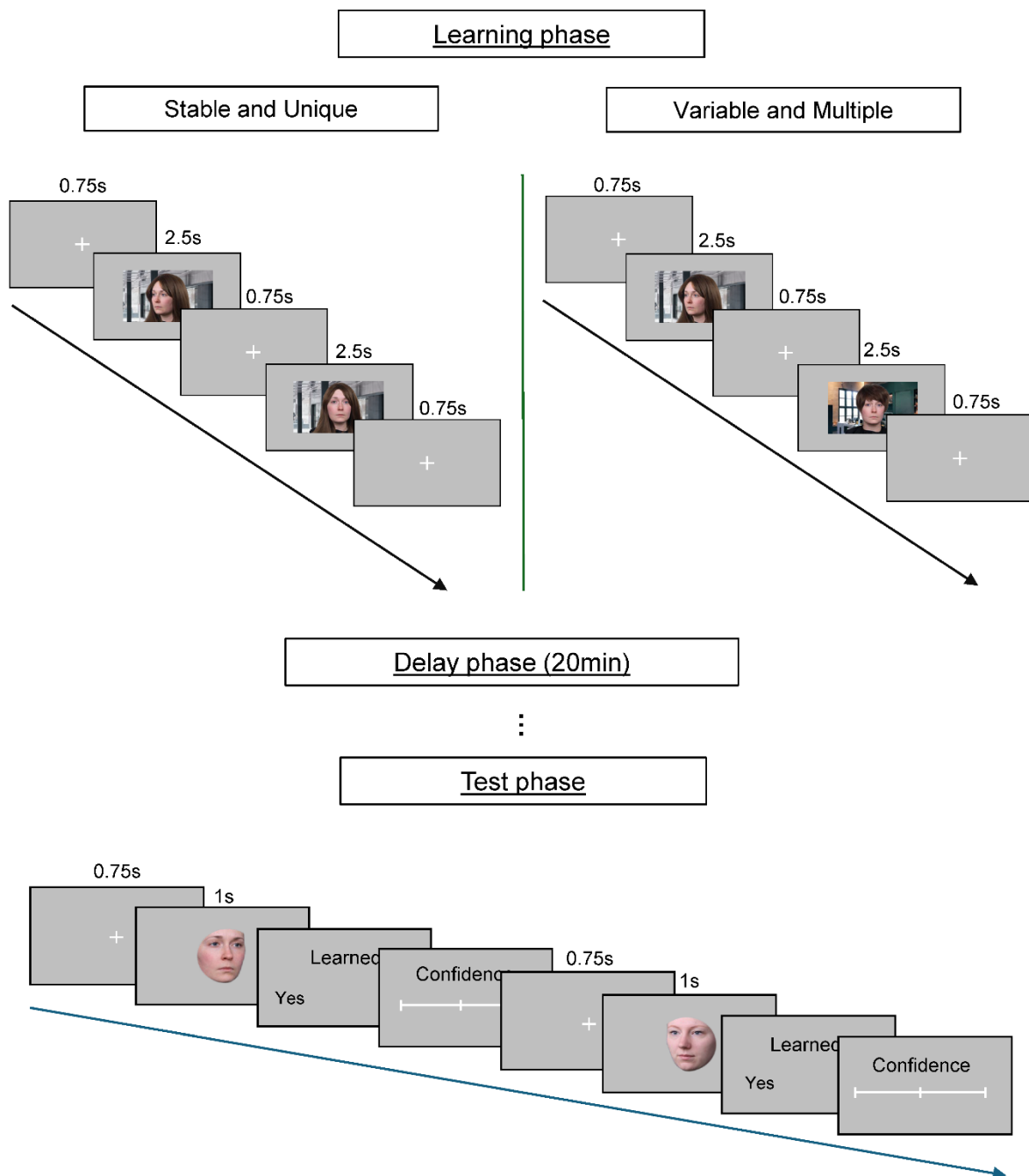
Notes. Images on the top row show an actress in the stable appearance and multiple context conditions (same hairstyle, different scenes across pictures). Images on the bottom row show the same actress in the variable appearance and unique context conditions (different hairstyle, same scene across pictures). In both experiments, participants learned eight identities via six images each. The combination of conditions in which each actress was learned was counterbalanced across participants.

Procedure

Participants were tested individually in a windowless laboratory. Their head was maintained at a distance of 60 cm from the screen by means of a chin rest. Tasks were programmed on Python 3.8 using Psychopy packages (Peirce et al., 2019) and presented on a 52.7x29.65 cm LCD Philips monitor, with a screen resolution of 1,920 pixels × 1,200 pixels and a refresh rate of 60 Hz. Eye movements were recorded with a Tobii Pro X3-120 eye-tracker attached at the bottom of the screen, to investigate the distribution of overt attention

during learning. The Tobii Pro X3-120 eyetracker was set at a sampling rate of 120Hz to track binocular eye movements of the participants at 60cm from the screen. The eye tracker was calibrated using the Tobii Pro Eye Tracker Manager with an 8-point calibration procedure on a grey background (matching the default background of PsychoPy). Under optimal conditions, the apparatus has an accuracy of 0.24° . The experiment consisted of two main phases (learning and test; see **Fig. 2**) separated by a twenty-minute delay and followed by a rating phase.

During the learning phase, participants were explicitly instructed to learn the face of eight women who would appear within a scene, without further indication about these latter. Each woman was presented via an individual learning block of six consecutive images presented in a random order. Within a learning block, a central fixation cross triggered the presentation of the stimulus image upon fixation for 750ms. Each stimulus was then shown for 2500ms. The eight individual learning blocks were presented in a random order. Half of the identities were learned with different hairstyles (variable condition) and the other half with the same hairstyle (stable condition). In addition, half of the identities in each appearance condition were shown against one single scene (unique context condition), and the other half against six different scenes (multiple context condition; see **Fig. 1**). The pairing between identities and scenes was determined pseudo-randomly, so that each scene was never attributed twice, and with the following constraints. In the unique context condition, faces were presented on the same indoor (two identities) or outdoor scene (two identities) across presentations. In the multiple contexts condition, faces were shown on a three different outdoor and three different indoor scenes on each presentation. Finally, the number of individual association between identities and scenes was equivalent across participants.

Fig. 2*Illustration of the experimental procedure*

Notes. The learning phase presents eight women identity with one or multiple hairstyles (stable, variable appearance) in one or multiple scenes (unique, multiple context) across pictures. Two resulting combinations are shown in the top panel. The delay phase consists of

a 20min video devoid of face. The test phase consists of an Old/New recognition task followed by a confidence assessment.

In sum, participants learned two identities in each combination of conditions and each identity appeared in the different conditions in a counterbalanced manner across participants to rule out any uncontrolled effect of idiosyncratic facial features. Following learning, participants watched a twenty-minute nature video devoid of human faces, which served as delay to ensure the investigation of long-term memory.

During the recognition phase, participants' memory was tested with an old/new recognition task. Test stimuli were 48 novel images of cropped inner facial features of the eight actresses and 48 images of eight foil identities. Those 96 images were equally distributed into two experimental blocks of 48 trials each. Each trial started with a cross that needed to be fixated for 750ms to trigger the presentation of a test image for 1s. Participants were asked to respond as fast and as accurately as possible by pressing the "z" or "p" keys on the keyboard with their index fingers. If no response was recorded, the stimulus was replaced by a grey screen with the question "Learned face?" ("Visages appris ?" in French). The keys for "Yes" and "No" responses were counterbalanced across participants with the same handedness. After a response was made, participants rated their confidence in their response via a mouse click on a 1-3 Likert scale. Finally, an instruction to reposition fingers on the response keys terminated the trial. The task was preceded by a training recognition session with images of teddy bears to familiarize the subjects with the task.

Immediately after the recognition test, we asked participants to rate how dynamic, bold, and interesting each identity seen during the experiment seemed to them on a 7-point Likert scale. This task was implemented for exploratory purposes and will not be discussed further.

Finally, we conducted a brief informal post-experimental interview, asking participants how they processed variability across the learning images.

Eyetracking data processing

We averaged the position of each binocular recording to get single middle gaze points and attributed an event label (“fixation” or “saccade”) to each sample using an adapted python version of the MAD adaptative algorithm developed by Voloh et al. (2020). For each trial, the algorithm iteratively calculates an adaptative velocity threshold for saccade detection based on the median and median absolute deviation of the putative fixation velocities of the trial. First, we linearly interpolated missing positions under 25ms of missing values. Then, we applied a Savitsky-Golay filter to calculate positions, velocities and accelerations and determined a velocity threshold. The onset and offset of each saccade were established as the first sample that exceeded or dropped below the velocity threshold adjusted for the local noise surrounding the saccade. All samples between saccades were classified as fixations and subsequently grouped into individual fixation events. In order to eliminate noise, fixation events below 80 ms were excluded (i.e. 2.60% of fixations)(Nyström & Holmqvist, 2010). For each learning picture, we defined three areas of interest (AOI) as geometrical zones using the free software Labelme (<https://labelme.io>, Russell et al., 2008). We used the label “face” to designate inner facial features, “hair” for hairstyles and “scene” for the background scene. Recorded points that fell outside or on the edges between these areas were labelled “other” (these represented 2.27% of total fixations). In this manner, each individual fixation event was assigned an area of interest corresponding to the area most frequently fixated across its constituent samples. Then, we calculated the percentage of fixation and saccade durations within each AOI to facilitate comparisons across conditions.

