The impact of stability in appearance on the development of facial representations

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

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

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

Manuscript accepted for publication in Cognition (July 2023)

*Corresponding author: School of Psychology, Victoria University of Wellington, Kelburn Parade, Wellington 6012, New Zealand

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

Author's notes: 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: https://osf.io/qd5y3/register/564d31db8c5e4a7c9694b2be; Amended plans for Experiment 2 and 3: https://osf.io/h5f6s/register/564d31db8c5e4a7c9694b2c0 and https://osf.io/afm4e. We posted a former draft of our manuscript on the preprint server PsyArXiv under the following URL https://psyarxiv.com/fqd3v/ on 21 December 2021.

Acknowledgements: We thank Serge Brédart, Arnaud D'Argembeau and Valentine Vanootighem for proofreading and commenting on an earlier version of on the manuscript, as well as two reviewers for their constructive feedback. We thank the Susilo lab at Victoria University of Wellington for their support with this research.

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 present three experiments serving as proof of concept for a cost-efficient mechanism of face learning describing how facial representations form over time and accounting for recognition errors. We propose that the encoding of facial information is dynamic and modulated by the intrinsic stability in individual faces' appearance. We drew on a robust and ecological method using a proxy of exposure to famous faces in the real world and manipulated test images to assess the prediction that recognition of famous faces is affected by their relative stability in appearance. We consistently show that stable facial appearances (like Tom Cruise's) facilitate recognition in early stages of familiarisation but that performance does not improve much over time. In contrast, variations in appearance (like Jared Leto's) hinder recognition at first but improve performance with further media exposure. This pattern of results is consistent with the proposed cost-efficient face learning mechanism whereby facial representations build on a foundation of large-scale diagnostic information and refine over time if needed. When coarse information loses its diagnostic value through the experience of variations in appearance across encounters, diagnostic facial details and/or their spatial relationships must receive more weights, leading to refined representations that are more discriminative and reliable than representations of stable faces.

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

Highlights

- We propose that face learning is cost-efficient and modulated by the relative stability of individual faces' appearance over episodic encounters, over and beyond changes in viewing conditions, and that encoding of facial information operates in a flexible and coarse-to-fine manner.
- We predicted that actors' stability in appearance and their levels of media exposure should interact to produce representations that are more or less reliable in a face recognition task.
- Results suggest a contribution of the most typical peripheral information to holistic facial representations.
- Data show an overall recognition advantage for famous faces with a stable appearance compared to faces displaying more variable looks.
- This advantage is due to stability facilitating recognition in earlier stages of familiarisation but performance is reversed to the benefit of variable faces with higher degrees of media exposure.
- This pattern is in line with a cost-efficient encoding mechanism yielding coarser representations for stable faces and promoting refinement when variations in appearance are observed over episodic encounters.
- This pattern seems in contrast with demonstrated benefits of exposure to variations during face learning but we discuss the role of learning supervision or lack therefore in laboratory and real-world conditions to explain the unexpected benefits of stability in appearance we found.
- Research on familiar face recognition and face learning must consider individual facial characteristics to help refine existing theoretical accounts.

The impact of stability in appearance on the development 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). Understanding this transition in performance between these extremes is the number one challenge to move research on face learning and recognition 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; Patterson & Baddeley, 1977; 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 (R. 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 in isolation, internal features are judged as more diagnostic of identity than external features (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-specialized 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 be the by-product of unfair comparisons. Unlike famous or personally familiar faces that have been learned in rich conditions (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. The latter learning conditions are insufficient to form three-dimensional representations of complex objects like faces and it was established that exposure to multiple viewpoints and/or movement improves learning (Etchells et al., 2017; 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. Recent studies examining neural changes produced after brief real-life encounters with new persons suggest these encounters are sufficient to yield activation in areas devoted to perception and memory for

faces and image-independent representations (Campbell & Tanaka, 2021; Popova & Wiese, 2023; Sliwinska et al., 2022).

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 (Jenkins & Burton, 2011). 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).

Recent human data partly support this rationale in 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). Interestingly, a recent study shows that a combination of changes in facial appearance, clothing and context help yield more reliable and durable facial representations than just systematic changes in viewpoints (Corpuz & Oriet, 2022).

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 and universally 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 valuable to humans. It seems that in humans, inner features are not always necessary nor sufficient to trigger recognition of familiar faces and so they might not always carry the most diagnostic information for a given face.

To resolve these apparent contradictions and explain how facial representations evolve as familiarity unfolds, we propose a parsimonious mechanism of face learning that integrates multiple aspects of existing theoretical accounts. 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 the encoding of coarser information (e.g., head silhouette, hairstyle and colour, light/dark pattern of inner features) is prioritised over that of finer details (e.g., details of the lips) because a coarse-to-fine prioritisation during encoding should incur fewer storage resources (Gao et al., 2013). This flexible and dynamic encoding mechanism would create *cost-effective* memory representations that start off as coarse but refine over time, particularly if changes in appearance are encountered (Corpuz & Oriet, 2022) 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) of people's facial appearance affects the quality of facial representations, specifically in terms of their spatial *resolution*. Coarse-to-fine processing is ubiquitous in scene, object, and face perception and depends on factors such as the type of categorisation task (e.g., Morrison & Schyns, 2001; Nakashima et al., 2008; Schyns & Oliva, 1994; Yan et al., 2022). Therefore, we propose that the scale of information encoded for a given face depends on its intrinsic characteristics, and on what is diagnostic enough to distinguish it from other faces. We further assume that the refinement of facial representations unfolds over longer time scale than the coarse-to-fine processing occurring during visual perception, and that memory refinement takes place across episodic encounters as a function of one's relative stability in appearance. When we view a face with a stable appearance, large-scale peripheral features and coarse information (e.g. hair colour

and style) are diagnostic and receive substantial representational weight. Moreover, details of inner features need not be encoded, yielding low-resolution representations. By contrast, when we observe that people change their appearance frequently through variations in hairstyle, hair colour, makeup or facial hair, the set of large-scale diagnostic features one might rely on decreases. 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 must receive more representational weight, thereby producing representations that include areas with higher resolution. 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.

