A cost-efficient face learning mechanism: The impact of stability in appearance on the resolution of facial representations

Christel Devue[1,2]*, Sofie de Sena[2], and Jade wright[2]

[1]Psychology Department, Psychology and Neuroscience of Cognition, University of Liège, Belgium

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

Draft version posted on 21 December 2021. This paper has not been peer reviewed.

Word count: 11,106

Author notes:

*Corresponding author: Psychology Department, Psychology and Neuroscience of Cognition, Place des Orateurs 2, 4000 Liège 1, Belgium ; Phone : +32 366 9282; Email: cdevue@uliege.be.

Image stimuli and datasets for all experiments are available on https://osf.io/8znw5/files/.

Pre-registrations of the study designs and analyses plans are available on the Open Science Framework: Experiments 1 to 3: https://osf.io/qd5y3/register/564d31db8c5e4a7c9694b2be; Amended plans for Experiment 2 and 3: https://osf.io/h5f6s/register/564d31db8c5e4a7c9694b2c0 and https://osf.io/afm4e; Experiment 4: https://osf.io/ethbd/.

We thank Serge Brédart, Arnaud D'argembeau and Valentine Vanootighem for proofreading and commenting on an earlier version of on the manuscript.

Abstract

The way faces become familiar and what information is represented as familiarity develops has puzzled researchers in the field of human face recognition for decades. In this paper, we propose a cost-efficient mechanism of face learning to describe how facial representations form over time and that explains why recognition errors occur. Encoding of diagnostic facial information would follow a coarse-to-fine trajectory, modulated by the intrinsic stability in individual faces' appearance. In four experiments, we draw on a robust and ecological method using a proxy of exposure to famous faces in the real world to test hypotheses generated by the model and we manipulate test images to probe the nature of facial representations. We consistently show that stable facial appearances help create more reliable representation in early stages of familiarisation but that their resolution remains relatively low and therefore less discriminative over time. In contrast, variations in appearance hinder recognition at first but encourage refinement of representations with further exposure. Consistent with the cost-efficient face learning mechanism we propose, facial representations built on a foundation of large-scale coarse information. When coarse information loses its diagnostic value through the experience of variations across encounters, facial details and their spatial relationships receive additional representational weights.

Keywords: face recognition, familiarisation, representational weight, identification, face processing

A cost-efficient face learning mechanism: The impact of stability in appearance on the resolution of facial representations

While most of us recognise a large number of familiar faces effortlessly and with great accuracy (Brédart & Devue, 2006; Devue et al., 2007; Jenkins, Dowsett, & Burton, 2018; Tong & Nakayama, 1999), learning new faces is difficult and highly error-prone (Hancock et al., 2000; Young & Burton, 2018). Experts in the field of face recognition have pointed out that understanding this transition in performance between these extremes is the number one challenge to move research in the area forward (O'Toole et al., 2018; Young & Burton, 2018).

In fact, we know surprisingly little about which facial cues we memorise and draw on to recognise people, and whether and how what we memorise changes over time. Seminal research showed that upon viewing novel faces, we rely by default on external or peripheral features, like hairstyle, even if this strategy is suboptimal and leads to poor recognition performance (Ellis, Shepherd, & Davies, 1979; Young, Hay, McWeeny, Flude, & Ellis, 1985; see also Bruce et al., 2001; Hill et al., 1997; Longmore et al., 2017; White et al., 2014). By contrast, recognition of highly familiar faces would rely on both internal and external features or favour the former (Campbell et al., 1995; Ellis et al., 1979; Kramer, Manesi, et al., 2018). The two categories of features would be part on the same holistic representation (see e.g., Andrews, Davies-Thompson, Kingstone, & Young, 2010), even though when presented by themselves, internal features are judged as more diagnostic of identity than external features alone (Kramer, Manesi, et al., 2018). The way representations transition from a suboptimal reliance on external features to a more optimal reliance on both internal and external features in familiar faces thus remains to be established.

One obstacle to understanding how familiarity with faces develops has been a tendency to study the processing of new faces and familiar faces separately (Burton, 2013). A possible reason for that tendency is an interpretation of observed differences in performance between unfamiliar and familiar faces as the manifestation of qualitatively different processes (for a review, see Johnston & Edmonds, 2009). Upon encounter with new faces, we would form simplistic pictorial representations that do not generalise well to new views and that fail to match new percepts of the same face resulting from changes in lighting or physical appearance (Burton, Bruce, & Hancock, 1999; Longmore et al., 2017). Once familiar, faces would benefit from "face-like" processing—i.e., view-invariant, holistic, or centred on inner-features and their configuration, depending on theories—allowing their recognition despite changes in viewing conditions. However, this dichotomy between unfamiliar and familiar faces may derive from unfair comparisons between them. Unlike familiar faces that have been learned in rich conditions in the real world (e.g., in motion, with changes in lighting and context), unfamiliar faces have traditionally been learned from one or a limited number of photographs in artificial laboratory conditions. We now know that such learning conditions are insufficient to form three-dimensional representations of complex objects like faces and that exposure to multiple viewpoints and/or movement improves learning (Etchells et al., 2017; A. Johnston et al., 2013; Lander & Bruce, 2003; Pilz et al., 2006). Therefore, it is plausible such dichotomy does not apply to real-world situations where we learn new faces in rich circumstances similar to those in which we also encountered faces that have become familiar.

More recently, computational models have refined what plausible mechanisms of familiarisation might entail. They suggest that we come to remember familiar faces by focusing on stable inner features (e.g., eyes, nose, mouth) and ignoring changeable

peripheral ones (Burton, Jenkins, Hancock, & White, 2005; Burton, Bruce, & Hancock, 1999; Jenkins & Burton, 2011; Kramer, Young, & Burton, 2018; Robins, Susilo, Ritchie, & Devue, 2018). Somehow, we would incorporate or average out variations in lighting, viewpoint, appearance, and expression to form robust memory representations that include stable inner aspects and unique ways in which a given face varies. Once formed, these abstract representations would enable recognition of novel instances of an individual (Burton, Kramer, Ritchie, & Jenkins, 2016; Kramer et al., 2018). Principal component analysis (PCA) models predict that the quantity and the quality of variations observers are exposed to should gradually improve recognition performance (Kramer, Young, et al., 2018).

Human data partly support this rationale in recent research focused on the development of familiarity in more ecological conditions. For example, we showed that increased exposure times to faces of actors learned incidentally in a TV show led to linear increases in recognition (Devue et al., 2019). Moreover, familiarisation with new faces in laboratory conditions is facilitated by exposure to large ranges of variations in natural images, mixing environmental (e.g., lighting, background, camera lens, camera angle) and facial (e.g., expression, age, weight, look/appearance) factors, and more so than by mere increases in exposure time (Baker, Laurence, & Mondloch, 2017; Menon, Kemp, & White, 2018; Menon, White, & Kemp, 2015; Murphy, Ipser, Gaigg, & Cook, 2015; Ritchie & Burton, 2017; Robins, Susilo, Ritchie, & Devue, 2018).

However, conclusions drawn from PCA models on the crucial role of inner features are sometimes in conflict with human data. Most strikingly, people occasionally fail to recognise highly familiar people, including themselves, when peripheral features deviate from their usual appearance, even if inner features are clearly visible (Brédart & Young, 2004; Carbon,

2008; Devue et al., 2019; Sinha & Poggio, 1996). Further, some famous individuals are better recognised from their peripheral features alone than from their inner features alone (see Table 1 in Ellis and Davies, 1979). These observations are incompatible with the notion that representations of familiar faces heavily rely on invariant internal features. This inconsistency between human and computer data could occur partly because most computational models ignore peripheral features by design, thereby discounting information that is valuable to humans. The examples above demonstrate that in humans, inner features are not always necessary nor sufficient to trigger recognition of familiar faces and that they do not always carry the most diagnostic information for a given face.

To resolve these apparent contradictions and explain how facial representations evolve as faces become familiar, we propose a parsimonious mechanism of face learning based on cost-efficiency. First, we assume that any feature (e.g., hair colour, ear or nose shape) can be more or less diagnostic of individual facial identity, regardless of its location and of the face's familiarity (see also Abudarham & Yovel, 2018). Rather than systematically relying on a costly encoding of all inner features and their details, representations are weighted based on the relative *stability* of different features over time, which make them more or less diagnostic (e.g., invariable nose vs. changing aspect of eyes due to variable makeup). Second, we take limitations in storage abilities inherent to humans into account and assume that coarser information (e.g., head silhouette, hairstyle and colour, light/dark pattern of inner features) is prioritised over finer details (e.g., details of the lips) because a coarse-to-fine prioritisation during encoding incurs fewer storage resources (Gao et al., 2013). This flexible and dynamic encoding mechanism creates *cost-effective* memory representations that start off as coarse but refine over time, particularly if appearance changes and/or if demands for recognition out of context increase.

From these two basic assumptions, we hypothesise that the relative stability of changeable aspects (e.g. hairstyle, hair colour, facial hair) affects the *resolution* at which a face is encoded. When a face has a stable appearance, large-scale peripheral features and coarse information are diagnostic and receive substantial representational weight. Moreover, details of inner features need not be encoded in details, yielding low-resolution representations. By contrast, people who change their appearance frequently through variations in hairstyle, hair colour, makeup or facial hair have a more restricted set of large-scale diagnostic features. Therefore, finer aspects that remain stable over time or that are less likely to be occluded by changes in hair, facial hair or make-up (e.g., nose or mouth shape, inner part of the eyes) must receive more representational weight, yielding higher resolution representations. In this framework, recognition errors like a failure of recognition following unexpected changes in appearance in a well-known person or false recognitions of strangers based on gross resemblance with familiar faces are thus viewed as the flipside of an otherwise efficient mechanism.

We tested hypotheses generated by this framework in four recognition experiments using faces of actors. One advantage of using actors is that their faces were learned through a rich variety of viewing conditions, over extended time periods, and without explicit instructions to do so, giving very ecological encoding conditions. Simultaneously, we can operate a strict selection of individual actors based on their physical appearance. Half of the actors had a stable appearance (e.g., Harrison Ford) and half had a variable appearance (e.g., Brad Pitt). Further, to probe the nature of representations observers relied on during recognition, we manipulated the type of information available in test images, and measured how this affected recognition performance. Because peripheral information should have more weight in stable than in variable faces, we predicted that variable faces should be

better recognised than stable faces when peripheral features are occluded (Experiments 1 and 2) or when changeable features, which can be internal or peripheral, deviate from their most common appearance (Experiment 3). We also predicted that stable faces would be better recognised than variable faces from coarse features in blurred images, and that the presence of fine-grained information would benefit more to variable faces than to stable faces (Experiment 4).

Finally, to examine how stability in appearance modulates the evolution of representations over time, we drew on a method developed recently that uses a proxy of exposure to actors in the real world (see Devue et al., 2019). Specifically, we controlled that the amounts of exposure stable and variable actors had received were comparable, based on an objective measure of public visibility available on the Internet Movie Database (IMDb). We compared recognition performance for actors with two levels of popularity (popular and less popular). Based on previous research that suggests a shift of focus from external features for newly encountered faces, towards a conjoint use of internal and external features or favouring the former as a face becomes familiar (Ellis et al., 1979; Young et al., 1985), we hypothesised that representational weights initially set on external features would converge towards inner features over time to become more evenly distributed over the whole face. We thus expected that differences in performance linked to stability should be weaker for popular actors than for less popular ones and that stability would interact with popularity.

General Methods

Participants. Based on power analyses (see **Supplementary Materials**) and to minimise the impact of individual differences in face recognition skills or in exposure to

actors, we recruited a large sample of 100 first year psychology students in Experiment 1. Sample sizes were adapted in subsequent experiments. In all experiments, we excluded participants who did not comply with instructions (i.e., who failed more than 50% of attention checks; see procedure below), and/or who responded too fast (<600 ms or under -2SD from the sample's overall mean reaction time). The study was approved by the School of Psychology Ethics Committee at Victoria University of Wellington.