To ensure that individual participants had sufficient data, we calculated the standard deviation and average ratio of missing samples pre and post interpolation and the ratio of

interpolated data across all files. We then counted the number of trials per participant that exceeds 2SD above the mean of at least one of these criteria, with resulting threshold being 30,1% of missing data before interpolation, 21,1% of data interpolated, 17.6% of missing data after interpolation. This enabled us to exclude participants with a number of trials greater than 2SD above the mean number of trials excluded ($M = 5.70$; $SD = 9.33$) across participants. We also decided to exclude participants who had less than 8 valid trials out of 12 per condition. This resulted in the exclusion of 16 participants. All the trials from the remaining 68 participants were retained for the eyetracking analyses. The preprocessing code is publicly available on OSF (osf.io/msruw).

Results

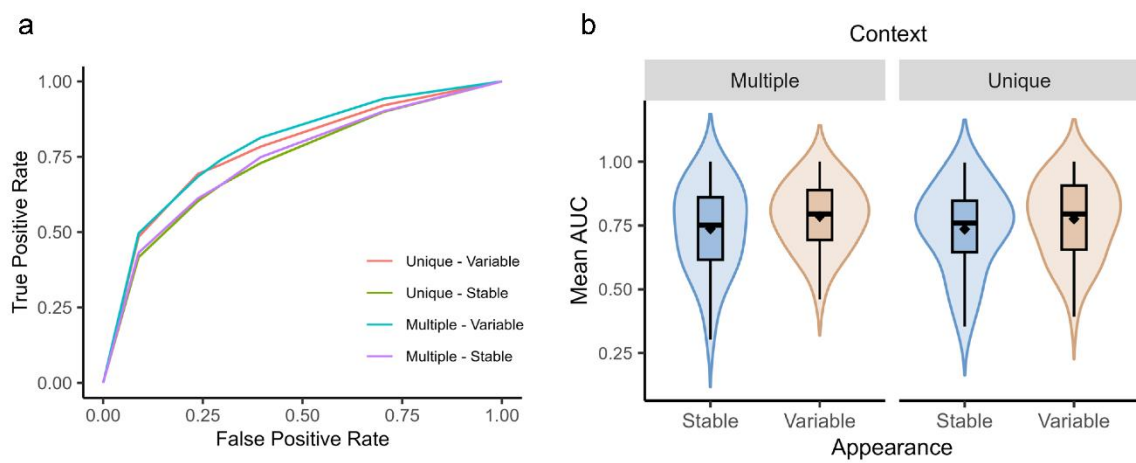
For each participant and condition, ROC curves were calculated with R using the pROC library (Robin et al., 2011) with response and confidence ratings as input. Response times (RT) below 300ms or exceeding three standard deviations from the participant's mean were treated as outliers. This resulted in the exclusion of 152 trials (i.e. 1.88%). Recognition performance was analysed using a two-way repeated measures Analysis of Variance (ANOVA) with variability in appearance (stable, variable) and context (unique, multiple) as within-subject factors on mean Area Under ROC Curves (AUC) and mean correct response time. As a reminder, the corrected alpha for of the analyses is $\alpha=0.0294$. In addition to pre-planned analyses, criterion data and proportion of fixation durations on each AOI were also analysed to investigate response bias and explore the distribution of attention across variability levels during learning. For the sake of space and clarity, additional analyses are reported as Supplementary Information. These include confidence ratings analyses using repeated measures ANOVA, as well as complementary generalized linear mixed-effects models analyses on AUC, correct response times and hit rates with fixation duration on inner features as cofactors.

Accuracy

Results are shown on **Fig. 3**. As predicted, there was a significant main effect of appearance, with a substantial effect size. AUCs were larger for variable faces ($M_{variable} = 0.78$, $SD_{variable} = 0.14$) than for stable faces ($M_{stable} = 0.74$, $SD_{stable} = 0.17$), $F(1,83) = 11.98$, $p < 0.001$, $\eta^2_p = 0.13$. However, in contrast with our hypotheses, there was no main effect of context, faces associated with multiple contexts ($M_{multiple} = 0.76$, $SD_{multiple} = 0.15$) were not significantly recognized better than faces associated with a unique context ($M_{unique} = 0.76$, $SD_{unique} = 0.16$), $F(1,83) < 1$, $p = 0.65$, $\eta^2_p < 0.01$, and the interaction was not significant, $F(1,83) < 1$, $p = 0.77$, $\eta^2_p < 0.01$.

Fig. 3

Accuracy results of Experiment 1



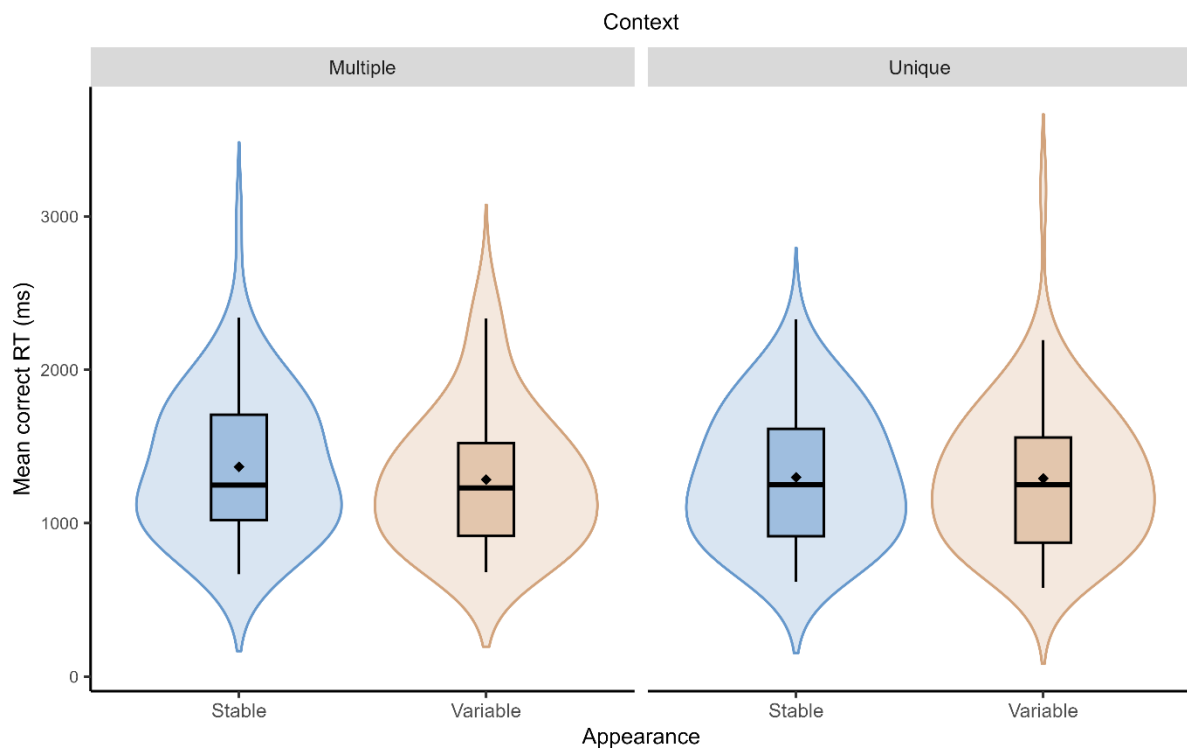
Notes. (a) ROC curves and (b) Mean Area Under ROC curve (AUC) per appearance and context condition within a blocked learning procedure. Mean values are represented by the black dot, boxplots show the distribution in quartiles, while the violins' width represents performance distributions across participants.

Correct response time

Results are shown on **Fig. 4**. Two participants were not taken into account in this analysis because they had no correct response in one of the conditions. There was no significant main effect of appearance ($M_{stable} = 1300.55\text{ms}$, $SD_{stable} = 660.99\text{ms}$, $M_{variable} = 1266.53\text{ms}$, $SD_{variable} = 655.60\text{ms}$), $F(1,81) = 1.56$, $p = 0.22$, $\eta^2p < 0.01$, or of context ($M_{unique} = 1263.06\text{ms}$, $SD_{unique} = 644.02\text{ms}$, $M_{multiple} = 1302.18\text{ms}$, $SD_{multiple} = 671.80\text{ms}$), $F(1,81) = 0.34$, $p = 0.56$, $\eta^2p < 0.01$, on correct RT. There was also no significant interaction between appearance and context, $F(1,81) = 2.06$, $p = 0.16$, $\eta^2p = 0.02$.

Fig. 4

Correct response time results of Experiment 1



Notes. Mean response time per appearance and context condition within a blocked learning procedure. Mean values are represented by the black dot, boxplots show the distribution in quartiles, while the violins' width represents performance distributions across participants.