Figure 1. Illustration of differences in stability in appearance in two individuals in terms of hairstyle, make-up and accessories (variable appearance on top and stable appearance at the bottom), over and beyond natural variations in facial expression, viewpoint, lighting, clothing and context, or variations due to camera artefacts. [Due to copyright restrictions, pictures used in the experiments are not shown. The people depicted here have provided permission.]

As a first step for testing the assumptions of this model, we present a series of three recognition experiments using actors' faces. 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, leading to very ecological encoding conditions compared to classical lab-based face learning tasks. Simultaneously, we can operate a strict selection of individual actors based on their physical appearance (see an illustration on **Figure 1**). Half of the actors had a stable appearance (e.g., Tom Cruise) and half had a variable appearance (e.g., Jared Leto). 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.

We examined whether and how the relative stability of one's appearance over encounters changes the likelihood that they are recognised. If efficient face encoding relied on averaging inner features as a group, as implicitly suggested by PCA models, one could expect comparable recognition performance regardless of variations in facial appearance, because recognition should rest upon robust representations of invariant inner features. Instead, we predicted different recognition performance for actors with a stable appearance and those with a more variable appearance. This pattern would in and of itself demonstrate that face encoding operates in a flexible and dynamic manner as a function of a face's relative stability.

Finally, to examine how stability in appearance modulates the evolution of representations over time and with increased exposure, 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 media exposure stable and variable actors

had received were comparable, based on an objective measure of public visibility available on the Internet Movie Database (IMDb). To get a snapshot of how reliable representations may be at different levels of exposure, 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 large-scale external features would converge towards inner features and their details over time. In other words, the encoding and consolidation of representations would operate based on statistical learning of stable elements of varying scales over encounters. We thus expected 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 individual exposure to individual actors amongst participants, we recruited a large sample of 100 first year psychology students in Experiment 1 (i.e. ten times more than needed from a priori power calculations). 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 local Ethics Committee.

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 (e.g. similar hairstyle, hair colour, facial hair across photos and no changes involving several 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 changeable dimensions. Actors' popularity was determined via the StarMeter ranks on IMDb pro, which reflect current popular interest for an actor and their visibility or exposure in the media—smaller ranks reflect higher popularity. Since we could not measure individual participants' exposure to individual actors, the logic here is that the more media exposure an actor has, the more likely it is that participants *as a group* will have been exposed to them, and so the better they should be recognised. These ranks were found to predict recognition performance of actors' faces 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 be 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).

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

[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.

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.

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 (i.e. yes, the face looks familiar, it has been seen before) or not via two response keys (1 = yes and 2 = no). Instructions thus emphasised visual familiarity with a face and specified that there was no need to remember the person's name or identity to judge that their 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) was manipulated between-subject in experiments 2 and 3a, and within-subject in Experiment 3b. We calculated *d'* based on hit rate (i.e., proportion

of “familiar” responses to actors) in each famous actor face category (i.e. popular variable, popular stable, less popular variable, less popular stable) and on false alarm rate (i.e., proportion of “familiar” responses to unfamiliar faces) in the corresponding image condition. Descriptive statistics (means and standard deviations) for familiarity judgments (i.e. hit and false alarm rates) and reaction times 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 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]. 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.

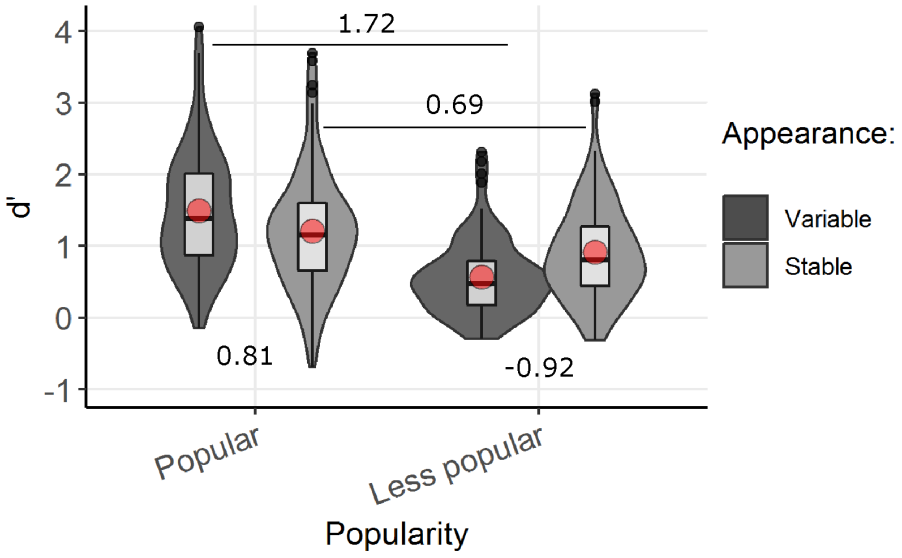

Figure 2. Results of Experiment 1. Discrimination performance (d') for images of cropped inner features as a function of popularity and appearance. Red circles show the mean, and boxplots show distribution in quartiles. Violin size is proportional to the distribution of performance in each condition. Values on the plot are effect sizes (Cohen's *d*) for paired-comparisons.