Materials. ***Actor selection.*** Stability in appearance of prospective actors within given ranges of popularity based on StarMeter ranks (see below) was determined from a visual inspection by authors CD and SD of the pictures on the right-hand side thumbnails returned from a Google web search and in the first five to six rows of Google image searches. Prospective actors were first rated by the authors as displaying low, moderate or high levels of variations based on the appearance of changeable dimensions like hairstyle, hair colour, facial hair, makeup, and accessories (e.g., glasses, hats) across images in the two search results. CD and SD then agreed on a selection based on those ratings while ensuring equivalent sex and age distributions in four different conditions (2 stability x 2 popularity). For the stable condition, we selected 48 actors (24 women, 24 men; *Mean* age = 41.65 years, *SD* = 13.04) whose pictures showed similar appearance on changeable dimensions. For the variable condition, we selected 48 actors (24 women, 24 men; *Mean* age = 40.77 years, *SD* = 10.5) whose pictures markedly varied through various combinations of changes on the same dimensions. Actors' popularity was determined via the StarMeter ranks on IMDb pro, which reflect current popular interest for an actor and their visibility—smaller ranks reflect higher popularity. These ranks were found to predict recognition performance in a recent study (Devue et al., 2019). We selected 48 actors (24 variable, 24 stable) with starMeter ranks between 1 and 500 for the "popular" condition. Importantly, startMeter

ranks of variable actors (Mean = 170.5 ± 104.5, range = 1 - 385) and of stable actors (Mean = 168.5 ± 116, range = 5 - 407) were overall similar[1]. We selected 48 actors (24 variable, 24 stable) with starMeter ranks between 1000 and 1500 for the "less popular" condition, so that the ranks of variable (Mean = 1208.2 ± 162, range = 1006 - 1480) and stable actors (Mean = 1199.6 ± 114, range = 1015 - 1470) were overall similar. Actors and their ranks are listed in **Table S1**.

Unfamiliar faces (48 women and 48 men) were actors with very low popularity on IMDb (i.e., ranks >100,000; *Mean* = 246,309; *SD* = 354,212) from non-English speaking countries and/or who worked in theatre, so that they would not been known by our participants. Their average age (*Mean* age = 39.49 years, *SD* = 11.04) overall matched that of known actors (*Mean* age = 41.21 years, *SD* = 11.8).

Image stimuli. For each of the 96 actors, we selected one image showing their most typical appearance—where the aspect of changeable features shows the most overlap across Google search images (e.g., no facial hair and short grey hair for Harrison Ford; blond short beard and semi long hair for Brad Pitt). For Experiment 3, we also selected 96 atypical pictures. We used the same approach as in Devue et al. (2019) and selected pictures with the most deviations possible from the usual appearance, including hair length, colour, and/or style, presence of facial hair, glasses, and differences in make-up that did not conceal internal features (e.g., goatee and earring for Harrison Ford; dark short hair and moustache for Brad Pitt).

[1]Note that since individual ranks are by definition unique, it is not feasible to pair stable and variable actors based on exact matched ranks. Moreover, actors that follow one another in the ranking do not necessarily display the desirable degree of stability/variability in appearance, gender, or age to achieve a perfect matching.

The set of 288 images (96 typical images of actors, 96 atypical images of actors, and 96 images of strangers) showed faces in a frontal or slightly angled view and with a neutral or happy expression (all evenly distributed across conditions). Images were rotated to align the eyes on a horizontal axis. They were then cropped, so that the hairstyle was apparent while minimising the amount of visible clothing, and resized to 399 by 476 pixels.

We created a "headshot" version of each image, in which the background was concealed with a grey field. For Experiment 1 and 2 we also created a "cropped inner features" version of typical images, where inner features appeared within a truncated ellipse (width = 264, height = 260 pixels), so that bangs and other external features were concealed by a grey field. For Experiment 4, we created a blurred version of typical headshots in which high spatial frequency details were removed by applying a Gaussian filter with a radius of 36 pixels—giving 11 cycles per face width (Goffaux & Rossion, 2006).

Procedure. Participants performed a recognition test online via Testable.org. The 96 pictures of actors and 96 pictures of strangers were presented in a random order at the centre of the screen—until a response was provided or for up to 3 seconds—and participants judged as accurately and fast as possible if they knew the face or not via two response keys (1 and 2). Instructions emphasised familiarity and that there was no need to remember the person's name or identity to judge that a face was familiar. A 1500-ms central fixation cross separated individual trials. Four attention checks—image with instructions to press a specific key (i.e., 5, 6, 7, or 8) instead of the two response keys—and four breaks were dispersed randomly through the trials. Participants performed three practice trials before the test.

Design and measures. Popularity (popular, less popular) and appearance (variable, stable) were manipulated within-subject in all experiments. Image condition (inner features/headshot, typical/atypical, blurred/intact) was manipulated between-subject in experiments 2 and 3a, and within-subject in Experiment 3b and Experiment 4. We calculated *d'* based on hit rate (i.e., correct recognition of actors) in each condition and on false alarm rate (i.e., incorrect recognition of unfamiliar actors) in the corresponding image condition. Descriptive statistics for familiarity judgments and reaction times (means and standard deviations) of all experiments, as well as reaction times analyses for Experiments 2 and 3 appear in **Supplementary materials**.

Transparency and openness. We preregistered the experimental design, analyses and hypotheses for the first series of three experiments with in-built replication on the Open Science Framework before data collection, the document is visible at [https://osf.io/qd5y3/register/564d31db8c5e4a7c9694b2be - 31 July 2018]. Following unexpected results in Experiment 1, analyses plans for Experiments 2 and 3 were amended and preregistered on 10 September 2018 [https://osf.io/h5f6s/register/564d31db8c5e4a7c9694b2c0]. The use of a within-subject design for Experiment 3b was preregistered on 5 December 2018 [https://osf.io/afm4e]. Preregistration for Experiment 4 is available at [https://osf.io/ethbd/ - 6 May 2019]. Image stimuli and datasets for all experiments are available on [https://osf.io/8znw5/files/].

Experiment 1

Methods. Participants were all tested with cropped images of inner features and so familiarity judgments relied exclusively on those features. Of the 100 participants recruited, 96, aged between 18 and 40 years (72 women, 22 men, 2 non-binary; *Mean age* = 19.81

years, *SD* = 3.62), completed the experiment in exchange of course credits. None of them was excluded.

Results and discussion. We conducted a two-way repeated measure Analysis of Variance (ANOVA) with appearance (stable, variable) and popularity (popular, less popular) as within-subject factors on *d'*. As expected, popular actors (*Mean* = 1.36, *SD* = 0.85) were better discriminated from strangers than less popular actors (*Mean* = 0.74, *SD* = 0.58), $F(1,95) = 226.755$, $p < .001$, $\eta_p^2 = .705$. The predicted main effect of appearance (i.e., variable > stable) was not significant, $F(1,95) = .342$, $p = .56$, $\eta_p^2 = .004$, because of a crossed interaction with popularity, $F(1,95) = 132.184$, $p < .001$, $\eta_p^2 = .582$, see **Figure 1**.

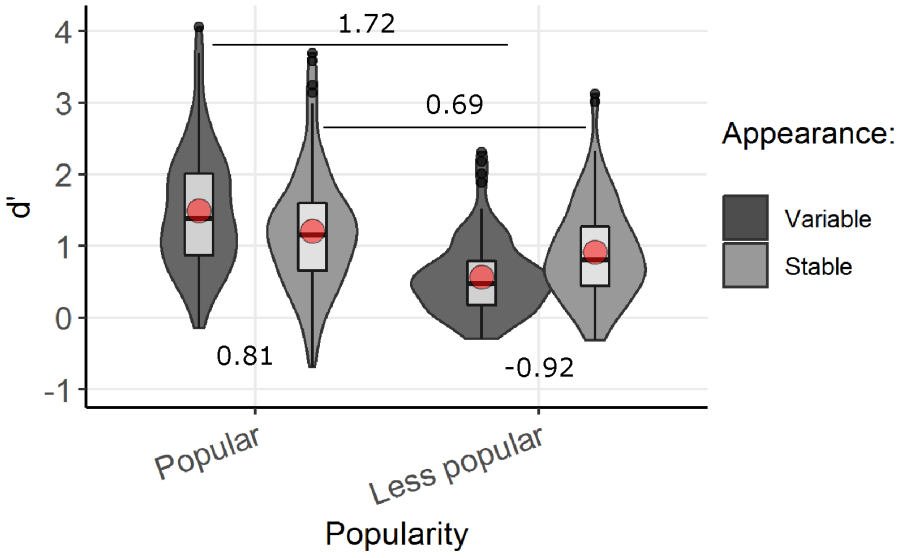

Figure 1. Results of Experiment 1. Discrimination performance (d') for images of cropped inner features as a function of popularity and appearance. Red circles show the mean, and boxplots show distribution in quartiles. Violin size is proportional to the distribution of performance in each condition. Values on the plot are effect sizes (Cohen's *d*) for paired-comparisons.

Benefit of exposure. We followed up the interaction with one-tailed paired sample Student t-tests. As expected, sensitivity improved with increased exposure in both variable,

$t(95) = 16.825$, $p_{one\text{-}tailed} < .001$, $d = 1.717$ (95% C.I.$_{two\text{-}tailed} = 1.4 – 2.031$), and stable actors, $t(95) = 6.775$, $p_{one\text{-}tailed} < .001$, $d = 0.691$ (95% C.I.$_{two\text{-}tailed} = 0.467 – 0.913$). Cohen's *d* values and the lack of overlap between their respective confidence intervals[2] suggest that benefits of increased exposure were significantly larger for variable than for stable faces.

Impact of appearance. As predicted, for popular actors, inner features of variable faces were better recognised than those of stable faces, $t(95) = 7.958$, $p_{one\text{-}tailed} < .001$, $d = 0.812$ (95% C.I. = 0.58 – 1.041). Unexpectedly, for less popular actors, sensitivity to inner features was higher for stable faces than for variable faces, and so the one-tailed test based on our expectation of the opposite pattern was not significant, $t(95) = -9.029$, $p_{one\text{-}tailed} = 1$, $d = 0.921$ (95% C.I. = 0.681| – |1.159). This advantage for stable faces contrasts with findings in face learning studies where exposure to increased levels of variability leads to immediate increases in recognition rates compared to less variable viewing conditions (Baker et al., 2017; Ritchie & Burton, 2017). We assume that this difference is due to unsupervised learning conditions in which actors' faces are often learned. Before actors get leading roles and become famous, we might see them in different support roles without the explicit knowledge that they are the same person. Our results suggest that stability helps "put faces together" during these early stages of familiarisation and that we are more likely to recognise someone who had similar appearances in different movies than someone who has changed. In turn, stability may allow us to consolidate the representation of newly learned faces and of their inner features.

[2] We present two-tailed confidence intervals for comparison purposes as the one-tailed version's upper limit is infinite.

Although the interaction between appearance and popularity did not take the anticipated shape—where a disadvantage of stable faces compared to variable faces would decrease over time if representations of all faces converged towards the same levels of refinements—the results of this experiment remain consistent with a cost-efficient face encoding mechanism. Faces that vary more are ultimately better recognised from inner features than faces that are more stable, suggesting that these features are represented in a more reliable manner—at a higher resolution. The larger improvement in recognition performance that variable faces display over time compared to stable faces suggests that representations of the former become more fine-tuned than representations of stable faces, which tend to remain coarser.

Experiment 2

This experiment replicates and expands on Experiment 1. We compared recognition from images of inner features and headshots where external features are visible. We expected that in popular actors, the presence of coarse external features would reduce the disadvantage of stable faces and compensate for lower resolution representations of inner details.

Methods. Because of the unexpected pattern with less popular actors in Experiment 1, we pre-registered an amended analysis plan before data collection [https://osf.io/h5f6s/register/564d31db8c5e4a7c9694b2c0 – 10 Sept 2018]. The design and variables remain identical to those described in the original pre-registration.