Response bias

We examined criterion *c* using the same ANOVA. Participants were more conservative for stable faces ($M_{stable} = 0.06$, $SD_{stable} = 0.50$) than variable faces ($M_{variable} = -0.05$, $SD_{variable} = 0.46$), as indicated by a significant main effect of appearance with a moderate effect size, $F(1,83) = 9.54$, $p = 0.003$, $\eta^2p = 0.10$. However, there was no significant main effect of context ($M_{unique} = 0.01$, $SD_{unique} = 0.49$, $M_{multiple} = -0.01$, $SD_{multiple} = 0.47$, $F(1,83) < 1$, $p = 0.73$, $\eta^2p < 0.01$) and no significant interaction, $F(1,83) < 1$, $p = 0.86$, $\eta^2p < 0.01$.

Eyetracking data

We analysed proportion of fixation durations using linear mixed-effects models rather than repeated-measures ANOVA, as mixed models provide greater flexibility in handling missing data and unbalanced designs. To assess whether Appearance and Context influenced the distribution of visual attention across areas of interest (AOIs), we first fitted a global model including AOI as a factor along with its interactions with Appearance and Context. However, this model showed substantial multicollinearity and instability (e.g., high variance inflation factors and boundary fit), which limited the interpretability of higher-order interactions. Therefore, to obtain more stable and interpretable estimates, we conducted separate linear mixed-effects models for each AOI. To account for multiple comparisons across AOIs, *p*-values were adjusted using the Holm–Bonferroni correction. For completeness, the results of the global model are reported in the Supplementary Materials.

For each AOI, we first fitted a model including the main effects of variability in appearance and context as fixed effects and participants' intercept as a random effect. We then added their interaction to test whether the effect of one factor depended on the other. First, we found that variability in appearance was a good predictor of the proportion of time spent on inner features ($\beta = -0.0223$, $SE = 0.007$, $t(3194) = -3.32$, $p = 0.002$) with stable faces

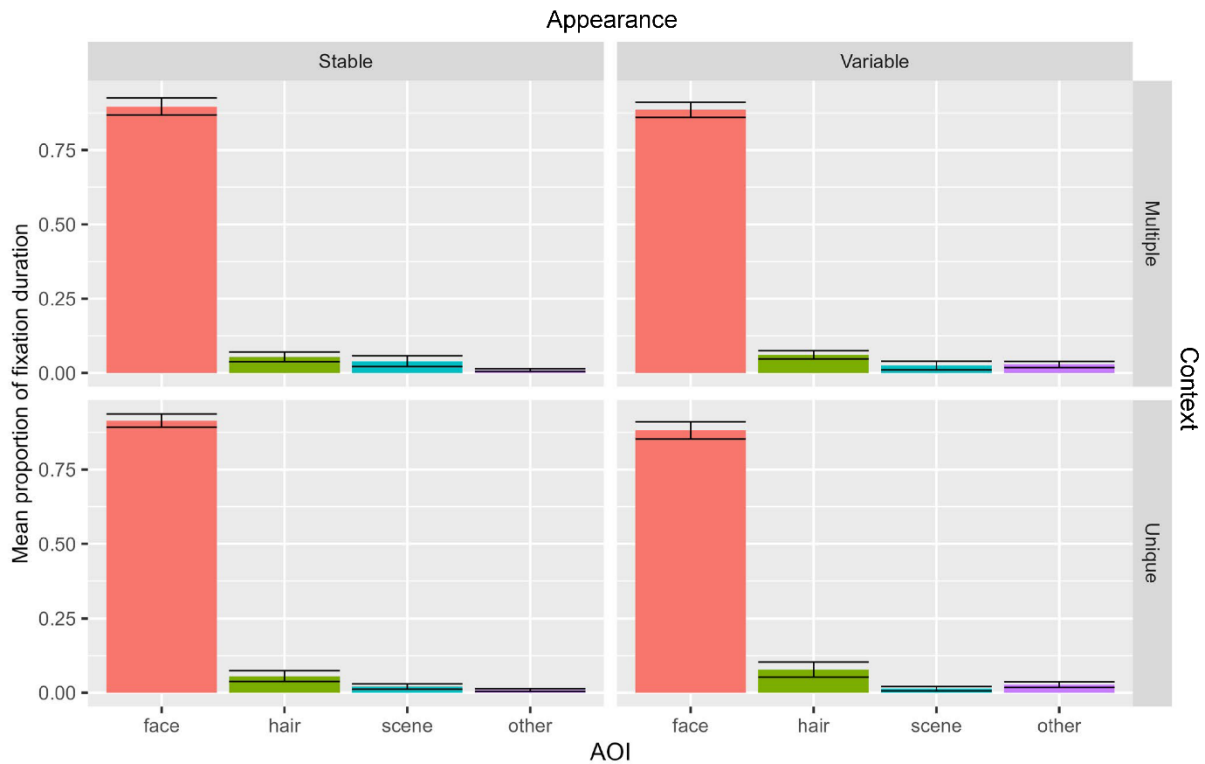
($M_{stable} = 0.91$, $SD_{stable} = 0.18$) fixated more than variable faces ($M_{variable} = 0.88$, $SD_{variable} = 0.23$), but we found no significant effect of variability of context ($\beta = -0.007$, $SE = 0.007$, $t(3194) = 1.07$, $p = 0.57$). There was no interaction between appearance and context ($\beta = -0.022$, $SE = 0.013$, $t(3193) = -1.65$, $p = 0.40$). Second, variability in appearance was also a good predictor of the ratio of time spent on the hair ($\beta = 0.014$, $SE = 0.005$, $t(3194) = 2.84$, $p = 0.004$) with hair of stable faces ($M_{stable} = 0.06$, $SD_{stable} = 0.14$) fixated less than the hair of variable faces ($M_{variable} = 0.08$, $SD_{variable} = 0.17$), suggesting that changing hairstyles attracted the eyes at the expense of inner features. We found no significant effect of context ($\beta = -0.009$, $SE = 0.005$, $t(3194) = 1.85$, $p = 0.19$) and no significant interaction ($\beta = 0.015$, $SE = 0.010$, $t(3193) = 1.50$, $p = 0.40$) on the fixation duration on hair. Third, variability in appearance ($\beta = -0.011$, $SE = 0.003$, $t(3194) = -3.89$, $p < 0.001$) was a good predictor of proportion of time spent on scenes ($\beta = -0.015$, $SE = 0.003$, $t(3194) = -5.47$, $p < 0.001$) with stable faces' scenes ($M_{stable} = 0.03$, $SD_{stable} = 0.10$) fixated more than variable faces' scenes ($M_{variable} = 0.02$, $SD_{variable} = 0.08$). We also found a significant effect of variability of context ($\beta = -0.015$, $SE = 0.003$, $t(3194) = -5.47$, $p < 0.001$) with scenes in multiple condition ($M_{multiple} = 0.03$, $SD_{multiple} = 0.11$) fixated more than scenes in unique condition ($M_{unique} = 0.02$, $SD_{unique} = 0.07$). The interaction was not significant, ($\beta = 0.008$, $SE = 0.006$, $t(3193) = -1.37$, $p = 0.40$).

We also checked whether our experimental conditions influenced the ratio of time spent outside of the AOIs, expecting no influence. Nevertheless, we found a significant effect of appearance ($\beta = 0.018$, $SE = 0.004$, $t(3194) = 4.78$, $p < 0.001$) with stable faces ($M_{stable} = 0.01$, $SD_{stable} = 0.06$) associated with less fixations outside of the image than variable faces ($M_{variable} = 0.03$, $SD_{variable} = 0.14$). This result moderate the conclusion that participants' attention was solely directed to the hair whenever it was variable. No significant effect of

context ($\beta = -0.001$, $SE = 0.004$, $t(3194) = -0.32$, $p = 0.75$) nor significant interaction ($\beta = -0.001$, $SE = 0.008$, $t(3193) = -0.13$, $p = 0.90$) were found. Results are shown in **Fig. 5**.

Fig. 5

Time spent on each AOI per condition in Experiment 1



Notes. The “face” area refers to the inner facial features, “hair” to the hairstyle, “other” to the boundary between hair and faces or neck and shoulders, “scene” to the background scene.

Error bars show 95% CI.

Discussion

In Experiment 1, we examined the distinct effect of appearance and context on accuracy and recognition speed. As expected, and in accordance with previous face learning studies suggesting that variability allows face representation to move beyond image-level properties (Baker et al., 2017; Corpuz & Oriet, 2022; Hancock et al., 2000; Kramer, 2021;

Murphy et al., 2015), faces with a variable appearance were better recognized than stable ones. The cost-efficient encoding hypothesis suggests that variations in appearance induce a more resolute encoding of the internal features by disengaging resources from changeable external or internal components. However, this disengagement was not observed through eye movement patterns during learning as participants spent less time looking at the internal traits of variable faces than of stable faces, and more time looking at hair for variable than for stable faces. As pictures were shown by block of identity, it is possible that the gaze of the participant was attracted to the hair which was the most salient feature to change across images. In line with Murphy et al. (2015), this suggests that the time spent viewing a face is not as decisive to predict success in recognition and further indicates that inner features of variable faces were somewhat processed more efficiently than inner features of stable faces. In addition, participants were also more prone to a liberal response bias and more confident in telling that a face was old based on inner features when it was learned with variations of appearance compared to a stable appearance (see Supplemental materials; see also Menon et al., 2015).