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 2**.

Benefit of increased media exposure/popularity. 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 media 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, even if Cohen's *d* indicates a large effect, $t(95) = -9.029$, $p_{one\text{-}tailed} = 1$, $d = 0.921$ (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; Corpuz & Oriet, 2022; Ritchie & Burton, 2017). We assume that this difference is due to unsupervised learning conditions in which actors' faces are often learned, contrasting with face learning in typical lab situations. Before actors become extremely famous, we might see them in multiple support roles without the explicit knowledge that they are the same person. In that situation, stability could help "put faces together" in that we are more likely to recognise someone who had similar appearances in different movies than someone who has changed. Stability may thus allow us to consolidate the representation of newly learned faces and of their inner features across episodic encounters.

[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 their representations develop to be 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 in compensating 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 when $p < .0182$. 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 3**. 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 media 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).

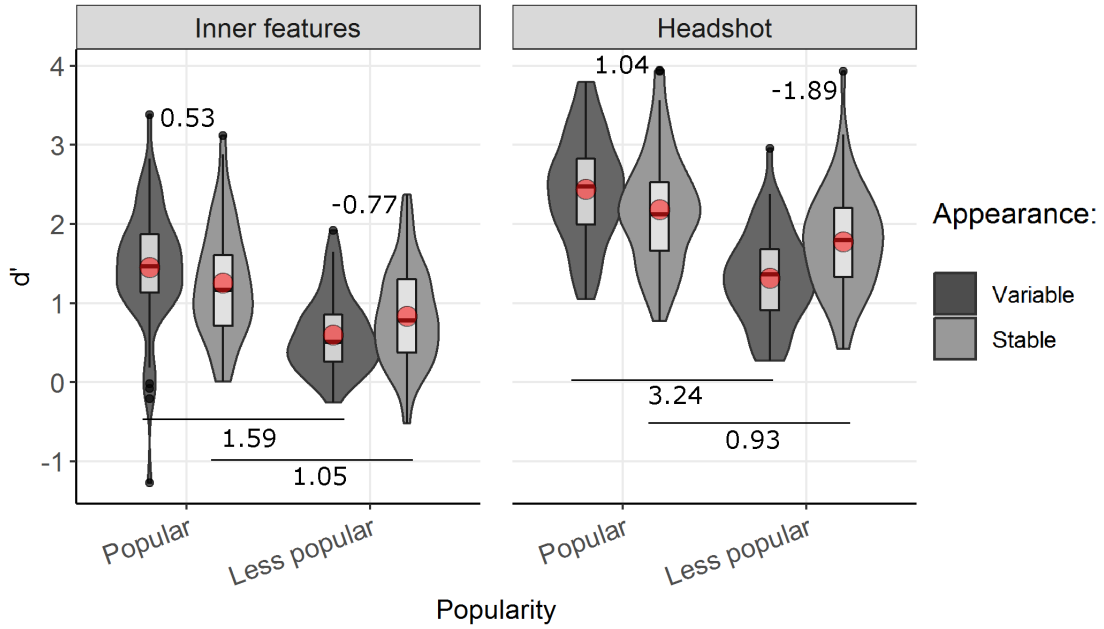

Figure 3. Results of Experiment 2. Discrimination performance (d') as a function of popularity, appearance of actors and image type(inner features vs. headshot). Red circles show the mean, and boxplots show distribution in quartiles. Violin size is proportional to the distribution of performance in each 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 media 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) 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; Toseeb et al., 2012). 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 seems to facilitate familiarisation since it improves recognition compared to variable appearances. 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. Further, this could be down to the mere fact that typicality effects are more subtle than effects from the removal of external features and that a between-subject design was underpowered in this situation despite a sample size similar to that in Experiment 2. To address these possibilities, 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 eight 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 provides 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 4** (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 4** (bottom panel).