We used sequential analyses (Lakens, 2014)—details are presented in **Supplementary Materials**—and recruited a total of 123 participants, 3 of whom replaced participants who did not follow instructions (N = 2) or responded too fast (N = 1). Participants completed an

online recognition task either in the "inner features" condition (39 women, 18 men, 3 non-binary; *Mean age* = 18.9 years, *SD* = 1.32) or in the "headshot" condition (41 women, 19 men; *Mean age* = 19.15 years, *SD* = 1.87).

Results and discussion. The critical *p* value for our sequential analyses was set at .0182. We present uncorrected *p* values and so an effect must be interpreted as significant at $p < .0182$. We conducted a three-way mixed effect ANOVA with appearance (variable, stable) and popularity (popular, less popular) as within-subject factors, and image condition (inner features, headshot) as between-subject factor on *d'*. We found the expected main effect of image condition, $F(1,118) = 73.97$, $p < .001$, $\eta_p^2 = .385$, as sensitivity was higher with headshots (*Mean* = 2.023, *SD* = 0.59) than with images of inner features (*Mean* = 1.121, *SD* = 0.556). The three-way interaction between appearance, popularity, and image condition was significant[3], $F(1,118) = 9.875$, $p = .002$, $\eta_p^2 = .077$, see **Figure 2**. We then examined performance separately in each image condition and tested whether we replicated findings of Experiment 1 in the inner features condition.

Inner features. As in Experiment 1, a two-way repeated measure ANOVA showed a main effect of popularity, $F(1,59) = 216.173$, $p < .001$, $\eta_p^2 = .786$, qualified by an interaction with appearance, $F(1,59) = 43.161$, $p < .001$, $\eta_p^2 = .422$. The main effect of appearance was not significant, $F(1,59) = 1.539$, $p = .22$, $\eta_p^2 = .025$.

Follow-up t-tests showed that sensitivity to inner features increased with popularity for both variable, $t(59) = 12.31$, $p < .001$, $d = 1.589$ (95% C.I. = 1.204 – 1.967), and stable actors, $t(59) = 8.136$, $p < .001$, $d = 1.05$ (95% C.I. = 0.732 – 1.363). Effects sizes suggest

[3] Results of the same ANOVA conducted at step 1 and step 2 of sequential analyses followed a similar pattern and are visible in Supplementary Materials.

numerically larger improvements from increased exposure for variable than for stable actors but Cohen's d confidence intervals overlap and so the size of the improvement is not significantly different. In popular actors, variable faces were better recognised than stable ones, $t(59) = 4.125$, $p < .001$, $d = 0.533$ (95% C.I. = 0.26 – 0.801), while in less popular actors, stability facilitated recognition compared to variability, $t(59) = -5.967$, $p < .001$, $d = 0.77$ (95% C.I. = 0.479 - 1.056).

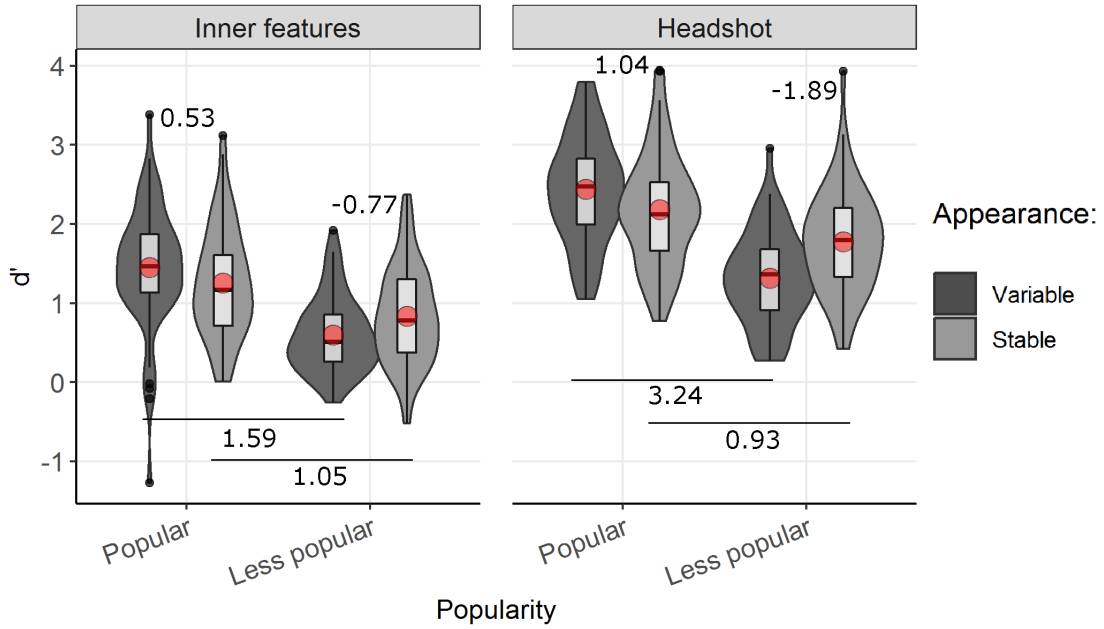

Figure 2. Results of Experiment 2. Discrimination performance (d') as a function of popularity, appearance of actors and image type(inner features vs. headshot). Red circles show the mean, and boxplots show distribution in quartiles. Violin size is proportional to the distribution of performance in each condition. Values on the plot are Cohen's d for paired-comparisons.

Headshots. The same ANOVA in the headshot condition yielded a roughly similar pattern, except that the main effect of appearance was significant, $F(1,59) = 21.59$, $p < .001$, $\eta_p^2 = .268$. Overall, headshots of stable faces (*Mean* = 2.14, *SD* = 0.65) were better discriminated from strangers than headshots of variable faces (*Mean* = 2, *SD* = 0.615). Here

too a significant main effect of popularity, $F(1,59) = 433.31$, $p < .001$, $\eta_p^2 = .88$, was qualified by an interaction with appearance, $F(1,59) = 269.56$, $p < .001$, $\eta_p^2 = .82$.

Follow-up analyses showed that sensitivity improved with popularity for both variable, $t(59) = 25.066$, $p < .001$, $d = 3.236$ (95% C.I. = 2.598 – 3.868), and stable faces, $t(59) = 7.186$, $p < .001$, $d = 0.928$ (95% C.I. = 0.622 – 1.228). Cohen's d values and their respective confidence intervals indicate that benefits of increased exposure were much stronger for variable faces than for stable faces. Amongst popular actors, variable faces were better recognised than stable ones, $t(59) = 8.05$, $p < .001$, $d = 1.039$ (95% C.I. = 0.722 – 1.351), whereas in less popular actors, stability improved recognition compared to variability, $t(59) = -14.665$, $p < .001$, $d = 1.893$ (95% C.I. = 1.466 - 2.315). The significantly larger advantage of stable faces over variable faces in less popular actors relative to the advantage of variable faces over stable faces in popular actors must be driving the overall advantage of stable faces over variable faces shown in the main effect of appearance above.

Gain from peripheral information. **Figure 5** (panel A) presented below illustrates gains in sensitivity from images of inner features to full headshots. We examined the gains provided by peripheral features in each actor category with four independent sample t-tests. We hypothesised that external information is more diagnostic in stable faces than in variable faces and so we expected larger gains—reflected by larger Cohen's d—for stable faces than for variable faces. Peripheral features helped recognition in all the conditions and Cohen's d values were numerically larger for less popular stable faces, $t(118) = 8.853$, $p < .001$, $d = 1.616$ (95% C.I. = 1.201 – 2.027), than in the three other categories, in which gains were all in the same ballpark: popular variable, $t(118) = 7.735$, $p < .001$, $d = 1.412$ (95% C.I. = 1.009 – 1.81); popular stable, $t(118) = 7.477$, $p < .001$, $d = 1.365$ (95% C.I. = 0.965 – 1.761);

less popular variable, $t(118) = 7.772$, $p < .001$, $d = 1.419$ (95% C.I. = 1.016 – 1.817). However, the overlap of the four Cohen's d confidence intervals suggest that the numerical difference was not significant.

Consistent gains from inner features to headshots suggests that the presence of peripheral information supports recognition of both variable and stable faces, probably because it is part of a holistic representation (Andrews et al., 2010; Tanaka & Simonyi, 2016). Nevertheless, **Figure 5** (panel A) suggests that proportionally, peripheral information seems particularly useful to both variable and stable faces in earlier stages of familiarisation. During that same stage, stability in appearance improves recognition and facilitates familiarisation compared to variability in appearance. Over time however, further exposure to a stable appearance seems less effective in increasing the reliability of representations than when appearance has varied more. In other words, although increased variations in appearance initially slow down familiarisation, they eventually lead to more robust representations.

Experiment 3

Here we compared recognition of typical and atypical headshots, following the same design as Experiment 2, except that atypical headshots replaced images of inner features. We expected that once an actor is popular, atypical changes in appearance should be less disruptive for variable than for stable faces because recognition could be based on fine-tuned representation of invariable features.

Methods. *Experiment 3a.* We tested 59 first year psychology students and 67 additional New Zealanders recruited via social media or amongst colleagues. We aimed to have at least 60 participants per group like in Experiment 2. We excluded two participants

who failed more than two attention checks and one participant whose accuracy was below 50%. The final combined sample consisted of 123 participants (78 women, 45 men) aged between 18 and 55 (*Mean* = 22.93 years, *SD* = 6.8). There were 62 participants in the typical condition (35 women; *Mean* = 22.34 years, *SD* = 6.4) and 61 in the atypical condition (43 women; *Mean* = 23.52 years, *SD* = 7.2). We did not find the expected advantage of typicality, which could have been due to individual differences in exposure to actors or in face recognition skills between groups. To address this possibility, we ran an additional experiment where image condition was manipulated within-subject.

Experiment 3b. Here we aimed to collect data from 80 participants and tested 89 Mechanical Turk workers located in the US. We excluded 8 participants who failed attention checks, responded too fast, and/or whose accuracy was below 50%. The final sample consisted of 81 participants (35 women, 45 men, 1 non-binary) aged between 18 and 67 (*Mean* = 37.19 years, *SD* = 10.62). As this sample had different demographics than those in the other experiments, it also provided an opportunity to test the generalisability of our findings.

We presented typical and atypical headshots of the 96 actors to the same participants in a random order. Images of the 96 strangers were presented twice to maintain the ratio of trials with actors and strangers, giving a total of 348 trials. Eight breaks and four attention checks were dispersed throughout. The instructions specified that familiarity judgments concerned pre-experimental familiarity, and that any person that appeared multiple times but was unknown prior the experiment should still be judged unfamiliar.

Results. ***Experiment 3a.*** We conducted a three-way mixed effect ANOVA with appearance (variable, stable) and popularity (popular, less popular) as within-subject

factors, and image condition (typical, atypical headshot) as between-subject factor on *d'*. Although performance was numerically lower with atypical headshots (*Mean* = 1.74, *SD* = 0.7) than with typical ones (*Mean* = 1.91, *SD* = 0.09), we did not find the expected typicality effect, $F(1,121) = 1.595$, $p = .21$, $\eta_p^2 = .013$. There was a main effect of popularity, $F(1,121) = 813.399$, $p < .001$, $\eta_p^2 = .871$, and of appearance, $F(1,121) = 22.705$, $p < .001$, $\eta_p^2 = .158$, with stable faces (*Mean* = 1.92, *SD* = 0.83) overall being better recognised than variable ones (*Mean* = 1.82, *SD* = 0.95). The three-way interaction between appearance, popularity, and image condition was not significant, $F(1,121) = 1.491$, $p = .224$, $\eta_p^2 = .012$. Nevertheless, **Figure 3** (top panel) shows a similar pattern in each image type as in Experiment 2. For the sake of space, we do not report follow-up analyses and move on to Experiment 3b.