Contrary to our predictions, performance for faces shown on multiple and unique scenes was similar. This contrasts with previous experiments that showed a reliable influence of contextual information during face learning (Andrews et al., 2015; Dalton, 1993; Smith et al., 2014; Young et al., 1985). However, this null finding aligns with recent findings by Nguyen et al. (2026) in a similar face learning task. They manipulated background variability with naturalistic or artificial images and found it does not improve generalization. Crucially, in their experiment and in the current one, individual faces were learned via separate identity blocks. Such learning blocks could in themselves produce a supervised learning context, reducing the need to process external clues to support learning (Cavazos et al., 2019). We tested this hypothesis in a second experiment, by moving from a blocked presentation of

individual identities to a randomized one. By doing so, we switched from supervised to unsupervised learning as the participants would now need to tell pictures from the same identity together themselves while learning. In this experiment, we expected a unique context and stable hairstyle to serve as identifying cues that would help participants match different images of the same person during learning, thereby providing an anchor for building reliable facial representations. Therefore, we predicted greater recognition performance for unique context than for multiple contexts and for stable appearance compared to variable appearance.

Experiment 2

This second experiment follows the same analyses plan as the first one, except that alpha risk was set back to $\alpha=0.05$ since no sequential analysis was planned. The experimental design and hypotheses were pre-registered on OSF (<https://osf.io/58fsz/>).

Methods

Participants

The final sample size was identical to that used in Experiment 1, i.e. 84 subjects (79 women, mean Age = 19.69, SD = 2.44), which proved powerful enough to discover meaningful effects within our learning paradigm. We increased the exclusion criterion from 2SD to 2.5SD around the mean of response times as the previous criterion was too conservative. This resulted in the replacement of 4 participants.

Material and procedure

The materials and procedure were identical to those used in the first experiment, except that the learning phase presented individual identities in a random order instead of a blocked presentation. Response times (RT) faster than 300ms or exceeding three standard deviations from the participant's mean were identified as outliers. This resulted in the exclusion of 182 trials (i.e. 2.25%). Due to practical organizational constraints, eye tracking

data could not be acquired and fixation crosses between each trial were therefore presented for a fixed duration of 750 ms. The final rating task was also removed.

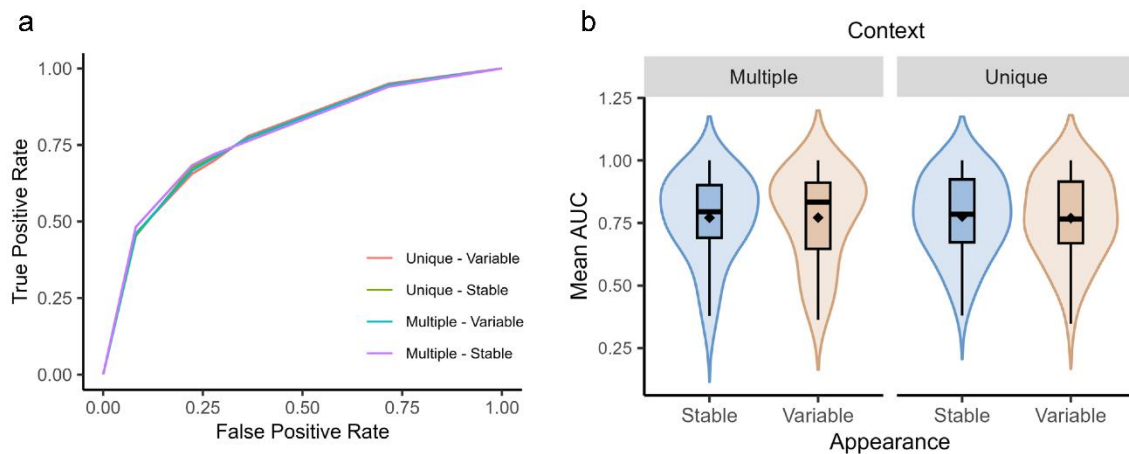
Results

Accuracy

Results are shown in **Fig. 6**. AUC, used to reflect accuracy, for stable faces ($M_{stable} = 0.77$, $SD_{stable} = 0.16$) and variable faces ($M_{variable} = 0.77$, $SD_{variable} = 0.17$), was not significantly different, $F(1,83) < 1$, $p = 0.86$, $\eta^2p < 0.01$. AUC for faces associated with unique ($M_{unique} = 0.77$, $SD_{unique} = 0.16$) and multiple contexts ($M_{multiple} = 0.77$, $SD_{multiple} = 0.17$) was also comparable, $F(1,83) < 1$, $p = 0.91$, $\eta^2p < 0.01$, and there was no significant interaction, $F(1,83) < 1$, $p = 0.81$, $\eta^2p < 0.01$.

Fig. 6

Accuracy results of Experiment 2



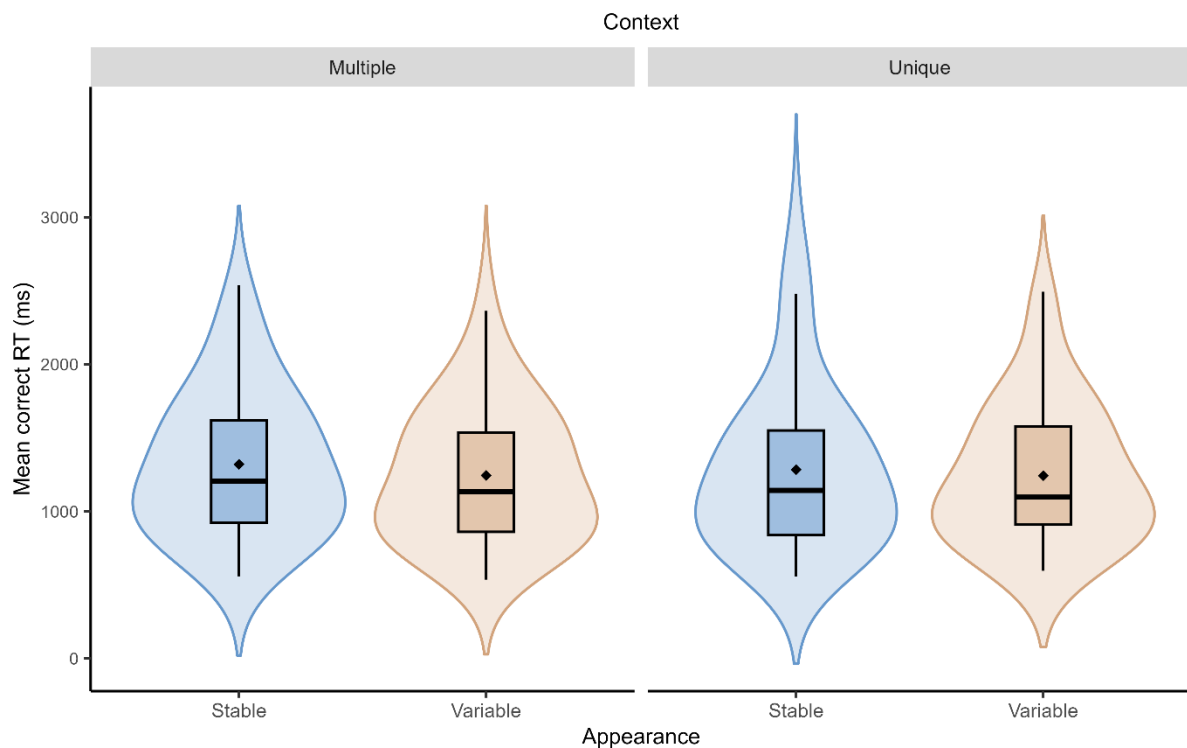
Notes. (a) ROC curves and (b) Mean Area Under ROC curve (AUC) per appearance and context condition within a randomized learning procedure. Mean values are represented by the black dot, boxplots show the distribution in quartiles, while the violins' width represents performance distributions across participants

Correct response time

Results are shown in **Fig. 7**. One participant was not taken into account in this analysis because they had no correct response in one of the conditions. There was no significant main effect of appearance ($M_{stable} = 1256.37\text{ms}$, $SD_{stable} = 705.07\text{ms}$, $M_{variable} = 1219.45\text{ms}$, $SD_{variable} = 656.47\text{ms}$)², $F(1,82) = 3.37$, $p = 0.07$, $\eta^2p = 0.04$, and no significant main effect of context ($M_{unique} = 1229.02\text{ms}$, $SD_{unique} = 701.50\text{ms}$, $M_{multiple} = 1247.22\text{ms}$, $SD_{multiple} = 660.68\text{ms}$), $F(1,82) < 1$, $p = 0.49$, $\eta^2p < 0.01$. There was no significant interaction between appearance and context, $F(1,82) < 1$, $p = 0.40$, $\eta^2p < 0.01$.