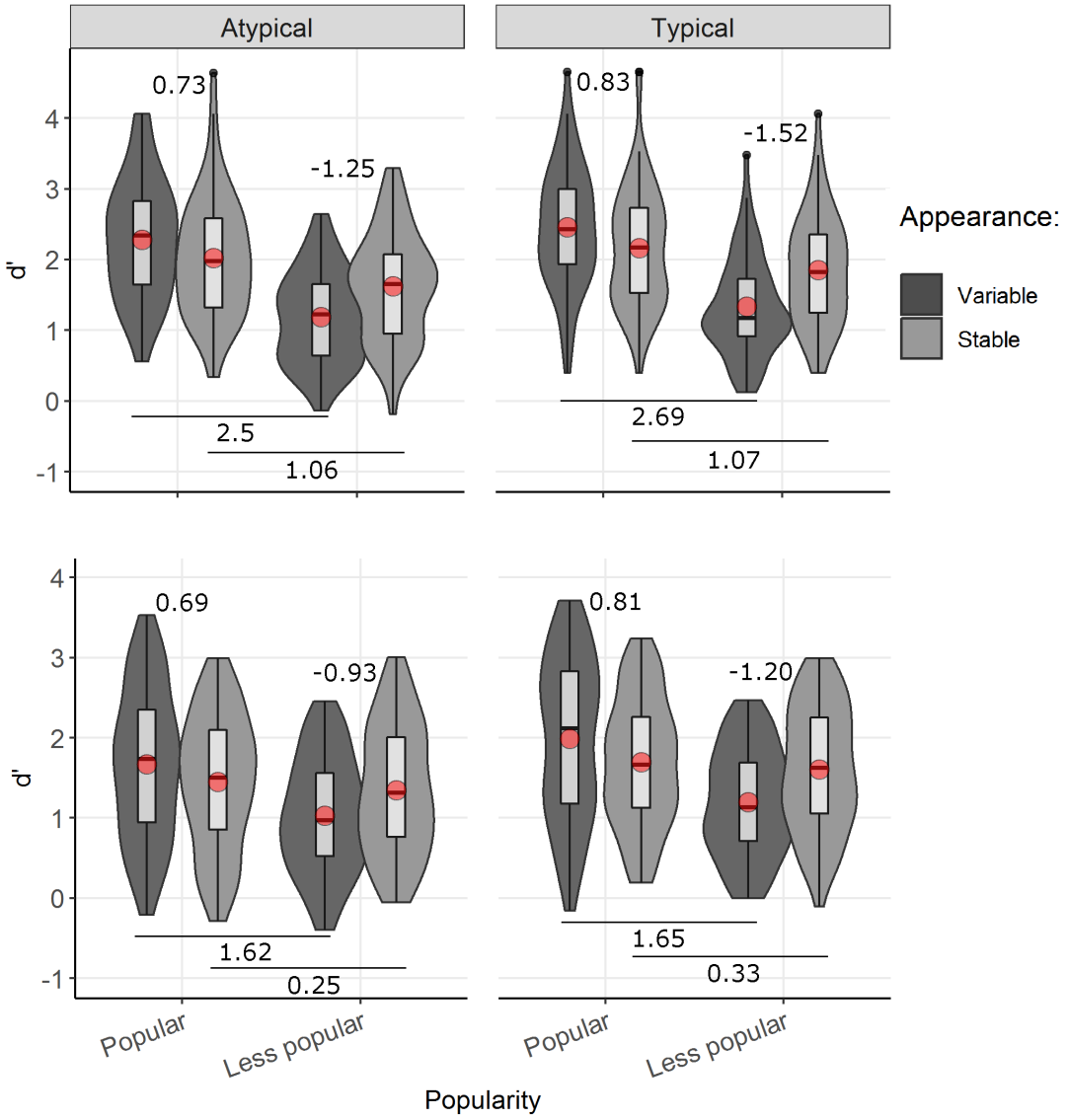

Figure 4. Results of Experiment 3a (top) and 3b (bottom). Discrimination performance (d') as a function of popularity and appearance, for typical and atypical images of actors. Red circles show the mean, and boxplots show distribution in quartiles. Violin size is proportional to the distribution of performance in each condition. Values on the plot are Cohen's d from paired-comparisons.

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).

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 internal features, recognition would have been equally good from typical and atypical images.

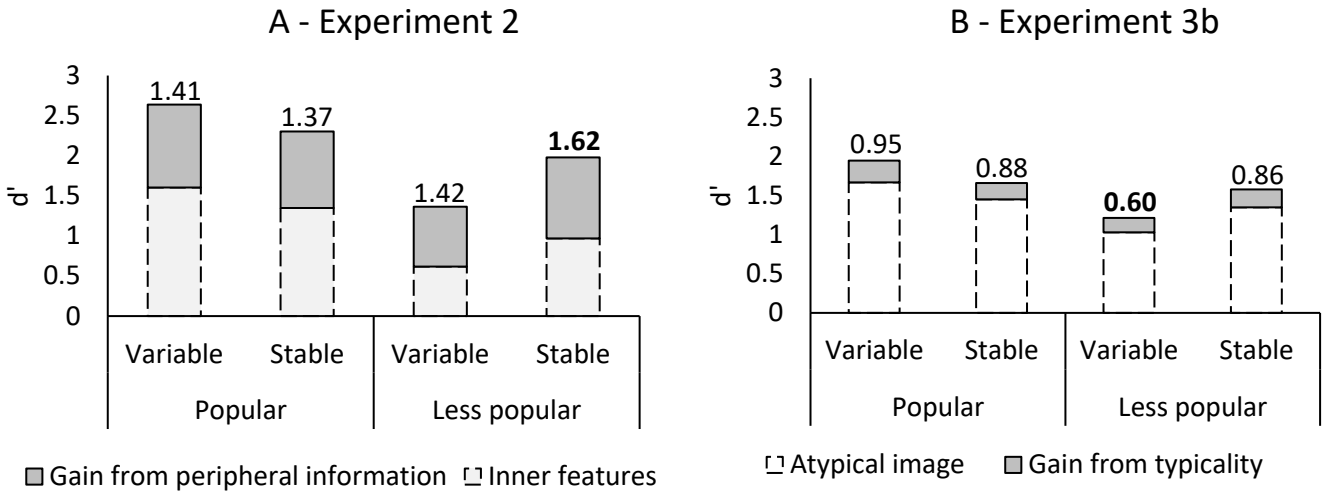

Figure 5. Comparisons of sensitivity to actors' faces in different image conditions, as a function of popularity and appearance. The top of the grey bar represents sensitivity to typical headshots, which was compared to recognition from inner features in Experiment 2 (N = 123 New Zealanders), from atypical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

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.