Experiment 3b. Using a fully within-subject design, we found the expected main effect of image type, $F(1,80) = 210.898$, $p < .001$, $\eta_p^2 = .725$. Typical images (*Mean* = 1.62, *SD* = 0.86) were now significantly better discriminated from strangers than atypical images (*Mean* = 1.37, *SD* = 0.86). There was a main effect of popularity, $F(1,80) = 203.147$, $p < .001$, $\eta_p^2 = .717$, and of appearance, $F(1,80) = 8.673$, $p = .004$, $\eta_p^2 = .098$, with an overall advantage for stable faces (*Mean* = 1.52, *SD* = 0.81) compared to variable ones (*Mean* = 1.47, *SD* = 0.92). The three-way interaction between image type, popularity and appearance was significant, $F(1,80) = 7.652$, $p = .007$, $\eta_p^2 = .087$, see **Figure 3** (bottom panel).

Atypical headshots. A follow-up 2-way ANOVA on atypical headshots showed a main effect of popularity, $F(1,80) = 115.579$, $p < .001$, $\eta_p^2 = .591$, and of appearance, $F(1,80) = 5.092$, $p = .027$, $\eta_p^2 = .06$, and an interaction between the two, $F(1,80) = 88.601$, $p < .001$, $\eta_p^2 = .526$. Paired comparisons showed that sensitivity improved with increased popularity for all actors, with a significantly larger improvement for variable faces, $t(80) = 14.563$, $p < .001$,

d = 1.618 (95% C.I. = 1.284 – 1.948), than for stable ones, *t*(80) = 2.244, *p* = .028, *d* = 0.249 (95% C.I. = 0.027 – 0.47). In popular actors, variable faces were recognised better than stable ones, *t*(80) = 6.181, *p* < .001, *d* = 0.687 (95% C.I. = 0.443 – 0.927). In less popular actors, stable faces were better recognised than variable ones, *t*(80) = -8.342, *p* < .001, *d* = 0.927 (95% C.I. = 0.664| –|1.186).

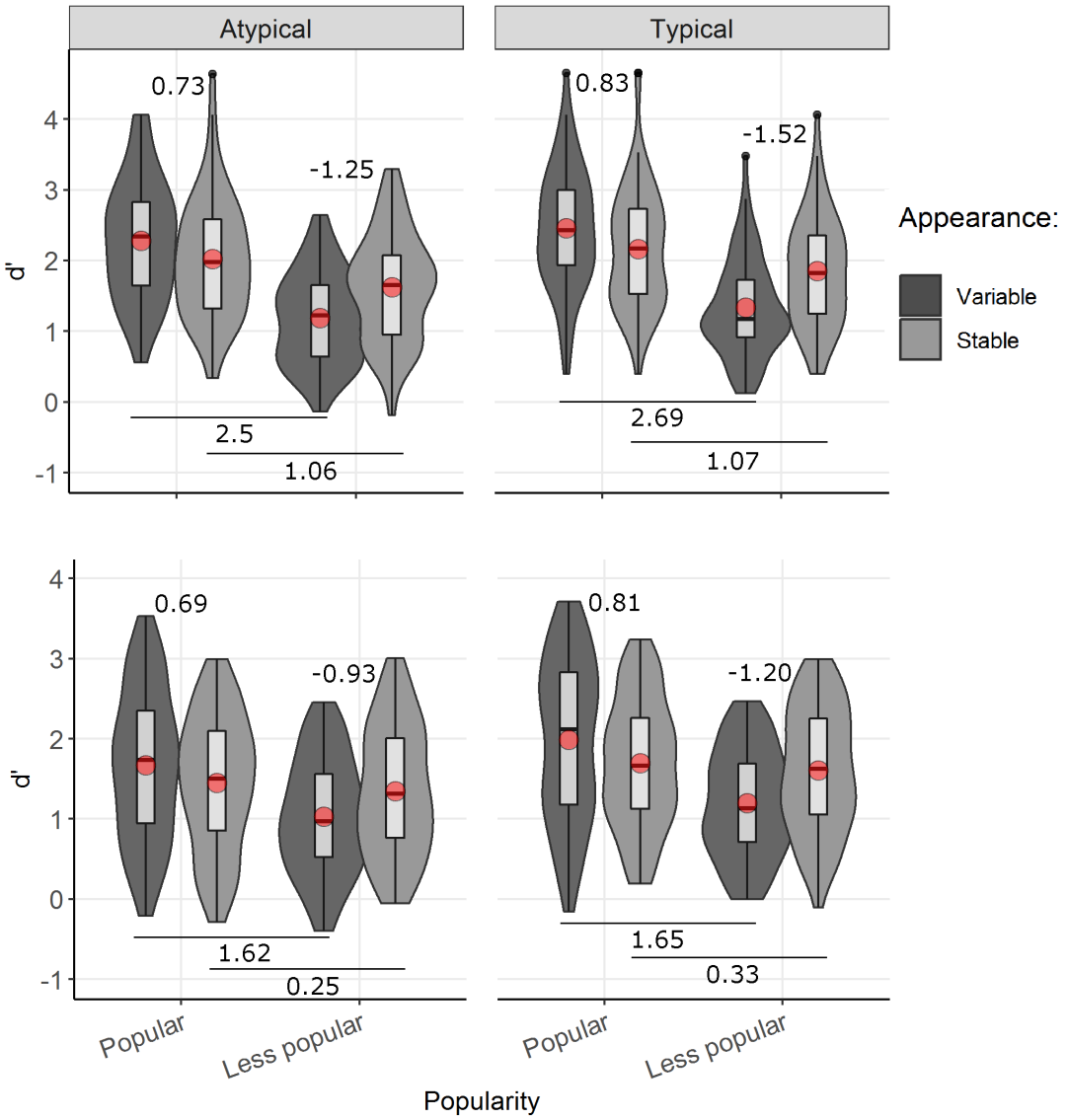

Figure 3. Results of Experiment 3a (top) and 3b (bottom). Discrimination performance (d') as a function of popularity and appearance, for typical and atypical images of actors. Red circles show the mean, and boxplots show distribution in quartiles. Violin size is proportional to the distribution of performance in each condition. Values on the plot are Cohen's d from paired-comparisons.

Typical headshots. The same 2-way ANOVA on typical headshots also showed a main effect of popularity, $F(1,80) = 188.779$, $p < .001$, $\eta_p^2 = .702$, and of appearance, $F(1,80) = 4.284$, $p = .042$, $\eta_p^2 = .051$, and an interaction between the two, $F(1,80) = 144.981$, $p < .001$, $\eta_p^2 = .644$, replicating results with headshots in Experiment 2. Like in Experiment 2, performance improved with popularity for stable faces, $t(80) = 2.95$, $p = .004$, $d = 0.328$ (95% C.I. =0.103 – 0.55), and improved significantly more for variable faces, $t(80) = 14.8513$, $p < .001$, $d = 1.65$ (95% C.I. = 1.312 – 1.983). In popular actors, variable faces were better recognised than stable ones, $t(80) = 7.26$, $p < .001$, $d = 0.807$ (95% C.I. = 0.554 – 1.056). In less popular actors, stable faces were better recognised than variable ones, $t(80) = -10.773$, $p < .001$, $d = 1.197$ (95% C.I. = 0.909| - |1.481).

Gain from typicality. We examined the gain in performance from typicality (i.e., typical vs. atypical) in each actor category with paired sample t-tests, see **Figure 5** (panel B). Typical facial information improved performance in all actor categories: Popular variable, $t(80) = 8.552$, $p < .001$, $d = 0.95$ (95% C.I. = 0.685 - 1.211); popular stable, $t(80) = 7.955$, $p < .001$, $d = 0.884$ (95% C.I. = 0.625 - 1.139); less popular variable, $t(80) = 5.398$, $p < .001$, $d = 0.6$ (95% C.I. = 0.362 - 0.835); and less popular stable, $t(80) = 7.705$, $p < .001$, $d = 0.856$ (95% C.I. = 0.599 - 1.109). Effects sizes and the fact that they overlap indicate that gains from typicality were comparable in all actor categories. This may suggest that although representations of variable actors refine over time and more so than those of stable actors, they also tend to incorporate more typical aspects of elements that vary (e.g., most frequent hairstyle or makeup) into a holistic representation. Indeed, if recognition of variable faces only relied on invariable elements, recognition would have been equally good from typical and atypical images.

Experiment 4

Finally, in this experiment, we more directly tested the hypothesis that representations of variable faces incorporate finer grain information than representations of stable faces. We compared recognition performance from intact images (typical headshots) and from blurred images which only contain coarse information and where facial details are removed. Because it was unclear whether running this experiment online would allow sufficient control over image presentation and viewing conditions, we first ran it in a controlled laboratory environment (Experiment 4a) and then ran an online version (Experiment 4b).

Methods. The experiment and analyses plans were pre-registered before data collection [https://osf.io/ethbd/][4].

In Experiment 4a, we recruited 81 psychology students (12 men, 69 women) aged between 18 and 21 (*Mean* = 18.39 years, *SD* = 0.72). None of them was excluded. In Experiment 4b, we recruited 102 psychology students (72 women, 29 men, 1 non-binary) aged between 18 and 34 years (*Mean* = 18.97 years, *SD* = 2.1) who completed the task online. We excluded one participant who responded too fast and two who responded "unfamiliar" to all the blurred images[5], giving a final sample of 99 participants (71 women, 27 men, 1 non-binary; *Mean age* = 18.94 years, *SD* = 2.11).

[4]This experiment was initially part of a separate project and pre-planned analyses were structured slightly differently—i.e., we were planning to follow up on a three-way interaction with two separate 2-way repeated measure ANOVA on each popularity level with image type and appearance as within-subject factors. We decided to incorporate this experiment to this project and adopted the same analysis plan as in the other three. The paired-comparisons presented here tackle the same questions as originally planned.

[5]We did not anticipate that possibility in our pre-registration but decided that this responding strategy did not comply with our instructions.

To accommodate for changes in popularity of actors over time (i.e, 9 months since Experiment 1 and 5 months since Experiment 3), we updated actors' IMDb ranks and selected a subset of identities (i.e., 18 identities per actor category instead of 24, and 72 strangers instead of 96) that could still fit within two distinct popularity categories (i.e., popular: range 26 to 475; less popular: range 594 to 3230) while keeping ranks and age similar between variable and stable actors. The updated list of actors is visible in Supplementary Materials (**Table S2**).

The procedure was identical to that in Experiment 3b with the following exceptions. There were 144 trials. Each trial started with the blurred image of a person, followed by the intact image of the same person. Participants responded to both images and were told that they could give a different response to each[6]. Each image in the pair was presented until a response was provided or for maximum 3 seconds, and they were separated by a 1-second interval. An inter-trial interval of 2 seconds with a central fixation cross separated trials with different identities. Finally, in Experiment 4a, participants were run in groups of maximum 4 in separate booths and viewing distance (57.3 cm) was controlled by means of a chin rest. Stimuli had a visual angle of 7.9 x 9.4 degrees.

Results. ***Experiment 4a.*** We found the expected main effect of image type, $F(1,80) = 219.342$, $p < .001$, $\eta_p^2 = .733$, as intact images (*Mean* = 2.07, *SD* = 0.85) were better discriminated than blurred images (*Mean* = 1.26, *SD* = 0.63). There was a main effect of

[6]Note that as in the other experiments, we are interested in the overall comparison between image conditions (blurred vs. intact) and so we do not analyse data contingent on changes in response from blurred to intact images. For familiar actors, changes from "familiar" responses to blurred images to "unfamiliar" responses to intact images represented only 4.4% of trials in Experiment 4a, and 4% of trials in Experiment 4b. Keep in mind that in any recognition task, there is always a possibility that a hit does not reflect a genuine familiarity with the face, whether participants judge two images and change their mind as in this experiment, whether they judge two different images of the same actor presented in a random order as in Experiment 3b, or whether they judge only one image as in a typical recognition experiment. This possibility is balanced out by similar response tendencies to strangers and accounted for in the measure of sensitivity.

popularity, $F(1,80) = 464.912$, $p < .001$, $\eta_p^2 = .853$, and of appearance, $F(1,80) = 8.592$, $p < .001$, $\eta_p^2 = .097$, with an overall advantage for stable faces (*Mean* = 1.7, *SD* = 0.77) compared to variable ones (*Mean* = 1.63, *SD* = 0.93). The three-way interaction between image type, popularity and appearance was significant, $F(1,80) = 16.731$, $p < .001$, $\eta_p^2 = .173$, see **Figure 4** (top panel).