Fig. 7

Correct response time results of Experiment 1



² Note that the linear mixed-effect model shown in Supplementary materials indicates a small but significant effect of appearance on correct RTs, with variable faces recognised faster than stable faces. We opted to not discuss this finding further, to maintain a conservative interpretation of our results.

Notes. Mean response time per appearance and context condition within a blocked learning procedure. Mean values are represented by the black dot, boxplots show the distribution in quartiles, while the violins' width represents performance distributions across participants.

Response bias

Participants' criterion was neutral for stable faces ($M_{stable} = 0.001$, $SD_{stable} = 0.52$) and for variable faces ($M_{variable} = 0.02$, $SD_{variable} = 0.49$), with no significant difference between them, $F(1,83) < 1$, $p = 0.67$, $\eta^2p < 0.01$. It was also similarly neutral between unique ($M_{unique} = 0.002$, $SD_{unique} = 0.50$) and multiple contexts ($M_{multiple} = 0.02$, $SD_{multiple} = 0.51$), $F(1,83) < 1$, $p = 0.67$, $\eta^2p = 0.01$. There was no significant interaction between our two variables, $F(1,83) < 1$, $p = 0.85$, $\eta^2p < 0.01$.

Discussion

In Experiment 2, we investigated the influence of variability in appearance and context with a randomized presentation of images of individual identities during learning. Contrary to our predictions, we did not find any significant effect of appearance or of context. Our results are at odds with other lab-based studies that reliably demonstrate a benefit of uncontrolled variability (e.g. Burton et al., 2016; Murphy et al., 2015). However, studies using more ecological learning procedure tend to have more nuanced conclusions. In a seminal study, Devue and de Sena (2023) capitalized on natural exposure to celebrities in everyday life to investigate the effect of real-life variability on face learning. They asked participants to recognize the inner features of celebrities according to their fame and their propensity to change appearance. For the most famous celebrities, they found that the exposure to more variations in appearance helps to build a more reliable representation of their internal facial features compared with the exposure to less variability. However, among the lesser-known celebrities, they found the opposite pattern suggesting that stability is more

helpful in the early stages of familiarisation, but that variability becomes more valuable at later stages. In the current experiment, we tested the recognition of unfamiliar faces at an early stage of familiarisation but with the explicit incentives to focus on learning the faces with a short delay between learning and recognition. Therefore, it is possible that the testing phase occurred at the point when familiarisation shifted from stability-reliant to variability-sensitive, thus leading to no observable difference between conditions.

Contrary to our predictions, the performance for faces associated with unique and multiple contexts during learning remained comparable. This could indicate that the new learning procedure did not imbued background scenes with the expected additional value of visual learning anchor. In an informal post-experimental survey, a considerable number of participants reported that they voluntarily ignored the context because it seemed distracting. Some of them also reported that they started ignoring hairstyle after realising that it could change from one image to another. This raises the question of whether the randomisation yield other strategies than the hypothesised learning anchor ones, and whether performance has improved or decreased in Experiment 2 compared to Experiment 1.

Cross experiment analyses

We conducted additional exploratory analyses to compare recognition performance between blocked learning (Experiment 1) and randomized learning (Experiment 2). The goal was to assess the influence of learning procedure on face recognition and its interaction with appearance. Two scenarios could account for the differences in appearance variability effects observed across experiments. First, performance for variable faces could be superior in Experiment 1 compared to both stable and variable faces in Experiment 2, suggesting that appearance variability benefits face learning only when it is salient, as in blocked learning conditions. Alternatively, performance for stable faces could be lower in Experiment 1 compared to both stable and variable faces in Experiment 2, indicating that stable appearance

may hinder learning, potentially due to reduced engagement across successive similar images. Therefore, in order to compare the two experiments, the data were submitted to a mixed ANOVA with appearance (stable, variable) and context (unique, multiple) as within-subject factors and learning procedure (blocked, randomized) as between-subject factor on our three measures of interest.

Results

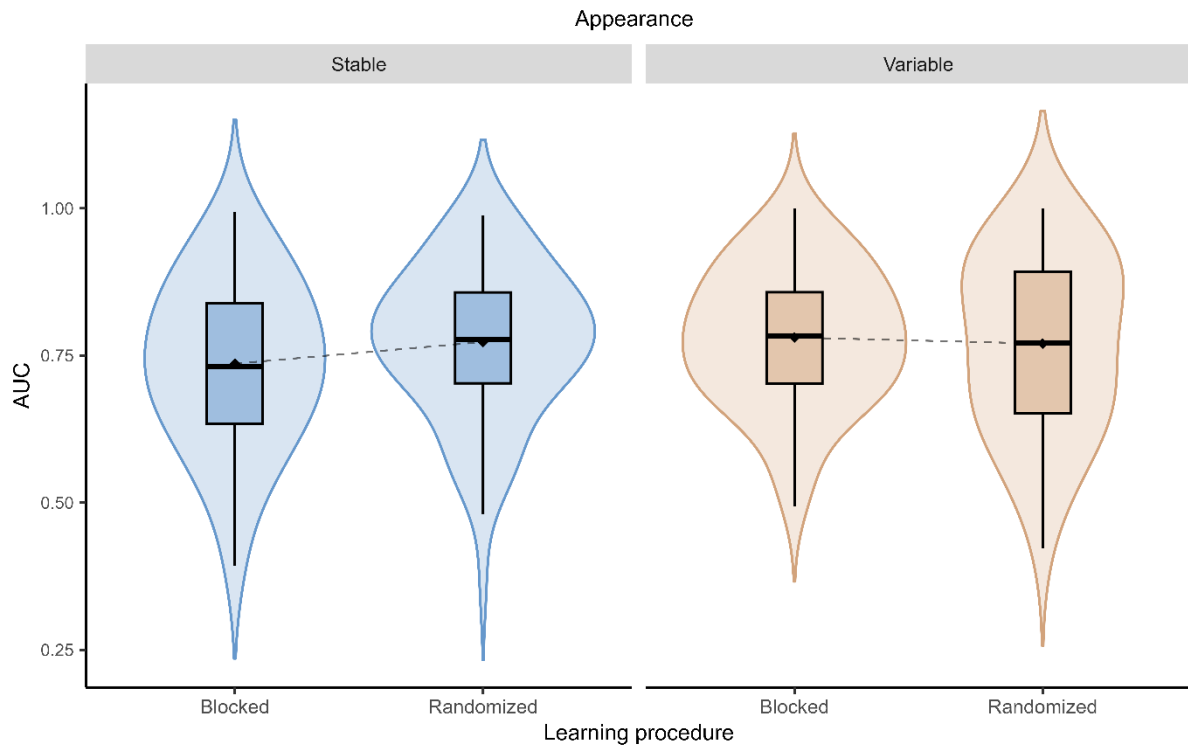
Accuracy

There was no significant main effect of context, $F(1,166) < 1$, $p = 0.84$, $\eta^2p < 0.01$, with comparable mean AUC for unique ($M_{unique} = 0.76$, $SD_{unique} = 0.19$) and multiple contexts ($M_{multiple} = 0.77$, $SD_{multiple} = 0.16$). There was no significant main effect of learning procedure, $F(1,166) < 1$, $p = 0.46$, $\eta^2p < 0.01$, with comparable performance following blocked learning ($M_{blocked} = 0.76$, $SD_{blocked} = 0.15$) and randomized learning ($M_{randomized} = 0.77$, $SD_{randomized} = 0.17$). However, there was a significant main effect of appearance ($M_{stable} = 0.75$, $SD_{stable} = 0.16$, $M_{variable} = 0.78$, $SD_{variable} = 0.16$), $F(1,166) = 4.12$, $p = 0.04$, $\eta^2p = 0.02$, qualified by a modest but significant interaction with learning procedure, $F(1,166) = 5.29$, $p = 0.02$, $\eta^2p = 0.031$. No other two-way interactions were significant (appearance by learning procedure, $F(1,166) < 1$, $p = 0.71$, $\eta^2p < 0.01$, variability of context by learning procedure, $F(1,166) < 1$, $p = 0.71$, $\eta^2p < 0.01$). The three-way interaction was not significant either, $F(1,166) < 1$, $p = 0.97$, $\eta^2p < 0.01$.

We explored the interaction between appearance and learning procedure by means of pairwise t-tests with Bonferroni-Holm correction, which revealed that AUCs for stable faces in blocked learning ($M_{blocked} = 0.74$, $SD_{blocked} = 0.17$) was smaller than AUCs for stable faces in randomized learning ($M_{randomized} = 0.77$, $SD_{randomized} = 0.16$), $t(1,167) = -1.98$, $p = 0.049$, $d = -0.15$. In contrast, AUC for variable faces in both learning procedures were not significantly different, $t(1,167) = 0.66$, $p = 0.51$, $d = 0.05$. Results are shown in **Fig. 8**.

Fig. 8

Accuracy results across experiments, as a function of appearance and learning procedure



Notes. Mean AUC per appearance condition (stable, variable) across blocked (Experiment 1) and randomized (Experiment 2) learning procedure. Mean values are represented by the black dot, boxplots show the distribution in quartiles, dotted lines join the mean AUC values across experiment for each appearance condition, while the violins' width represents performance distributions across participants.