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 S7). In all experiments using intact images, 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 media exposure, discrimination of stable male faces did not improve with media 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, 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
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

† 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 display 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 the *actual* exposure of 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 individual participants' exposure to individual actor faces. In the current situation, we were unable to calculate screen times of individual actors or to objectively measure individual exposure of participants to each of them. We thus calculated Pearson's correlations between average hit rates per actor in each image condition of each experiment and their StarMeter rank to explore if these latter predict recognition performance and are thus a valid proxy of *probable* exposure for participants in uncontrolled learning conditions. The correlations ranged from -.604 to -.788 and were all significant, see **Table 2**.

These results thus validate our use of StarMeter ranks as a proxy measure of exposure since smaller StarMeter ranks, which indicate a higher media visibility, are associated with

higher recognition rates[4]. 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.

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
Inner features	1	NZ	96	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

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.

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

[4] Note that as part of another study, we collected familiarity ratings (1 = "not familiar" to 7 = "very familiar") from 35 independent judges on the same set of 96 actors' headshots. These mean ratings were also significantly correlated with StarMeter ranks, r = -0.422, p < 0.001, 95% C.I. = -0.591 – -0.265. This smaller correlation is likely due to Likert scales allowing for more variable ranges of responses than the "yes/no" responses compiled to yield hit rates in our recognition tasks and to the smaller sample used to collect those familiarity ratings.

experiments. Results indicate large positive associations between hit rates in the two populations, with correlation coefficients ranging from 0.83 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 two 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

General discussion

We conducted three famous face recognition experiments on a total of 420 participants to provide preliminary evidence for the two main assumptions of a cost-efficient mechanism of face learning, namely that the quality of facial representations specifically depends on the relative stability in appearance of individual faces, and that representations evolve 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 specifically affects recognition performance. Unexpectedly and in apparent contrast to previous research on lab-based face learning (e.g. Baker et al., 2017; Kramer, Young, et al., 2018; Ritchie & Burton, 2017), 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 media exposure, confirming that facial representations evolve to become more reliable.

Our manipulation of popularity levels by means of an objective index of media visibility/exposure (i.e. the StarMeter ranks on IMDb) allows nuancing these results. We show that, all else being equal, stability in appearance 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 variable faces are more likely to be recognised than stable faces, consistent with the idea that variations in appearance yield more reliable representations by encouraging more refinement[5]. Further, while sensitivity to both variable and stable faces increases with media 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 refine as much as representations of variable faces. The relative benefits of stability compared to variability in earlier stages of familiarisation are also larger than the benefits of variability compared to stability 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 suggest that the quality of facial representations is the product of a given face's

[5] 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/exposure.

stability in appearance and its interplay with exposure, in line with hypotheses drawn from a cost-efficient mechanism of face learning.

Unlike what we found here, recent lab-based face learning studies have shown that exposure to high degrees of natural variations 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 typically 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 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 comments from peers. If we are correct in assuming that face encoding operates parsimoniously, then a default assumption an observer makes 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 varies, which typically occurs readily in lab-based face learning studies. 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 the opportunity to recognise them based on the same coarse

representation from one movie to another. By contrast, if an emerging actor sports a different appearance in different movies, we may fail to recognise them as the same person across encounters. The benefit of associating various depictions of a face with a single identity was demonstrated in the lab. 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). Consistently, our data suggest that with low levels of learning supervision, variability in appearance has a negative impact on learning compared to stability, probably because differences in appearance are interpreted as differences in identity.

More generally, the reasoning derived from our framework also explains the poor performance classically observed with new faces learned in non-ecological laboratory conditions (see e.g. Hancock et al., 2000). 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 process would yield low cost representations with low generalisability and lead to poor performance in a subsequent memory test that uses images where the appearance, viewpoint, accessories or pictorial artefacts have changed and/or where distractors display gross resemblances with learned faces (see Flack et al., 2019a for a recent example with viewpoint; see Hsiao et al., 2022; Noyes et al., 2021 for recent examples with face masks). 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. On the flip side, stimuli in face learning or face perception studies in which external features are removed may force a finer processing of facial features than what would occur naturally since there are no coarse peripheral features to rely on. Such conditions may inflate the difficulties of participants that are over reliant on external features as in acquired or developmental prosopagnosia (see e.g., Jansari et al., 2015; Towler et al., 2018).

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 encoded 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 (R. 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 (e.g. most frequent hairstyle), even for variable aspects when someone changes their appearance from one encounter to another. We can speculate that at the neural level, activations associated with changeable aspects are more likely to consolidate for those patterns of activations that reoccur and overlap more over time (Sekeres et al., 2018).