Blurred images. A follow-up 2-way ANOVA on blurred images showed a main effect of popularity, $F(1,80) = 203.937$, $p < .001$, $\eta_p^2 = .718$, qualified by an interaction with appearance, $F(1,80) = 27.44$, $p < .001$, $\eta_p^2 = .255$. Contrary to our expectation (i.e., blurred stable faces > blurred variable faces), there was no significant main effect of appearance, $F(1,80) = 2.062$, $p = .155$, $\eta_p^2 = .025$. Sensitivity improved with popularity for all faces, with comparable improvements for variable faces, $t(80) = 12.6$, $p < .001$, $d = 1.4$ (95% C.I. = 1.09 – 1.705), and for stable ones, $t(80) = 8.09$, $p < .001$, $d = 0.899$ (95% C.I. = 0.639 – 1.155). Contrary to our expectation, in popular actors, variable faces were still better recognised than stable ones from blurred images, $t(80) = 2.77$, $p = 0.007$, $d = 0.308$ (95% C.I. = 0.084 – 0.53), showing that representations of both types of faces somewhat rely on coarse information. However, this effect of appearance was the smallest across all experiments. In less popular actors, stable faces were better recognised than variable ones, $t(80) = -5.77$, $p < .001$, $d = 0.641$ (95% C.I. = 0.4 - 0.879).

Intact images. We found a similar pattern with intact headshots as with headshots in Experiment 2 and 3. There was a main effect of popularity, $F(1,80) = 497.97$, $p < .001$, $\eta_p^2 = .862$, and of appearance, $F(1,80) = 10.28$, $p < .001$, $\eta_p^2 = .114$, as well as an interaction between the two, $F(1,80) = 129.95$, $p < .001$, $\eta_p^2 = .619$. Again, sensitivity improved with popularity, and significantly more so for variable faces, $t(80) = 21.194$, $p < .001$, $d = 2.355$

(95% C.I. =1.928 – 2.777), than for stable faces, $t(80) = 12.342$, $p < .001$, $d = 1.371$ (95% C.I. = 1.065 – 1.673). In popular actors, variable faces were better recognised than stables faces, $t(80) = 5.674$, $p < .001$, $d = 0.63$ (95% C.I. = 0.39 – 0.867), while the opposite was true in less popular actors, $t(80) = -12.862$, $p < .001$, $d = 1.429$ (95% C.I. = 1.116 - 1.737).

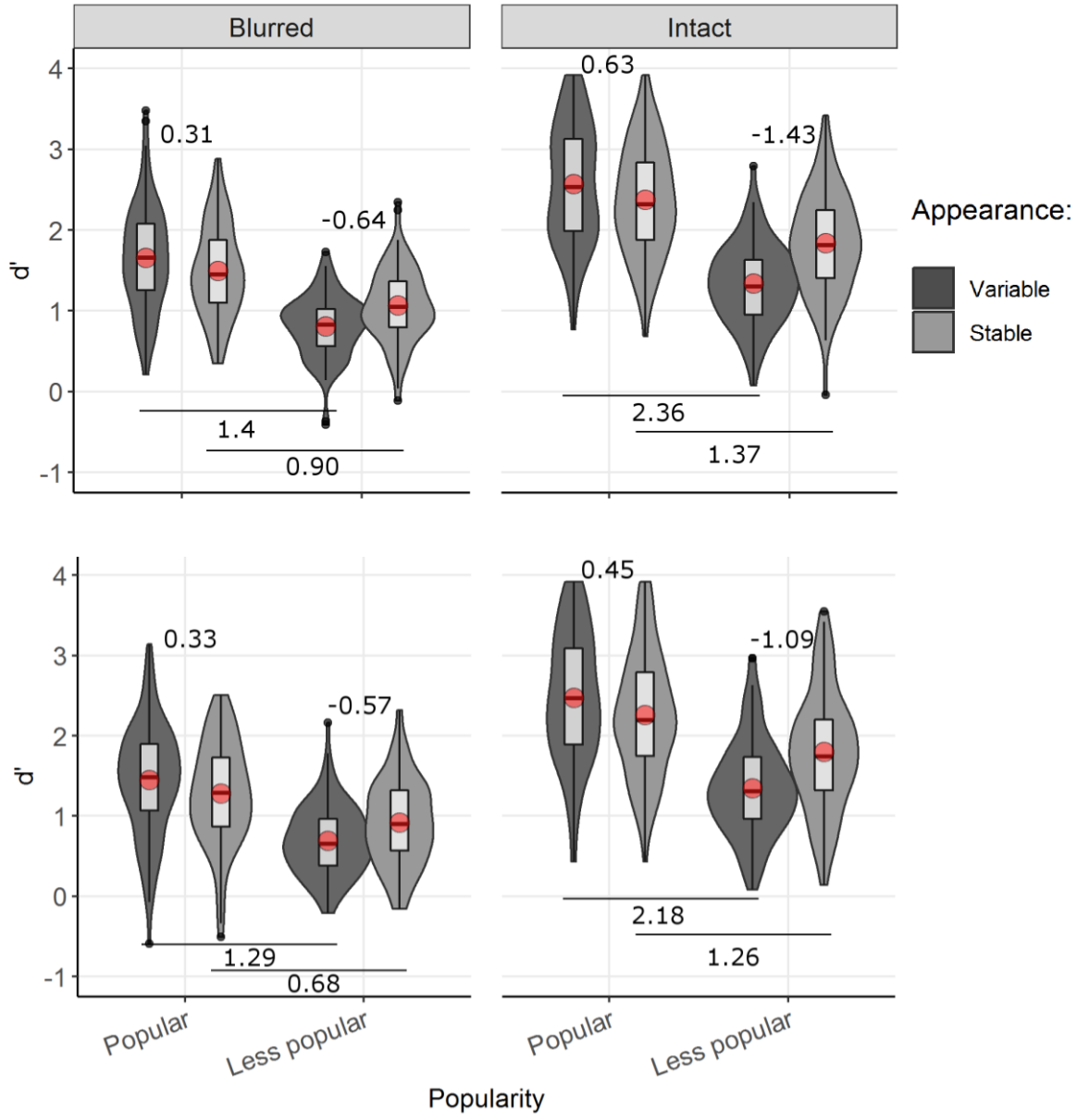

Figure 4. Results of Experiment 4a (lab-based, top) and 4b (online, bottom). Discrimination performance (d') as a function of popularity and appearance, for blurred and intact images of actors. Red circles show the mean, and boxplots show distribution in quartiles. Violin size is proportional to the distribution of performance in each condition. Values on the plot are Cohen's d from paired-comparisons.

Gain from fine-grain information. We examined improvement in recognition from the inclusion of fine details in intact images compared to blurred images in each actor category, see **Figure 5** (panel C). The addition of details consistently improved recognition performance compared to coarse information alone: Popular variable, $t(80) = 13.669$, $p < .001$, $d = 1.519$ (95% C.I. = 1.196 – 1.837); popular stable, $t(80) = 12.568$, $p < .001$, $d = 1.396$ (95% C.I. = 1.087 – 1.701); less popular variable, $t(80) = 9.46$, $p < .001$, $d = 1.051$ (95% C.I. = 0.777 – 1.321); and less popular stable, $t(80) = 12.451$, $p < .001$, $d = 1.383$ (95% C.I. = 1.076 – 1.686).

Experiment 4b. We found the expected main effect of image type, $F(1,98) = 350.835$, $p < .001$, $\eta_p^2 = .782$, as intact images (*Mean* = 2.02, *SD* = 0.92) were better recognised than blurred images (*Mean* = 1.09, *SD* = 0.65). There was a main effect of popularity, $F(1,98) = 389.54$, $p < .001$, $\eta_p^2 = .799$, and of appearance, $F(1,98) = 10.416$, $p = .002$, $\eta_p^2 = .096$, with an overall advantage for stable faces (*Mean* = 1.59, *SD* = 0.9) compared to variable ones (*Mean* = 1.52, *SD* = 0.96). The three-way interaction between image type, popularity and appearance was significant, $F(1,98) = 12.648$, $p < .001$, $\eta_p^2 = .114$, see **Figure 4** (bottom panel).

Blurred images. We replicated the findings of Experiment 4a and found a main effect of popularity, $F(1,98) = 149.069$, $p < .001$, $\eta_p^2 = .603$, qualified by an interaction with appearance, $F(1,98) = 34.671$, $p < .001$, $\eta_p^2 = .261$. There was no significant main effect of appearance, $F(1,98) = 0.797$, $p = .374$, $\eta_p^2 = .008$. Sensitivity improved with popularity for all faces but with a significantly larger improvement for variable faces, $t(98) = 12.858$, $p < .001$, $d = 1.292$ (95% C.I. = 1.023 – 1.558), than for stable ones, $t(98) = 6.736$, $p < .001$, $d = 0.677$ (95% C.I. = 0.457 – 0.894). Again, in popular actors, variable faces were still better

recognised than stable ones, $t(98) = 3.248$, $p = 0.002$, $d = 0.326$ (95% C.I. = 0.123 – 0.528), but the simple effect of appearance was also relatively small compared to that obtained with different types of images across experiments. In less popular actors, stable faces were better recognised than variable ones, $t(98) = -5.652$, $p < .001$, $d = 0.568$ (95% C.I. = 0.354 - 0.779).

Intact images. We again replicated the pattern found with intact headshots in all the other experiments. There was a main effect of popularity, $F(1,98) = 506.77$, $p < .001$, $\eta_p^2 = .838$, and of appearance, $F(1,98) = 18.23$, $p < .001$, $\eta_p^2 = .157$, and an interaction between the two, $F(1,98) = 104.15$, $p < .001$, $\eta_p^2 = .515$. As per usual now, recognition performance increased with popularity, and significantly more so for variable faces, $t(98) = 21.644$, $p < .001$, $d = 2.175$ (95% C.I. =1.811 – 2.536), than for stable faces, $t(98) = 12.534$, $p < .001$, $d = 1.26$ (95% C.I. = 0.994 – 1.522). In popular actors, variable faces were better recognised than stables faces, $t(98) = 4.456$, $p < .001$, $d = 0.448$ (95% C.I. =0.24 – 0.653), while the opposite was true in less popular actors, $t(98) = -10.82$, $p < .001$, $d = 1.087$ (95% C.I. = 0.837 - 1.335).

Gain from fine-grain information. Increases in sensitivity due to the addition of details compared to coarse information alone are illustrated on **Figure 5** (panel D). Details consistently improved recognition performance and in comparable ways in all actor categories: Popular variable, $t(98) = 17.233$, $p < .001$, $d = 1.732$ (95% C.I. = 1.418 – 2.042); popular stable, $t(98) = 15.993$, $p < .001$, $d = 1.607$ (95% C.I. = 1.307 – 1.904); less popular variable, $t(98) = 12.276$, $p < .001$, $d = 1.234$ (95% C.I. = 0.97 – 1.494); and less popular stable, $t(98) = 15.746$, $p < .001$, $d = 1.583$ (95% C.I. = 1.284 – 1.877).