Correct response time

We found a significant effect of appearance with longer response time for stable faces ($M_{stable} = 1276.78\text{ms}$, $SD_{stable} = 670.60\text{ms}$) than for variable faces ($M_{variable} = 1241.09\text{ms}$, $SD_{variable} = 645.26\text{ms}$), $F(1,163) = 4.68$, $p = 0.032$, $\eta^2p = 0.03$. There was no significant main effect of context ($M_{unique} = 1243.07\text{ms}$, $SD_{unique} = 654.72\text{ms}$, $M_{multiple} = 1274.15\text{ms}$, $SD_{multiple} = 660.98\text{ms}$ $F(1,163) < 1$, $p = 0.35$, $\eta^2p < 0.01$) and no significant main effect of learning

procedure ($M_{blocked} = 1281.72\text{ms}$, $SD_{blocked} = 652.91\text{ms}$, $M_{randomized} = 1235.84\text{ms}$, $SD_{randomized} = 662.23\text{ms}$, $F(1,163) < 1$, $p = 0.58$, $\eta^2p < 0.01$). No significant interaction effect was found nor between variability in appearance and context, $F(1,163) = 2.12$, $p = 0.15$, $\eta^2p < 0.01$, nor between variability in appearance and learning procedure, $F(1,163) < 1$, $p = 0.80$, $\eta^2p < 0.01$, nor between variability of context and learning procedure, $F(1,163) < 1$, $p = 0.78$, $\eta^2p < 0.01$. The three-way interaction was not significant either, $F(1,163) < 1$, $p = 0.51$, $\eta^2p < 0.01$.

Response bias

We found no significant main effect of variability in appearance ($M_{stable} = 0.03$, $SD_{stable} = 0.51$, $M_{variable} = -0.02$, $SD_{variable} = 0.48$), $F(1,166) = 2.54$, $p = 0.11$, $\eta^2p = 0.02$. There was also no significant main effect of variability of context ($M_{unique} = 0.004$, $SD_{unique} = 0.50$, $M_{multiple} = 0.006$, $SD_{multiple} = 0.49$), $F(1,166) < 1$, $p = 0.94$, $\eta^2p < 0.01$. There was no significant main effect of learning procedure ($M_{blocked} = 0.0004$, $SD_{blocked} = 0.48$, $M_{randomized} = 0.01$, $SD_{randomized} = 0.51$), $F(1,166) < 1$, $p = 0.87$, $\eta^2p < 0.01$. There was a very small but significant interaction effect between appearance and learning procedure, $F(1,166) = 5.16$, $p = 0.03$, $\eta^2p < 0.01$. However, we found no significant interaction effect of appearance and variability of context, $F(1,166) < 1$, $p \approx 1.00$, $\eta^2p < 0.01$; nor between variability of context and learning procedure, $F(1,166) < 1$, $p = 0.58$, $\eta^2p < 0.01$. The triple interaction was not significant neither, $F(1,166) < 1$, $p = 0.80$, $\eta^2p < 0.01$.

We examined the interaction between appearance and learning procedure using pairwise t -tests with Holm–Bonferroni correction. However, this analysis revealed no significant difference of criterion for stable faces between blocked learning and randomized learning, $t(1,167) = 0.98$, $p = 0.33$, $d = 0.08$, and no significant difference in criterion for variable faces in both learning procedures, $t(1,167) = -1.36$, $p = 0.18$, $d = -0.10$.

Discussion

Across experiments, accuracy was higher in the variable condition than in the stable condition. However, the interaction between appearance and learning procedure indicated that this main effect was driven by a decline in performance for stable faces following the blocked learning procedure used in Experiment 1. Indeed, in Experiment 2, performance for stable faces increased to reach similar levels as with variable faces in both experiments. A key aspect of the second experiment is the unsupervised type of learning resulting from the randomised presentation. This implicitly required from participants to assess whether each picture matched any of the previously encountered identities for which they had started to build representations. We expected that a unique context and consistent hairstyle of stable faces would facilitate this process. In contrast, it seems that the blocked presentations of individual identities in Experiment 1 may have dissuaded participants from attending to internal features in the stable condition, perhaps because successive pictures did not challenge the budding representations of those faces. In other word, following a blocked learning procedure, stability in appearance hindered recognition performance, likely due to a lack of attentional engagement across presentations during learning. An alternative hypothesis is that variability in appearance and randomisation similarly benefit face abstraction but are not cumulative. While our experiment does not adjudicate between these two explanations, our findings suggest that randomization allowed the diagnostic internal features of all faces to be processed with similar efficiency. This interpretation aligns with some participants' self-reports, which indicated that they focused on internal features of all faces once they realized that hairstyles could change. Importantly, the cross-experiment analysis rules out the possibility that images were processed independently during learning due to difficulty matching different instances of the same face, as performance with variable faces was comparable across experiments.

General discussion

Previous face learning studies have provided convergent evidence toward a positive effect of exposure to uncontrolled variations on recognition performance using ambient pictures and videos (Baker et al., 2017; Corpuz & Oriet, 2022; Ritchie & Burton, 2017; Zhou et al., 2022). In the current study, we examined the specific effect of variability in appearance and variability of context, using rich learning materials that incorporated changes in viewpoint and lighting. Specifically, we compared recognition of faces learned with a variable or a stable appearance (variability in appearance, manipulated via hairstyle), against a unique or multiple background scenes (variability of context), in two experiments involving different learning procedures (blocked per identity in Experiment 1, randomised in Experiment 2). While faces were learned via full headshots, we tested recognition with images of inner features devoid of their peripheral features and context. We expected that conditions that promote refinement of inner facial features (i.e. variable appearance and multiple contexts in Experiment 1), or that provide learning anchors (i.e. stable appearance and unique context in Experiment 2) would lead to higher recognition performance. In Experiment 1, participants recognized inner features of variable faces better than those of stable faces, despite spending less time watching them during learning. In contrast, in Experiment 2, recognition performance for inner features was comparable in both appearance conditions. Contrary to our expectations, variability of context during learning did not significantly impact recognition performance in either experiment. Overall, although the effect of appearance depends on the learning procedure, our results suggest that appearance variability significantly contributes to finer encoding of inner features, beyond the benefits of other visual cues known to support the development of rich three-dimensional facial representations (i.e., changes in lighting and viewpoint; Hill et al., 1997; Liu et al., 2009; Longmore et al., 2008; Yamanashi Leib et al., 2014).

In Experiment 1, we employed a learning procedure in which stimuli were blocked by identity, similar to that used in most previous face learning studies (Baker et al., 2017; Corpuz & Oriet, 2022; Reedy & Devue, 2019; Ritchie & Burton, 2017). In this format, learning is effectively supervised, as all instances of a given identity are presented sequentially, without the need to recognize different instances or discriminate them from other identities across presentations. As such, blocked learning creates a context in itself, in which learning may be facilitated compared with real-life situations where identity information is not explicitly provided and must be inferred by the observer. Our results suggest that in this situation, face learning benefits from variability in appearance, which is in line with previous research manipulating within-person variations through ambient images (Andrews et al., 2015; Baker et al., 2017; Bruce, 1994; Corpuz & Oriet, 2022; Dowsett et al., 2016; Kramer, Young, et al., 2018; Matthews et al., 2024; Menon et al., 2015; Reedy & Devue, 2019; Ritchie & Burton, 2017) and with the cost-efficient hypothesis on which we based our predictions. In the absence of variability in appearance, face learning would prioritize global information, such as hairstyle, as these features are not expected to change and are less demanding in terms of cognitive resources. By contrast, exposure to variability in appearance would favour refinement of facial representations toward more costly internal features, explaining the benefit in recognition performance we found for internal facial features. Moreover, our eye-tracking data demonstrate that the benefit from variability on recognition performance occurs despite a decrease in overt attention towards inner facial features during learning. This suggests that variability may promote more exploration of the face, which would help the extraction of finer inner facial diagnostic details, whereas longer fixations on the inner features of stable faces may reflect a holistic processing, overly reliant on coarse facial information (i.e. peripheral features and gross facial configuration).

In Experiment 2, we moved from a blocked to a randomised learning procedure, in which multiple instances of an individual face interspersed amongst others must be matched with a budding representation for learning of that identity to occur. The aim was to investigate the effect of variability in appearance and in context within an unsupervised learning procedure, which better reflects the unpredictability of daily life encounters. By removing supervision, we expected more reliance on large-scale visual cues such as the hairstyle and context in order to match faces together across presentations. More specifically, our prediction was that stable hairstyle and unique context would serve as reliable cues for face matching, which would enable face learning to capitalize better on the exposure to multiple different pictures of a given face and allow a higher level of abstraction compared to variable faces and multiple contexts. Contrary to our predictions and to Experiment 1, we found no significant effect of variability in appearance on recognition of inner facial features. Like in Experiment 1, contextual variability did not significantly affect facial recognition either. In other words, in conditions that better reflect real-world temporal settings, face learning does not necessarily benefit from variability.