Integration of current findings. Altogether, our series of experiments suggest that faces are represented via holistic representations, and that these representations refine over time to become more reliable depending on levels of variations in appearance they display. Our data seem consistent with the suggestion that 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. In other words, compared to variable faces, stable faces may ultimately lack one crucial type of variation, i.e. variations in appearance, amongst the set of variations that have been shown to improve face learning (e.g., Andrews et al., 2015; Burton et al., 2016; Ritchie & Burton, 2017).

In practice, this is not necessarily a problem as long as the target person does not change appearance, and coarse representations have the benefit of being cheap in terms of memory resources. However, coarse representations also carry the risk of poor discriminability between similar individuals. For efficiency purposes, they are thus presumably 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 recognition 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.

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 have comparable levels of media exposure have generated highly replicable results despite differences in populations used, variations in experimental paradigms, and varying levels of performance in different image conditions (i.e. lower with just inner features or with atypical images than with typical headshots).

One may actually be surprised or concerned by the consistency of the pattern of findings across experiments (i.e. interaction between Popularity and Stability) and we were

initially as well. We believe it is useful to keep in mind that one should not expect such consistency to apply at the individual (participant or even actor) level, as individual participants will undoubtedly vary in their recognition skills and in the amount and nature of exposure they would have had to each actor. Even when exposure is controlled, individual performance with individual faces is not predictable. In a research using participants who had watched all the episodes of a specific TV show (Devue et al., 2019), participants recognised overlapping but different sets of individual actors despite having all been exposed to the same faces. This is why we believe it is crucial to use StarMeter ranks in combination with large samples of participants and high numbers of actors per category as we have done here. This approach allows for average performance in each actor category to average out uncontrolled individual variations[6].

In further support of this approach relying on StarMeter ranks combined with large samples, actor-level analyses have shown that mean hit rates from English-speakers' samples on different continents were very strongly correlated (*r* ranging from .83 to .873). More importantly, StarMeter ranks were strongly correlated with hit rates (i.e. ranging from *r* = .627 with atypical headshots to *r* = .788 with typical headshots). A study by Ritchie and colleagues (2018) provides a relevant point of comparison. They examined individual correlations between the reported level of familiarity with five actors and the reported

[6] Experiment 3b in which typical and atypical images of each actor were presented to the same participants provides an opportunity to check the consistency of responses of individual participants to individual pictures of actors. We examined the responses of each participant (N = 81) to images of each actor (N = 96), giving a total of 81 x 96 = 7,776 cases. There were 1,833 instances where the two images of a given actor did not receive the same response by a given participant (i.e. recognition of only one of the two images), that is 23.6 % of inconsistent responses. Responses were thus consistent 76.4% of the time despite the fact that both typical and atypical appearances were presented. Importantly, rates of inconsistent responses were similar for stable and variable actors (i.e., 23.48 % and 23.66 %, respectively) and so that would not have skewed the results. Despite what could be seen as an imperfect consistency at the individual participant/actor level (i.e. inconsistent responses in close to one out of four instances), group-level patterns of performance in the two image conditions were remarkably consistent. This speaks to the relevance of our approach examining average performance on large samples in order to average out uncontrolled observer-related variations.

number of movies seen with the same five actors. The mean correlation they found, expressed in Z-scores, was of 0.288. In Z-scores, the abovementioned correlations ranged from 0.736 to 1.066, which is quite remarkable given that they are between the pooled performance of participants (with all the observer-related variations involved) and an independent but objective measure of media exposure, instead of between two subjective ratings provided by the same individuals. In other words, providing that large enough samples are used to account for individual preferences and skills of observers, StarMeter ranks may even prove more reliable and predictive of recognition performance than more typical checks of pre-experimental exposition that rely on various aspects of observers' memory.

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 even more productive. Importantly, our data 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 behavioural and ERP research showed that the eyes are strong identity cues, more reliable than other features like the mouth (Hsiao et al., 2022; 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 media 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.

The series of experiments presented here only offer indirect support for our new theoretical framework. Future research should endeavour to more directly assess the contribution of coarse and fine-grained information to facial representations, for example via systematic manipulations of spatial frequencies in test images of celebrities with various levels of relative stability in appearance.

Finally, at the neural level, 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 manipulate stability in appearance to examine how it modulates ERP responses.

Conclusions

We proposed a new account of face learning and familiarisation that takes stability in appearance into account and a series of three experiments as a proof of concept. 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 Tom Cruise, 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 less reliable representations, stability facilitates recognition in earlier stages of familiarisation. Tom Cruise'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 interplay with exposure. 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.

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. Mike. (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. Mike, 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 Mike, 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, A., & Tanaka, J. W. (2021). When a stranger becomes a friend: Measuring the neural correlates of real-world face familiarisation. *Visual Cognition*, *29*(10), 689–707. https://doi.org/10.1080/13506285.2021.2002993

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* (Vol. 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

Corpuz, R. L., & Oriet, C. (2022). Within-Person Variability Contributes to More Durable Learning of Faces. *Canadian Journal of Experimental Psychology*, *76*(4), 270–282. https://doi.org/10.1037/cep0000282

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. (2018). New insights on real-world human face recognition. *Retrieved from Osf.Io/8g6tm*. https://doi.org/10.17605/OSF.IO/8G6TM

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

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

Hsiao, J. H. wen, Liao, W., & Tso, R. V. Y. (2022). Impact of mask use on face recognition: an eye-tracking study. *Cognitive Research: Principles and Implications*, *7*(1). https://doi.org/10.1186/s41235-022-00382-w