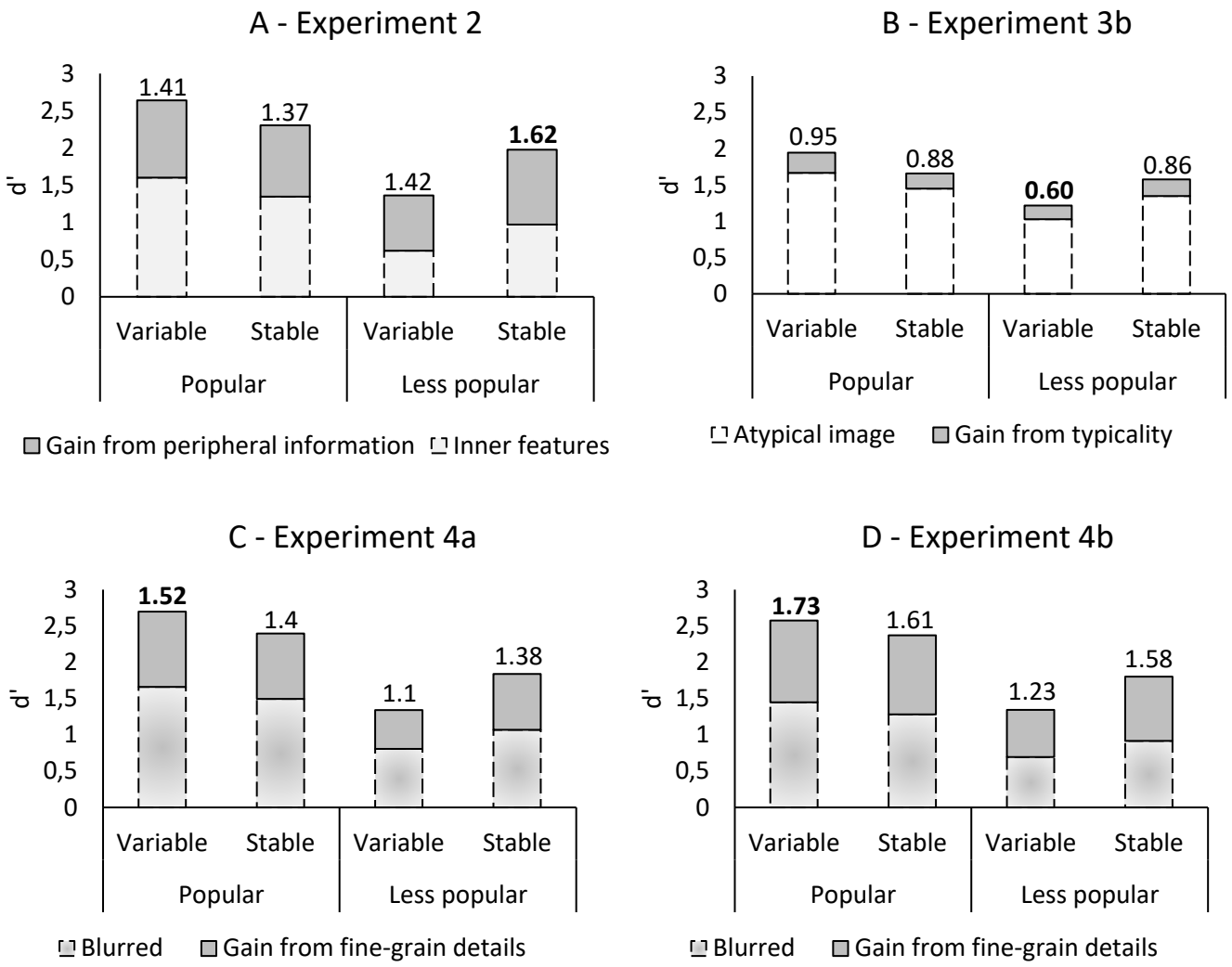

Figure 5. Comparisons of sensitivity to actors' faces in different image conditions, as a function of popularity and appearance. The top of the grey bar represents sensitivity to typical headshots, which was compared to recognition from inner features in Experiment 2 (N = 123 New Zealanders), from atypical headshots in Experiment 3b (N = 81 US Mechanical Turk workers), and from blurred headshots in Experiment 4. Experiment 4 was administered in a controlled laboratory environment (4a, N = 81 NZ students) and online (4b, N = 99 NZ students). Values in the plot area represent Cohen's d for paired-comparisons of performance in different image conditions.

Lab-based vs. online administration. It is worth noting that the patterns of results are remarkably similar in the lab-based and in the online versions of the experiment. An exploratory four-way mixed effect ANOVA conducted on the combined samples with

administration (lab-based, online) as a between-subject factor showed no significant main effect of administration, $F(1,178) = 1.903$, $p = .170$, $\eta_p^2 = .011$, and no significant interaction between administration and any of the other factors (image type, popularity, appearance). Past research had shown that low-pass filtering is not affected by viewing distance (Hayes et al., 1986) and our findings thus confirm that uncontrolled viewing distances during online testing with blurred images poses no notable issue. More generally, the fact that the lab-based experiment yields similar patterns of results as the other three experiments, despite substantial changes in the procedure and image set, also speaks to the validity of all the data collected online.

Exploration of the impact of actors' sex

In addition to initial pre-planned analyses, we explored whether the sex of the actors had an impact on discrimination performance (d') by means of 3-ways repeated measures ANOVA with popularity, appearance and sex as within-subject factors for each type of image in each experiment. For the sake of space, we report descriptive statistics and results of the three-way interactions in **Table 1**. Additional figures showing performance in each image condition are presented in Supplementary materials (Figures S1 to S11). In all experiments using intact images but one (i.e. Experiment 4b), we found a similar pattern of performance. In less popular actors, stable faces were better recognised than variable faces for both female and male actors. However, differences were observed between female and male faces for popular actors with a stable appearance. Whereas discrimination performance for stable women increased with exposure, discrimination of stable male faces did not improve with exposure or when it did, it did not as much as women's.

Similar patterns of interaction were also observed with cropped images of inner features and with blurred images, but not with atypical images.

Table 1. Mean discrimination performance (d') and standard deviations (in italics) as a function of sex, popularity and appearance in different image conditions of all experiments. Results of the associated three-way interactions appear in the three rightmost columns.

			Female								**Male**								**Popularity ✻ Appearance ✻ Sexe**		
			Popular				**Less popular**				**Popular**				**Less popular**						
Test image	**Exp.**	**N**	**Variable**		**Stable**		**Variable**		**Stable**		**Variable**		**Stable**		**Variable**		**Stable**		**F(1,N-1)**	**p**	**η^2_p**
Headshot	2	60	2.5	*0.9*	2.5	*1.0*	1.1	*0.7*	1.7	*0.7*	2.9	*0.7*	2.3	*0.6*	1.5	*0.7*	2.3	*0.6*	33.418	< .001	0.362
	3a	62	2.2	*0.9*	2.2	*1.0*	1.1	*0.8*	1.5	*0.9*	2.9	*1.0*	2.2	*0.9*	1.6	*0.8*	2.2	*0.8*	48.900	< .001	0.445
	3b	81†	1.7	*0.9*	1.6	*0.9*	0.9	*0.7*	1.4	*0.7*	2.5	*1.3*	1.8	*0.8*	1.4	*0.8*	1.9	*1.0*	14.810	< .001	0.156
	4a	81†	2.7	*1.0*	2.5	*0.8*	1.2	*0.5*	1.5	*0.8*	2.8	*0.9*	2.3	*0.8*	1.4	*0.7*	2.1	*0.6*	13.143	< .001	0.141
	4b	99†	2.6	*1.1*	2.4	*1.0*	1.1	*0.7*	1.5	*0.8*	2.6	*0.9*	2.4	*0.9*	1.4	*0.8*	2	*0.9*	1.487	0.226	0.015
Inner features	1	96	1.5	*1.0*	1.5	*1.0*	0.5	*0.7*	0.9	*0.7*	1.7	*1.0*	1	*0.9*	0.6	*0.6*	1	*0.8*	28.866	< .001	0.233
	2	60	1.5	*0.8*	1.5	*0.8*	0.6	*0.7*	0.9	*0.7*	1.8	*0.9*	1.3	*0.8*	0.6	*0.6*	1.1	*0.8*	14.300	< .001	0.195
Atypical	3a	61	2.2	*1.0*	2	*1.0*	1	*0.7*	1.5	*0.8*	2.5	*1.0*	2.1	*0.8*	1.4	*0.8*	1.8	*0.9*	0.069	0.793	0.001
	3b	81†	1.4	*0.9*	1.2	*0.9*	0.8	*0.6*	1.2	*0.8*	2	*1.3*	1.6	*0.9*	1.2	*0.9*	1.4	*0.9*	1.599	0.210	0.020
Blurred	4a	81†	1.2	*0.8*	1.3	*0.7*	0.5	*0.5*	0.7	*0.6*	1.9	*1.0*	1.4	*0.8*	0.7	*0.8*	1.1	*0.7*	19.568	< .001	0.197
	4b	99†	1.2	*0.9*	1.1	*0.8*	0.5	*0.6*	0.6	*0.7*	1.6	*0.9*	1.3	*0.9*	0.7	*0.9*	1	*0.8*	8.643	0.004	0.081

† Indicates participants in a given image condition of a within-subject experiment.

A likely explanation for that pattern is that the appearance of stable men is even more stable than that of stable women. Women with long hair can present small variations in hairstyle, even if the colour and length are constant, for example by tying their hair up or by straightening/waving it. By contrast, men with shorter hair cannot present this type of small variations. Consequently, on average, extra-facial features and coarse information could carry more weight in men than in women, and small variations in the appearance of women could help refine the representations of their face despite a relatively stable appearance.

Validation of IMDb StarMeter ranks as a proxy of exposure

We demonstrated in Devue and colleagues (2019) that StarMeter ranks available on the IMDb website were a good proxy of exposure to a set of actors from a specific TV show, as they correlated with screen times (r = -0.441, p = 0.001). Unlike here, we had used a selected sample of 32 participants who had watched the entirety of the TV show, providing an excellent control of exposure to individual actor faces. In the current situation, we were unable to calculate screen times of individual actors or to control individual exposure of participants to actors. We thus calculated Pearson's correlations between average hit rates per actor in each image condition of each experiment and their StarMeter rank (as used in Experiments 1 to 3, and Experiment 4, respectively) to explore how valid a measure of exposure StarMeter ranks are in uncontrolled learning conditions. The correlations ranged from -.604 to -.788 and were all significant, see **Table 2**.

Table 2. Associations between StarMeter ranks and mean hit rates per actor in different image condition and experiments.

Test image	Experiment	Sample origin	N	Number of actors	Pearson's r	p	Lower 95% CI	Upper 95% CI
Headshot	2	NZ	60	96	-0.773	< .001	-0.843	-0.678
	3a	NZ	62	96	-0.788	< .001	-0.853	-0.697
	3b	US	81†	96	-0.725	< .001	-0.808	-0.614
	4a	NZ	81†	72	-0.728	< .001	-0.821	-0.597
	4b	NZ	99†	72	-0.681	< .001	-0.788	-0.533
Inner features	1	NZ	100	96	-0.634	< .001	-0.740	-0.496
	2	NZ	60	96	-0.666	< .001	-0.764	-0.537
Atypical	3a	NZ	61	96	-0.680	< .001	-0.775	-0.556
	3b	US	81†	96	-0.627	< .001	-0.735	-0.488
Blurred	4a	NZ	81†	72	-0.629	< .001	-0.751	-0.465
	4b	NZ	99†	72	-0.604	< .001	-0.733	-0.433

Note. Sample origin and N refers to participants tested in our recognition tests. Number of actors refers to the number of individual actors used in a given test. † indicates participants in a given condition of an experiment with a within-subject design.

These results thus validate our use of StarMeter ranks as a proxy of exposure since smaller StarMeter ranks, which indicate a higher media visibility, are associated with higher recognition rates. Results of the same correlations calculated between mean hit rates on 90 actors and their StarMeter ranks from Devue et al. (2019)'s data were comparable, r = -0.628, p < 0.001, 95% C.I. = -0.739 – -0.484 with typical images (N = 16), and r = -0.448, p < 0.001, 95% C.I. = -0.600 – -0.266 with atypical images (N=16). We note that associations between mean hit rates and StarMeter ranks are numerically larger in the current series of experiments than in our previous work, but this is likely due to the larger samples we used to compensate for the aforementioned lack of control on individual exposure and on individual face recognition abilities.