Although not planned initially, supplementary cross-experiment analyses indicated that recognition performance for stable faces was higher in Experiment 2 (randomised learning) than in Experiment 1 (blocked learning) and reached levels comparable to those observed for variable faces in both experiments. This interaction reproduces the numerical pattern of results found in Cavazos et al. (2019). In a similar supplementary cross-experiment analysis, they compared two experiments in which the learning procedure (either blocked or randomized) was manipulated, with four different ambient images shown during learning (i.e. multiple images, Experiment 1) or a unique image repeated four times (i.e. single image, Experiment 2). They found a non-significant interaction between image variability and learning procedure, with worse recognition performance in the single image condition only

with a blocked learning procedure. They concluded that variability improves recognition by promoting a representation that supports generalization, provided that participants manage to group images of identities together. Rather than reflecting a benefit of variability for abstraction when learning is blocked, we propose that stable faces in such learning context induce an overreliance on image-specific encoding strategies, accompanied by reduced attentional engagement across instances. Indeed, both Cavazos et al. (2019) and our study show a performance decrease for stable faces rather than an increase for variable faces when switching from randomized to blocked learning. In contrast, randomized presentation appears to mitigate the disadvantage of stable faces by encouraging the re-engagement of resources across presentations and the formation of more robust, identity-level representations. Although exploratory, our findings indicate that multiple sources of variability may interact during face learning, highlighting the importance of examining each factor independently.

In a recent paper, Raviv et al. (2022) identified four types of variability that may promote generalization and abstraction across a wide range of domains (e.g., visual perception, motor learning, or language). Specifically, they distinguish the set size of exemplars presented in a task (i.e. numerosity), the magnitude of differences between exemplars (i.e. heterogeneity), the range of environments in which the exemplars are presented (i.e. contextual diversity) and the variability of presentation order and/or temporal spacing (i.e. scheduling). In line with Raviv et al. (2022), our findings suggest that both heterogeneity, operationalized through variations in hairstyles, and scheduling, manipulated via learning procedure, may lead to finer face learning. Raviv et al. (2022) proposed two mechanisms through which heterogeneity may benefit learning. First, exposure to a large set of diverse images may increase the likelihood that at least one image closely matches the test image. Our findings rule out this first explanation, given that test images included only internal facial features and excluded the hairstyle, which constituted the source of

heterogeneity. Second, exposure to heterogeneity may provide statistical information about the idiosyncratic variability of individual faces, which may help extract their diagnostic stable features and establish accurate identity boundaries. Besides, variations in scheduling may benefit learning by encouraging active retrieval and attention engagement across successive encounters with a face. Although both types of variability may yield similar benefits for recognition performance, their non-additive effects suggest that they rely on distinct yet partially overlapping cognitive mechanisms.

Raviv et al. (2022) also suggested that contextual diversity may facilitate generalization by decoupling stimuli from the context in which they are presented. In our study, we isolated the effect of contextual diversity by showing faces either on a unique or multiple background scenes. The presentation of isolated inner features at test allowed assessing whether these features had been successfully decontextualized. In Experiment 1 using a blocked learning procedure, we expected multiple background scenes to promote the decontextualization of inner facial features, hence improving their recognition, whereas in Experiment 2, we predicted that a unique background would serve as a staple cue to match the instances of a given face together to capitalize on the numerosity of exemplars. However, we did not observe any significant effect of context variability in either experiment. These findings are consistent with a recent face learning study by Nguyen et al. (2026), who reported across five experiments that neither contextual variability nor augmentation (i.e., artificially introduced visual noise applied to faces) impacts face learning. Overall, these results suggests that only heterogeneity pertaining to ingrained features of a person affects face learning.

We note however that our protocol may have biased attention toward the face at the expense of the context, potentially limiting contextual encoding. Indeed, the instructions to remember the face, as well as their prevailing size and position at the centre of the screen,

may have led participants to focus on the face at the expense of the context. In the second experiment, we modified the learning procedure by adopting a randomized procedure of presentation to give greater weight to the context in an implicit way. However, this manipulation had no significant impact on subsequent recognition performance. By adopting instructions that reinforce the association between face and context (Baddeley & Woodhead, 1982) one may uncover an effect of the variability of context as observed in real life. An alternative explanation resides in the dual-processing hypothesis, which stipulates that recognition and recollection rely on different mechanisms. While recollection necessitates the activation of the full contextual representation associated with a memory, recognition is driven exclusively by the familiarity signal elicited by the presented cue (Mandler, 1980; Smith & Vela, 2001). Thus, when the facial signal alone suffices for recognition, consulting contextual cues becomes unnecessary. Overall, our recognition task may thus not be optimal for detecting an effect of contextual variability.

An alternative explanation could have to do with the artificial nature of our contextual manipulation that undermined a potential effect. In real-world settings, encounters in different contexts tend to be distributed over longer time periods. Further, contextual cues are way richer, comprising three-dimensional visual information extending beyond the frame limits of a mere picture, but also information from different sensory modalities, as well as semantic and personally relevant information related to our objectives. In contrast, in our laboratory experiment, multiple contexts associated with each face were presented in successive images with brief temporal intervals between exposures. This condensed presentation of identities may have obscured the facilitating role of contextual cues as a support in face matching during learning. Future studies should examine whether and how the delay between presentations of different face image exemplars interact with the visual context to shape face learning.

Another limitation of our experiment resides in the artificial construction of our facial stimuli, whereby a series of wigs were superimposed on actresses. While we made efforts to match the orientation of our wigs (photographed from different angles) to the orientation of models in the Stirling database, some resulting images may have attracted attention more than others, overemphasising that impact of variability. In addition, one of the variable face images we used for each identity was an original picture showing natural hairstyle. With the exception of one identity, these images depict women with their hair pulled back and their ears visible, which was not the case for any of the photoshopped hairstyles. This additional information may have contributed to the development of more robust representations for variable faces. Hence, although all other hairstyles in the stable and variable conditions were photoshopped wigs, we cannot rule out the possibility that the benefit of variable appearance we found in the blocked learning experiment was partially due to participants seeing the identity with a more natural look. Nevertheless, the fact that stable and variable faces were recognised in a comparable manner in Experiment 2 mitigates that concern, and the cross experiment comparison points to more plausible explanations (i.e. in terms of learning procedure affecting the way stable and variable faces are processed).

A last limitation of our study is its exclusive use of Caucasian female faces. Western cultural norms allow for more variable hairstyle in women than in men, opening the door to recreate these variations in an experiment without introducing confounds due to unexpected or less natural looking appearances. This may however limit the generalizability of our findings, and future studies should investigate whether and how the effect of appearance variability vary across gender.

Conclusion

Variability has consistently been shown to improve face learning in studies using ambient images. The present study demonstrates for the first time that variability in appearance can

specifically elicit more refined facial representations that support recognition, over and beyond other rich sources of variations induced by changes in lighting and viewpoint. In addition, while past face learning studies have predominantly used blocked learning procedures, our findings suggest that learning procedure is an important modulator of face learning that may artificially inflate benefits from variations or artificially disadvantage learning from stable materials that is repeated in quick succession. Taken together, these results provide new insights into the role of variability in face learning and highlight the importance of carefully considering the learning procedures employed in future research.

Public Significance Statement

Our findings show that while changes in visual context play a limited role in a face learning paradigm, experiencing variability in a person's appearance specifically helps build more robust and flexible mental representations, improving our ability to recognize others across changing conditions. However, the way faces are learned—whether exposure to individual identities is grouped or randomised—strongly shapes how useful these variations in appearance are. Together, these results highlight that both what we see and how we encounter it are critical for effective face learning in everyday life.

Statements and Declarations

Ethical standards

The current study was approved by the Ethics Committee at the University of Liège and informed consent was given by each participant before participation, in accordance with the ethical standards laid down in the [1964 Declaration of Helsinki](#) and its later amendments.

Transparency and Openness

The two experiments of this study were pre-registered on Open Science Framework (<https://osf.io/msruw/>) following an As Predicted template. All the result data, pre-processing and analyses scripts from both experiments are accessible following the same link. We were unable to share face and scene stimuli due to copyright, but the modified face stimuli were sent back to the Stirling/ESRC 3D Face Database (“pics.stir.ac.uk.”) in accordance with the license agreement.

Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Author contributions

Kevin Nguy: Conceptualization; Methodology; Software; Formal analysis and investigation; Writing – original draft preparation; Writing – review and editing; Visualization; Supervision; Data Curation; Project administration.

Tom Lapadula: Formal analysis and investigation.

Christel Devue: Conceptualization; Methodology; Writing – review and editing; Visualization; Supervision; Project administration; Funding acquisition.