Jansari, A., Miller, S., Pearce, L., Cobb, S., Sagiv, N., Williams, A. L., Tree, J. J., & Hanley, J. R. (2015). The man who mistook his neuropsychologist for a popstar: when configural processing fails in acquired prosopagnosia. *Frontiers in Human Neuroscience*, *9*(July), 1–16. https://doi.org/10.3389/fnhum.2015.00390

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, Rob, & 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

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

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

Morrison, D. J., & Schyns, P. G. (2001). Usage of spatial scales for the categorization of faces, objects, and scenes. *Psychonomic Bulletin and Review*, *8*(3), 454–469. https://doi.org/10.3758/BF03196180

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

Nakashima, T., Kaneko, K., Goto, Y., Abe, T., Mitsudo, T., Ogata, K., Makinouchi, A., & Tobimatsu, S. (2008). Early ERP components differentially extract facial features: evidence for spatial frequency-and-contrast detectors. *Neuroscience Research*, *62*(4), 225–235. https://doi.org/10.1016/j.neures.2008.08.009

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

Noyes, E., Davis, J. P., Petrov, N., Gray, K. L. H., & Ritchie, K. L. (2021). The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science*, *8*(3). https://doi.org/10.1098/rsos.201169

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

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(4), 406–417. https://doi.org/10.1037/0278-7393.3.4.406

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

Popova, T., & Wiese, H. (2023). How quickly do we learn new faces in everyday life? Neurophysiological evidence for face identity learning after a brief real-life encounter. *Cortex*, *159*, 205–216. https://doi.org/10.1016/j.cortex.2022.12.005

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

Ritchie, K. L., Kramer, R. S. S., & Burton, A. M. (2018). What makes a face photo a 'good likeness'? *Cognition*, *170*(April 2017), 1–8. https://doi.org/10.1016/j.cognition.2017.09.001

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

Sliwinska, M. W., Searle, L. R., Earl, M., O'Gorman, D., Pollicina, G., Burton, A. M., & Pitcher, D. (2022). Face learning via brief real-world social interactions includes changes in face-selective brain areas and hippocampus. *Perception*, *April*, 1–18. https://doi.org/10.1177/03010066221098728

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

Towler, J., Fisher, K., & Eimer, M. (2018). Holistic face perception is impaired in developmental prosopagnosia. *Cortex*, *108*, 112–126. https://doi.org/10.1016/j.cortex.2018.07.019

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

Yan, X., Goffaux, V. C. D. S., & Rossion, B. (2022). Coarse-to-Fine(r) Automatic Familiar Face Recognition in the Human Brain. *Cerebral Cortex*, *32*(8), 1560–1573. https://doi.org/10.1093/cercor/bhab238

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

Supplementary materials

Sampling rationale

Experiment 1. In face learning paradigms, variability yields strong effects on recognition. Power calculations from effect sizes in a study comparing high and low variability learning conditions (d = 1.336 in Experiment 1, d =1.427 in Experiment 2; Robins et al., 2018) yields a sample size of 10 and 9 to replicate the effect with .95 power. This effect might be reduced for faces that became familiar over longer periods and when only appearance varies. Further, since exposure of individual participants to each actor cannot be measured or guaranteed, data should be noisier than in face learning paradigms. Finally, the strength of an interaction between appearance and popularity was not possible to anticipate at the start. In Experiment 1, we thus decided to use a large sample size of 100 participants. In all experiments, participants received information about the study and provided consent before taking part.

Experiment 2. Power calculation based on Experiment 1 yielded 8 participants to replicate the interaction between popularity and appearance with .95 power. For cost-efficiency purposes and because the effect size of image condition was unknown, we used sequential analyses where data can be analysed at pre-defined incremental sample sizes, but with stricter significance thresholds (Lakens, 2014). We set to examine data in four increments—after 20, 40, 60 or 80 participants in each image condition. We used the Pocock boundary to establish the critical p value of .0182 for the three-way interaction of interest (image condition x popularity x appearance). Although the three-way interaction was already significant with 20 and 40 participants per group, we decided to collect data from 60 participants per group to rule out potential differences in face recognition skills or in

exposure to actors between participants in the two groups. Note that increasing the sample size while an effect of interest is already significant can never inflate Type 1 error since the only possibility is that the effect becomes non-significant (Lakens, 2018, personal communication).

Additional analyses

Experiment 2.

Sequential analysis – Step 1. 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'* obtained from the first 40 participants (20 per group). We found the expected main effect of image condition, $F(1,38) = 19.59$, $p < .001$, $\eta_p^2 = .340$, and of popularity, $F(1,38) = 189.206$, $p < .001$, $\eta_p^2 = .833$, as well as a significant main effect of appearance, $F(1,38) = 7.708$, $p = .008$, $\eta_p^2 = .169$. The three-way interaction between appearance, popularity, and image condition was significant, $F(1,38) = 9.246$, $p = .004$, $\eta_p^2 = .196$. The pattern of results was similar to that obtained at Step 3.

Sequential analysis – Step 2. We conducted the same ANOVA after collecting data from 80 participants (40 per group). We found the expected main effect of image condition, $F(1,78) = 42.81$, $p < .001$, $\eta_p^2 = .354$, and of popularity, $F(1,78) = 384.029$, $p < .001$, $\eta_p^2 = .831$, as well as a significant main effect of appearance, $F(1,78) = 11.964$, $p < .001$, $\eta_p^2 = .133$. The three-way interaction between appearance, popularity, and image condition was significant, $F(1,78) = 6.998$, $p = .010$, $\eta_p^2 = .082$. Again, the pattern of results was similar to the one obtained at Step 3.