Finally, to check the validity of StarMeter ranks in different English speaking geographical areas, we calculated Pearson's correlations between hit rates per individual actor headshots in a sample from the US and in the different NZ samples used in different experiments. Results indicate large positive associations between hit rates in the two populations, with correlation coefficients ranging from 0.685 to .874, see **Table 3**. This confirms that actors we selected based on the US-based IMDb website have comparable visibility in both populations.

Table 3. Associations between hit rates for individual actor headshots in one US sample and in multiple NZ samples used in different experiments.

Experiment (US sample)	Experiment (NZ samples)	Number of actors	Pearson's r	p	Lower 95% CI	Upper 95% CI
3b	2	96	0.830	< .001	0.756	0.884
	3a	96	0.874	< .001	0.817	0.915
	4a	72	0.685	< .001	0.539	0.791
	4b	72	0.755	< .001	0.635	0.840

General discussion

We conducted four famous face recognition experiments on a total of 604 participants to test the two main assumptions of a cost-efficient mechanism of face learning, namely that representational weights are distributed contingent of the relative stability of individual faces, and following a dynamic coarse-to-fine encoding over the course of familiarisation.

Impact of stability and exposure on facial representations. We have considered the impact of intrinsic characteristics of famous faces on recognition performance and we show for the first time to our knowledge that the relative stability in appearance of individual faces affects recognition performance. Unexpectedly, in all experiments using intact headshots, we found that overall, famous faces with a stable appearance were better discriminated from strangers than faces that display looks that are more variable. In line with computer simulations (Burton et al., 2016) and recent studies on humans (Devue et al., 2019), we also found that recognition performance improves with increased exposure, confirming that humans build facial representations that evolve to become more reliable.

Our manipulation of popularity levels by means of an objective index of exposure (i.e. the StarMeter ranks on IMDb) allows nuancing these results and shows that, all else being equal, stability affects recognition performance in different ways along the course of familiarisation with faces. Specifically, in earlier stages of learning, stability in appearance supports recognition compared to variability, suggesting that stable faces benefit from representations that are more reliable at first. Over time, a shift in performance occurs and recognition of variable faces is better than that of stable faces, consistent with the idea that variations in appearance yields more reliable representations by encouraging more

refinement[7]. Further, while sensitivity to both variable and stable faces increase with exposure, the improvement is significantly larger with variable faces than with stable faces, suggesting that once a representation of a stable face is formed, it does not evolve as much and remains coarser compared to variable faces. The relative benefits of stability in earlier stages of familiarisation are also larger than the benefits of variability in later stages of familiarisation, a pattern that replicated across all experiments using intact images and that explains the overall advantage of stable faces over variable faces. In sum, our results consistently show that the quality of facial representations is the product of a given face's stability in appearance and its interplay with exposure, in line with hypotheses drawn from a cost-efficient mechanism of face learning.

Contrary to what we found here, recent lab-based face learning studies have shown that exposure to high degrees of variability in images of faces—both in viewing conditions and in appearance—improves recognition of newly learned faces relative to stable viewing conditions, even after a single brief learning session (Burton et al., 2016; Kramer, Manesi, et al., 2018, Robins et al., 2019). This seems inconsistent with the advantage for stable faces compared to variable faces we found in less popular actors and with the overall benefit of stability we observe. This apparent discrepancy is likely due to differences in learning supervision when learning new faces in the lab and when learning faces in the real world. In the lab, faces are often learned under supervised conditions, and so observers can take advantage of natural variations in images to refine their representations with the explicit knowledge that a set of images shows the same person. In contrast, when we encounter

[7] Note that during the original selection of actors, we purposefully left a gap in StarMeter ranks between popular and less popular ones. We can thus assume that recognition rates of variable and stable actors would be equivalent at some intermediate levels of popularity.

emerging actors in the real world, we often learn their faces incidentally and with low levels of supervision—for example, those can be in the form of credits or feedback from peers. If we are correct in assuming that face encoding operates parsimoniously, then a default assumption must be that the appearance of a newly encountered face is stable and will not change in the future, leading to the creation of a coarse representation. One can only revisit this assumption with repeated exposure to a person and the realisation that their appearance vary. This revision is most likely more challenging when an observer is not aware that they are viewing a person they have seen before than when they are explicitly told so. Therefore, if an emerging actor acts in several movies with the same appearance, we have more opportunity to recognise them based on the same coarse representation from one movie to another. By contrast, if an emerging actor appears with a different appearance in different movies, we may build multiple coarse representations that include different large-scale peripheral elements on each occasion and fail to recognise them as the same person across encounters. The benefit of associating various depictions of a face with a single identity occurs even when simultaneously viewing multiple images of a person in the lab. Indeed, when viewing a mix of different images of multiple people, participants are better at sorting images per identity when told how many different identities there are than when they are not informed or misinformed about it, in which case they tend to interpret singles identities as multiple identities (Andrews et al., 2015; Menon et al., 2018). Therefore, our data suggest that with low levels of learning supervision, variability in appearance has a negative impact on learning compared to stability, because differences in appearance are interpreted as differences in identity.

At the neural level, the benefits of learning stable faces with low levels of supervision across episodic encounters could translate in larger overlaps in activity patterns in the

anterior hippocampus from one encounter to another relative to variable faces, leading to a faster consolidation of that pattern (see e.g. Sekeres et al., 2018). In supervised conditions, influences from anterior cortical areas linked to the explicit knowledge that one is viewing multiple photos of the same person may help consolidate overlapping patterns of activity from one encounter to another. In faces with a variable appearance, the consolidation of more restricted overlapping patterns of activations compared to stable faces might correspond to the greater refinement of features that are common between images or encounters. Recent research has shown that refinement of representations with increased familiarity is indexed by the N250 component, and in similar ways for famous and personally familiar faces (Wiese et al., 2021). Future developments of that research could incorporate stability in appearance to examine how it modulates ERP responses.

More generally, the reasoning derived from our framework also explains the poor performance classically observed with new faces learned in non-ecological laboratory conditions. When an observer is learning a limited set of faces from single pictures, a cost-efficient encoding mechanism would lead to assume that the stimulus is stable, will not change in the future, and so to favour coarse elements of the person's appearance (e.g. the shape of the hairline in the given view, hair colour) or even diagnostic pictorial elements (e.g. a difference in background colour or a photographic artefact). This would then yield low cost representations with low generalisability and to poor performance in a subsequent memory test that uses images where the appearance, the view or pictorial artefacts have changed and/or where distractors display gross resemblances with learned faces (for a recent example with viewpoint, see Flack et al., 2019). The same reasoning can also help explain poor performance with new faces briefly encountered in the real world, for example, when one is witnessing a crime.

Content of facial representations. The comparison of recognition performance with typical headshots of actors and with images containing partial or atypical information gives us some clues on the content of facial representations and on the contribution of different types of information.

Peripheral and inner features. In initial stages of familiarisation, recognition of both stable and variable faces is greatly improved by the presence of peripheral information compared to internal features alone (Experiment 2). Contrary to the view drawn from PCA models that recognition of familiar faces relies on an average representation of inner features, the presence of peripheral features also improved recognition of more familiar faces. This suggests that all faces in our set were processed holistically, in line with studies showing that the holistic processing of unfamiliar faces is disrupted by the removal of external features (García-Zurdo et al., 2018; Toseeb et al., 2012) or that recognition of familiar faces is impaired when extra-facial features are altered (Carbon, 2008; Devue et al., 2019; Sinha & Poggio, 1996). Consistent with seminal studies showing a stronger reliance on peripheral features for less familiar faces than for more familiar ones (Campbell et al., 1995; Ellis et al., 1979), we observe that peripheral features facilitate the correct discrimination of familiar faces from strangers proportionally more for less popular faces than for more popular ones (see Figure 5A). The cost-efficient theory we have proposed provides a plausible encoding mechanism for present and past data: representational weights are broadly distributed over large-scale information at first, forming low-cost coarse representations, to converge towards internal information over time, giving more costly but more reliable refined representations.

Typical information. We show that headshots with the most typical individual appearance were better recognised than headshots deviating from that appearance, regardless of their popularity or relative stability (Experiment 3). This suggests that representations give more weight to facial information encountered more frequently, even for variable aspects in the case of faces that change appearance from one encounter to another. We can speculate that at the neural level, activations associated with these variable aspects are more likely to consolidate for those patterns of activations that reoccur more over time.

We note that discrimination performance was overall lower on the Mechanical Turk sample from the US in Experiment 3b than in other experiments, but that the pattern of performance with headshots seen in other experiments nonetheless replicated.

Coarse information. Like with other types of test images, stable faces were better recognised from blurred images than variable faces in earlier stages of familiarisation. Contrary to our expectations, we observed the same shift as in other image conditions whereby variable faces were better recognised from blurred images than stable ones in later stages of familiarisation. However, the size of that effect what the smallest across all the comparisons done between the two types of faces in different image conditions. We had hypothesised that if representations of stable faces are coarser than representations of variable faces, then those latter should be less well recognised from blurred images. This reasoning emerged from the assumption that representational weights might converge towards more detailed internal information, *to the detriment* of other areas at the periphery, as if the pool of representational weights available per individual face was finite. Instead, in line with research on face perception (Goffaux et al., 2005; Peters et al., 2018;

Weibert et al., 2018), current data suggest that representations of all faces have a foundation of coarse information, onto which finer information may be *added*.

Further, the fact that exposure increased sensitivity even from blurred images suggests that more discriminative information is extracted within coarse information over time. A likely candidate behind that improvement is the refined coding of configural information (i.e. spatial relationships between features) contained in blurred images (Goffaux et al., 2005), beyond larger scale and/or peripheral information (i.e. head shape, hair colour) encoded at first. The advantage of variables faces in later stages of familiarisation seems to rest both on that more refined coding of coarse information and the incorporation of facial details into representations. Future research where a more systematic manipulation of spatial frequencies available in test images is conducted should help confirm these hypotheses.

Integration of current findings. Altogether, our series of experiments suggest that faces are represented via holistic representations based on coarse information and that representational weights are added as needed to encode facial details and their relationships with higher resolution. When changeable features remain stable over time, representational weights remain broadly distributed over large-scale extra-facial information and internal features are encoded at lower resolution. Coarse representations are cheap but carry the risk of poor discrimination between similar individuals. For efficiency purposes, they must thus be favoured when we encounter new people and have no reason to assume that they will change or that we will see them again in the future. They could also be favoured when episodic encounters with an individual are consistently linked to a specific context and that gross information is discriminative enough in that context,

perhaps contributing to well-known difficulties when a person appears in a different context (Mandler, 1980). The more we experience variations in a person's appearance over encounters, the highest the resolution of invariant information needs to be to guarantee recognition. Finer representations are more costly but more discriminative, and the face recognition system must turn to them as we get to know people and demands for recognition out of context increase.

Holistic representations based on coarse information are adaptive not only because they are cheap but also because they give us a chance to recognise people from a distance when facial details are not available. Although recognition is not always accurate at large distances (Loftus & Harley, 2005; McKone, 2009), coarse information at least allows us to form an hypothesis on someone's identity, that finer information can confirm or infirm as it becomes available. Such coarse-to-fine verification process is also at play during scene identification (Schyns & Oliva, 1994).

Implications and future directions. Our series of experiments confirm that the large amount of data on celebrities available on the internet can be exploited to advance psychology research. The StarMeter ranks we have used to create sets of images of famous faces that are comparable in exposure have generated highly replicable results despite differences in populations used, variations in experimental paradigms (between vs. within-subject, lab-based vs. online, image types blocked or presented sequentially) and different items included in image sets across experiments. Importantly, as the StarMeter ranks are a dynamic measure, stimuli sets must evolve over time as well.

While recent research has emphasised the use of uncontrolled natural stimuli to study face recognition in a more ecological manner, we show that an approach maximising

internal and external validities may be more productive. Importantly, the results presented here show that studying face recognition based on averaged performance on indiscriminate heterogeneous sets of face images may muddy waters. This is striking through the interaction we consistently found between popularity and stability in appearance.