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.
<https://doi.org/10.1080/17470218.2014.1003949>
- Baddeley, A. D., & Woodhead, M. (1982). Depth of processing, context, and face recognition. *Canadian Journal of Psychology / Revue canadienne de psychologie*, 36(2), 148-164. <https://doi.org/10.1037/h0080635>
- 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. Scopus.
<https://doi.org/10.1016/j.cognition.2016.12.012>
- Bruce, V. (1982). Changing faces : Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, 73(1), 105-116.
<https://doi.org/10.1111/j.2044-8295.1982.tb01795.x>
- Bruce, V. (1994). Stability from variation : The case of face recognition the M.D. Vernon memorial lecture. *The Quarterly Journal of Experimental Psychology Section A*, 47(1), 5-28. <https://doi.org/10.1080/14640749408401141>
- Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition : The power of averages. *Cognitive Psychology*, 51(3), 256-284.
<https://doi.org/10.1016/j.cogpsych.2005.06.003>
- Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity From Variation : Representations of Faces Derived From Multiple Instances. *Cognitive Science*, 40(1), 202-223. <https://doi.org/10.1111/cogs.12231>

- 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. <https://doi.org/10.1111/1467-9280.00144>
- Cavazos, J. G., Noyes, E., & O'Toole, A. J. (2019). Learning context and the other-race effect : Strategies for improving face recognition. *Vision Research, Face perception: Experience, models and neural mechanisms*, 157, 169-183. <https://doi.org/10.1016/j.visres.2018.03.003>
- Corpuz, R. L., & Oriet, C. (2022). Within-Person Variability Contributes to More Durable Learning of Faces. *Canadian Journal of Experimental Psychology*, 76(4), 270-282. Scopus. <https://doi.org/10.1037/cep0000282>
- Davies, G., & Milne, A. (1982). Recognizing faces in and out of context. *Current Psychology*, 2(1), 235-246. <https://doi.org/10.1007/BF03186766>
- de Fockert, J., & Wolfenstein, C. (2009). Short article : Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, 62(9), 1716-1722. <https://doi.org/10.1080/17470210902811249>
- Devue, C., & de Sena, S. (2023). The impact of stability in appearance on the development of facial representations. *Cognition*, 239, 105569. <https://doi.org/10.1016/j.cognition.2023.105569>
- Devue, C., Wride, A., & Grimshaw, G. M. (2019). New insights on real-world human face recognition. *Journal of Experimental Psychology: General*, 148(6), 994-1007. <https://doi.org/10.1037/xge0000493>
- 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. <https://doi.org/10.1080/17470218.2015.1017513>

Gruppuso, V., Lindsay, D. S., & Masson, M. E. J. (2007). I'd know that face anywhere!

Psychonomic Bulletin & Review, 14(6), 1085-1089.

<https://doi.org/10.3758/BF03193095>

Hancock, P. J. B., Bruce, V., Burton, A. M., Hancock, P. J. B., Bruce, V., Burton, A. M.,

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, 4(9), 330-337. [https://doi.org/10.1016/S1364-](https://doi.org/10.1016/S1364-6613(00)01519-9)

[6613\(00\)01519-9](https://doi.org/10.1016/S1364-6613(00)01519-9)

Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in

face recognition. *Cognition*, 62(2), 201-222. [https://doi.org/10.1016/S0010-](https://doi.org/10.1016/S0010-0277(96)00785-8)

[0277\(96\)00785-8](https://doi.org/10.1016/S0010-0277(96)00785-8)

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1571), 1671-1683.

<https://doi.org/10.1098/rstb.2010.0379>

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition*, 121(3), 313-323.

<https://doi.org/10.1016/j.cognition.2011.08.001>

Koca, Y., & Oriet, C. (2023). From Pictures to the People in Them : Averaging Within-

Person Variability Leads to Face Familiarization. *Psychological Science*, 34(2),

252-264. <https://doi.org/10.1177/09567976221131520>

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.

<https://doi.org/10.1177/0301006617725242>

- Kramer, R. S. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person : A route to face learning. *Journal of Vision*, 15(4), 1. <https://doi.org/10.1167/15.4.1>
- Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*, 172, 46-58. <https://doi.org/10.1016/j.cognition.2017.12.005>
- Lakens, D., Pahlke, F., & Wassmer, G. (2021). *Group Sequential Designs : A Tutorial*. OSF. <https://doi.org/10.31234/osf.io/x4azm>
- Lenth R (2025). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.11.2-80001, <https://rvlenth.github.io/emmeans/>.
- Liu, C. H., Bhuiyan, Md. A.-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. <https://doi.org/10.1037/a0013710>
- Liu, C. H., Chen, W., & Ward, J. (2015). Effects of exposure to facial expression variation in face learning and recognition. *Psychological Research*, 79(6), 1042-1053. <https://doi.org/10.1007/s00426-014-0627-8>
- Liu, C. H., Chen, W., Ward, J., & Takahashi, N. (2016). Dynamic Emotional Faces Generalise Better to a New Expression but not to a New View. *Scientific Reports*, 6(1), 31001. <https://doi.org/10.1038/srep31001>
- 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. <https://doi.org/10.1037/0096-1523.34.1.77>
- Mandler, G. (1980). Recognizing : The judgment of previous occurrence. *Psychological Review*, 87(3), 252-271. <https://doi.org/10.1037/0033-295X.87.3.252>

- Matthews, C. M., Ritchie, K. L., Laurence, S., & Mondloch, C. J. (2024). Multiple images captured from a single encounter do not promote face learning. *Perception*, 53(5-6), 299-316. <https://doi.org/10.1177/03010066241234034>
- Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces : Evidence from a matching task. *Memory & Cognition*, 34(4), 865-876. <https://doi.org/10.3758/BF03193433>
- Menon, N., White, D., & Kemp, R. I. (2015). Variation in Photos of the Same Face Drives Improvements in Identity Verification. *Perception*, 44(11), 1332-1341. <https://doi.org/10.1177/0301006615599902>
- Mileva, M., Young, A. W., Jenkins, R., & Burton, A. M. (2020). Facial identity across the lifespan. *Cognitive Psychology*, 116, 101260. <https://doi.org/10.1016/j.cogpsych.2019.101260>
- 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. <https://doi.org/10.1037/xhp0000049>
- Nguyen, T. H. A., Baker, K. A., Nullmeyer, M. A. L., Damaso, K. L., & Mondloch, C. J. (2026). Noisy variability does not benefit face learning. *Journal of Experimental Psychology: General*. Advance online publication. <https://doi.org/10.1037/xge0001948>
- Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, 128(1), 56-63. <https://doi.org/10.1016/j.cognition.2013.03.006>
- Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1), 188-204. <https://doi.org/10.3758/BRM.42.1.188>

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, 38(15), 2351-2363.

[https://doi.org/10.1016/S0042-6989\(98\)00042-X](https://doi.org/10.1016/S0042-6989(98)00042-X)

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory*, 3(4), 406-417.

<https://doi.org/10.1037/0278-7393.3.4.406>

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2 : Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195-203. <https://doi.org/10.3758/s13428-018-01193-y>

Pexels. (n.d.). Free stock photos. <https://www.pexels.com/>

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, 26(6), 462-483.

<https://doi.org/10.1016/j.tics.2022.03.007>

Reedy, M., & Devue, C. (2019). New perspective on face learning : Stability modulates resolution of facial representations in the optimal observer.

<https://doi.org/10.31219/osf.io/6s8c7>

Righi, G., Peissig, J. J., & Tarr, M. J. (2012). Recognizing disguised faces. *Visual Cognition*, 20(2), 143-169. <https://doi.org/10.1080/13506285.2012.654624>

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

<https://doi.org/10.1080/17470218.2015.1136656>

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M.

(2011). pROC : An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1), 77. <https://doi.org/10.1186/1471-2105-12-77>

- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe : A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77(1), 157-173. <https://doi.org/10.1007/s11263-007-0090-8>
- Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6), 1582-1593. <https://doi.org/10.1037/xlm0000019>
- The GIMP Development Team. (2026). *GNU Image Manipulation Program (GIMP) (Version 3.2.4)* [Computer software]. <https://gimp.org/>
- Toseeb, U., Keeble, D. R. T., & Bryant, E. J. (2012). The Significance of Hair for Face Recognition. *PLOS ONE*, 7(3), e34144. <https://doi.org/10.1371/journal.pone.0034144>
- Voloh, B., Watson, M. R., König, S., & Womelsdorf, T. (2020). MAD saccade : Statistically robust saccade threshold estimation via the median absolute deviation. *Journal of Eye Movement Research*, 12(8). <https://doi.org/10.16910/jemr.12.8.3>
- White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied*, 20(2), 166-173. <https://doi.org/10.1037/xap0000009>
- White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport Officers' Errors in Face Matching. *PLOS ONE*, 9(8), e103510. <https://doi.org/10.1371/journal.pone.0103510>
- Xu, H., & Armony, J. L. (2024). Arousal level and exemplar variability of emotional face and voice encoding influence expression-independent identity recognition. *Motivation and Emotion*, 48(3), 464-483. <https://doi.org/10.1007/s11031-024-10066-1>
- Yamanashi Leib, A., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception : A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, 14(8), 26. <https://doi.org/10.1167/14.8.26>

- Young, A. W., Hay, D. C., & Ellis, A. W. (1985). The faces that launched a thousand slips : Everyday difficulties and errors in recognizing people. *British Journal of Psychology*, 76(4), 495-523. <https://doi.org/10.1111/j.2044-8295.1985.tb01972.x>
- Zhou, X., Vyas, S., Ning, J., & Moulson, M. C. (2022). Naturalistic Face Learning in Infants and Adults. *Psychological Science*, 33(1), 135-151. <https://doi.org/10.1177/09567976211030630>