The current series of experiments is not without its own shortcomings in that regard. For example, we referred to and studied the role of inner features as a group, although we explicitly assumed that single or multiple features within that group could carry more or less representational weight. For example, past ERP research showed that the eyes are strong identity cues, more reliable than other features like the mouth (Mohr et al., 2018; Nemrodov et al., 2014). Therefore, the eyes may carry more representational weight than other inner features, which could result from their central position in the head, allowing to take in surrounding coarse information. However, regular changes (e.g. make-up and/or swapping between glasses and contacts) or occlusions (e.g. with hair of sunglasses) of the eyes area in a given individual face could lead to refine representations of other aspects less affected by changes and occlusions (e.g. the nose). The role of individual facial features as a function of their intrinsic characteristics in terms of stability or of other aspects like their distinctiveness will thus be the object of future research. Moreover, exploratory analyses including the sex of the actors have suggested differences in discrimination patterns of popular male and female actors, whereby discrimination sensitivity to stable women increased with exposure more than discrimination sensitivity to stable men. In other words, women faces may have been driving the small improvement seen over time for stable faces. This might be due to stable women displaying more variations than stable men (e.g. larger differences in hair styling despite consistent length and colour in women than in men) and will warrant further investigations too.

Conclusions

We present a new account of face learning and familiarisation that takes stability in appearance into account. We posit that representations are cost-efficient and laid out differently depending on intrinsic characteristics of individual faces. We show that despite comparable levels of popularity of actors like Brad Pitt and Harrison Ford, the representation of people like the former, who have a variable look, are more refined than that of people like the latter, who have a more consistent appearance. Although it leads to maintain coarser representations, stability facilitates recognition in earlier stages of familiarisation. Harrison Ford's signature look helped us remember him from encounter to encounter, and his face must have become familiar faster than the face of Brad Pitt. This account is integrative in nature and resolves conflicting theoretical conceptions as to what type of facial information is encoded and whether qualitatively different processes are used for unfamiliar and familiar faces. Indeed, seemingly conflicting empirical data in past research may be the result of the same cost-efficient face learning mechanism and its consequences over time. This account also generates numerous hypotheses for future research, which will hopefully further our understanding of how most of us are able to recognise large amounts of faces despite large memory constraints.

Context of research

The reasoning behind the cost-efficient mechanism of face learning has emerged from unexpected findings reported in Christel Devue's 2019 paper [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]. We had found that recognition of familiar faces learned in the rich conditions of a TV show was impaired by superficial

changes in appearance regardless of the degrees of exposure they had received, including for very prominent actors. Results also showed that recognition errors (false recognition of strangers or confusion between people) were often due to gross resemblances in extra-facial features. The will to understand the root of those results, to close the gap between what is known of unfamiliar and familiar face processing and to take real-world conditions and memory constraints of humans into account led to the current theoretical developments and empirical work.

References

Abudarham, N., & Yovel, G. (2018). Same critical features are used for identification of familiarized and unfamiliar faces. *Vision Research*, *October 2017*, 1–7. https://doi.org/10.1016/j.visres.2018.01.002

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

Andrews, T. J., Davies-Thompson, J., Kingstone, A., & Young, A. W. (2010). Internal and External Features of the Face Are Represented Holistically in Face-Selective Regions of Visual Cortex. *Journal of Neuroscience*, *30*(9), 3544–3552. https://doi.org/10.1523/JNEUROSCI.4863-09.2010

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. https://doi.org/10.1016/j.cognition.2016.12.012

Brédart, S., & Devue, C. (2006). The accuracy of memory for faces of personally known individuals. *Perception*, *35*(1), 101–106. https://doi.org/10.1068/p5382

Brédart, S., & Young, A. W. (2004). Self-recognition in everyday life. *Cognitive Neuropsychiatry*, *9*(3), 183–197. https://doi.org/10.1080/13546800344000075

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, *23*(1), 1–31. https://doi.org/10.1016/S0364-0213(99)80050-0

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125

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., 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

Campbell, R., Walker, J., & Baron-Cohen, S. (1995). The development of differential use of inner and outer face features in familiar face identification. In *Journal of Experimental Child Psychology,* 59, Issue 2, pp. 196–210). https://doi.org/10.1006/jecp.1995.1009

Carbon, C. C. (2008). Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception*, *37*(5), 801–806. https://doi.org/10.1068/p5789

Devue, C., Collette, F., Balteau, E., Degueldre, C., Luxen, A., Maquet, P., & Brédart, S. (2007). Here I am: the cortical correlates of visual self-recognition. *Brain Research*, *1143*(1), 169–182. https://doi.org/10.1016/j.brainres.2007.01.055

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

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external feature : Some implications for theories of face recognition. *Perception*, *8*, 431–439. https://doi.org/10.1068/p080431

Etchells, D. B., Brooks, J. L., & Johnston, R. A. (2017). Evidence for view-invariant face recognition units in unfamiliar face learning. *Quarterly Journal of Experimental Psychology*, *70*(5), 874–889. https://doi.org/10.1080/17470218.2016.1248453

Flack, T. R., Harris, R. J., Young, A. W., & Andrews, T. J. (2019). Symmetrical viewpoint representations in face-selective regions convey an advantage in the perception and recognition of faces. *Journal of Neuroscience*, *39*(19), 3741–3751. https://doi.org/10.1523/JNEUROSCI.1977-18.2019

Gao, Z., Ding, X., Yang, T., Liang, J., & Shui, R. (2013). Coarse-to-Fine Construction for High-Resolution Representation in Visual Working Memory. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0057913

García-Zurdo, R., Frowd, C. D., & Manzanero, A. L. (2018). Effects of facial periphery on unfamiliar face recognition. *Current Psychology*, *2018*, 1–7. https://doi.org/10.1007/s12144-018-9863-1

Goffaux, V., Hault, B., Michel, C., Vuong, Q. C., & Rossion, B. (2005). The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. *Perception*, *34*(1), 77–86. https://doi.org/10.1068/p5370

Goffaux, V., & Rossion, B. (2006). Faces are "spatial"--holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(4), 1023–1039. https://doi.org/10.1037/0096-1523.32.4.1023

Hancock, P., Bruce, V., & Burton, A. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330–337. http://www.ncbi.nlm.nih.gov/pubmed/10962614

Hayes, T., Morrone, M. C., & Burr, D. C. (1986). Recognition of positive and negative bandpass-filtered images. *Perception*, *15*(5), 595–602. https://doi.org/10.1068/p150595

Jenkins, R., Dowsett, A. J., & Burton, A. M. (2018). How many faces do people know? *Proceedings of the Royal Society B: Biological Sciences*, *285*. https://doi.org/10.1098/rspb.2018.1319

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1571), 1671–1683. https://doi.org/10.1098/rstb.2010.0379

Johnston, A., Hill, H., & Carman, N. (2013). Recognising faces: Effects of lighting direction, inversion, and brightness reversal. *Perception*, *42*(11), 1227–1237. https://doi.org/10.1068/p210365n

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: a review. *Memory*, *17*(5), 577–596. https://doi.org/10.1080/09658210902976969

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., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*, *172*(June 2017), 46–58. https://doi.org/10.1016/j.cognition.2017.12.005

Lakens, D. (2014). Performing High-Powered Studies Efficiently With Sequential Analyses. *European Journal of Social Psychology*, *44*, 701–710.

Lander, K., & Bruce, V. (2003). The role of motion in learning new faces. *Visual Cognition*, *10*(8), 897–912. https://doi.org/10.1080/13506280344000149

Loftus, G. R., & Harley, E. M. (2005). Why is it easier to identify someone close than far away? *Psychonomic Bulletin and Review*, *12*(1), 43–65. https://doi.org/10.3758/BF03196348

Longmore, C. A., Santos, I. M., Silva, C. F., Hall, A., Faloyin, D., & Little, E. (2017). Image dependency in the recognition of newly learnt faces. *Quarterly Journal of Experimental Psychology*, *70*(5), 863–873. https://doi.org/10.1080/17470218.2016.1236825

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

McKone, E. (2009). Holistic processing for faces operates over a wide range of sizes but is strongest at identification rather than conversational distances. *Vision Research*, *49*(2), 268–283. https://doi.org/10.1016/j.visres.2008.10.020

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

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

Mohr, S., Wang, A., Engell, A. D., & Hall, S. M. (2018). Early identity recognition of familiar faces is not dependent on holistic processing. *BioRxiv Preprint*, *March 2018*, 1–27. https://doi.org/10.1093/scan/nsy079

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

Nemrodov, D., Anderson, T., Preston, F. F., & Itier, R. J. (2014). Early sensitivity for eyes within faces: A new neuronal account of holistic and featural processing. *NeuroImage*, *97*, 81–94. https://doi.org/10.1016/j.neuroimage.2014.04.042

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks. *Trends in Cognitive Sciences*, *22*(9), 794–809. https://doi.org/10.1016/j.tics.2018.06.006

Peters, J. C., Goebel, R., & Goffaux, V. (2018). From Coarse to Fine: Interactive feature processing precedes local feature analysis in human face perception. *Biological Psychology*, *138*(October 2017), 1–10. https://doi.org/10.1016/j.biopsycho.2018.07.009

Pilz, K. S., Thornton, I. M., & Bulthoff, H. H. (2006). A search advantage for faces learned in motion. *Experimental Brain Research*, *171*(4), 436–447. https://doi.org/10.1007/s00221-005-0283-8

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*(5), 1–9. https://doi.org/10.1080/17470218.2015.1136656

Robins, E., Susilo, T., Ritchie, K. L., & Devue, C. (n.d.). *Within-person variability promotes learning of internal facial features and facilitates perceptual discrimination and memory*. https://doi.org/10.31219/osf.io/5scnm

Robins, Elliott, Susilo, T., Ritchie, K., & Devue, C. (2018). *Within-person variability promotes learning of internal facial features and facilitates perceptual discrimination and memory*. https://osf.io/8tndq/

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition. *Psychological Science*, *5*(4), 195–200. https://doi.org/10.1111/j.1467-9280.1994.tb00500.x

Sekeres, M. J., Winocur, G., & Moscovitch, M. (2018). The hippocampus and related neocortical structures in memory transformation. *Neuroscience Letters*, *680*(August 2017), 39–53. https://doi.org/10.1016/j.neulet.2018.05.006

Sinha, P., & Poggio, T. (1996). I think I know that face... *Nature*, *384*(6608), 404–404. https://doi.org/10.1038/384404a0

Tanaka, J. W., & Simonyi, D. (2016). The “parts and wholes” of face recognition: a review of the literature. *Quarterly Journal of Experimental Psychology*, *69*(10), 1876–1889. https://doi.org/10.1080/17470218.2016.1146780.The

Tong, F., & Nakayama, K. (1999). Robust representations for faces: evidence from visual search. *Journal of Experimental Psychology: …*, *25*(4), 1016–1035. http://psycnet.apa.org/journals/xhp/25/4/1016/

Toseeb, U., Keeble, D. R. T., & Bryant, E. J. (2012). The significance of hair for face recognition. *PLoS ONE*, *7*(3), 1–8. https://doi.org/10.1371/journal.pone.0034144

Weibert, K., Flack, T. R., Young, A. W., & Andrews, T. J. (2018). Patterns of neural response in face regions are predicted by low-level image properties. *Cortex*, *103*, 199–210. https://doi.org/10.1016/j.cortex.2018.03.009

Wiese, H., Hobden, G., Siilbek, E., Martignac, V., Flack, T. R., Ritchie, K. L., Young, A. W., & Burton, A. M. (2021). Familiarity Is Familiarity Is Familiarity: Event-Related Brain Potentials Reveal Qualitatively Similar Representations of Personally Familiar and Famous Faces. *Journal of Experimental Psychology: Learning Memory and Cognition*, *November*. https://doi.org/10.1037/xlm0001063

Young, A. W., & Burton, A. M. (2018). Are We Face Experts? *Trends in Cognitive Sciences*, *22*(2), 100–110. https://doi.org/10.1016/j.tics.2017.11.007