Reaction times (final Step 3). 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 mean correct reaction time. There was a main effect of popularity, $F(1,118) = 29.314$, $p < .001$, $\eta_p^2 = .199$, and a main effect of image condition, $F(1,118) = 14.63$, $p < .001$, $\eta_p^2 = .110$, but no significant effect of appearance, $F(1,118) = 0.148$, $p = .701$, $\eta_p^2 = .001$. The three-way interaction of interest was not significant, $F(1,118) < 1$. There was also no interaction between image condition and appearance, $F(1,118) < 1$, between popularity and image condition, $F(1,118) = 1.337$, $p = .25$, $\eta_p^2 = .011$, nor between popularity and appearance, $F(1,118) = 1.435$, $p = .233$, $\eta_p^2 = .012$.

Experiment 3

Reaction times - 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 mean correct reaction time. There was a main effect of popularity, $F(1,121) = 55.373$, $p < .001$, $\eta_p^2 = .314$, and a main effect of image condition, $F(1,121) = 6.242$, $p = .014$, $\eta_p^2 = .049$, but no significant effect of appearance, $F(1,121) < 1$. The three-way interaction of interest was not significant, $F(1,121) < 1$. There was no interaction between image condition and appearance, $F(1,121) < 1$, or between popularity and image condition, $F(1,121) < 1$, but there was a significant interaction between popularity and appearance, $F(1,121) = 16.515$, $p < .001$, $\eta_p^2 = .12$. Mirroring the pattern found on sensitivity, in popular actors, responses to variable actors were faster (*Mean* = 968 ms, *SD* = 221) than to stable actors (*Mean* = 1026

ms, *SD* = 242), $t(122) = -4.01$, $p < .001$, $d = -0.361$. In less popular actors, responses to stable actors (*Mean* = 1075 ms, *SD* = 291) were faster than to variable actors (*Mean* = 1114 ms, *SD* = 332), $t(122) = 2.627$, $p = .010$, $d = 0.237$.

Reaction times - Experiment 3b. We conducted a three-way repeated measure ANOVA with appearance (variable, stable), popularity (popular, less popular), and image condition (typical, atypical headshot) as within-subject factors on mean correct reaction time. There was a main effect of popularity, $F(1,80) = 4.376$, $p = .040$, $\eta_p^2 = .052$, and a main effect of image condition, $F(1,80) = 7.421$, $p = .008$, $\eta_p^2 = .085$, but no significant effect of appearance, $F(1,80) < 1$. Here, there was a significant three-way interaction, $F(1,80) = 5.161$, $p = .026$, $\eta_p^2 = .061$. This was driven by slower responses to variable faces (*Mean* = 1133 ms, *SD* = 344) than to stable faces (*Mean* = 1030 ms, *SD* = 260) with typical images of less popular actors, $t(80) = -3.692$, $p < .001$, $d = -0.410$. There was also an interaction between image condition and appearance, $F(1,80) = 6.246$, $p = .014$, $\eta_p^2 = .072$. Finally, there was no significant interaction between popularity and image condition, $F(1,80) < 1$, nor between popularity and appearance, $F(1,80) < 1$.

Exploration of the impact of sex

The figures (S1 to S11) bellow illustrate discrimination performance as a function of popularity, appearance and sex and for which associated statistics are presented in **Table 1** of the main text. The figures display performance with intact images first, and then with other image conditions. All error bars represent 95% confidence intervals.

Figure S1. Experiment2 - Intact images

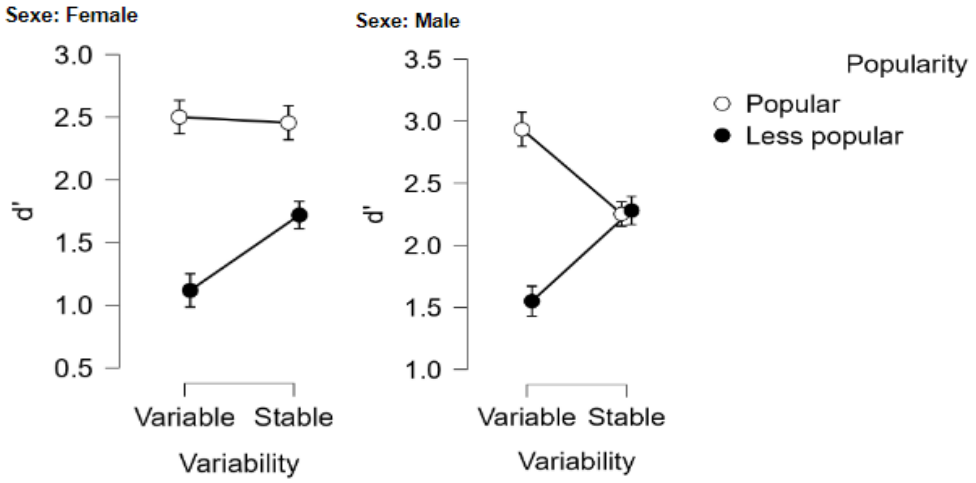

Figure S2. Experiment 3a - Intact images

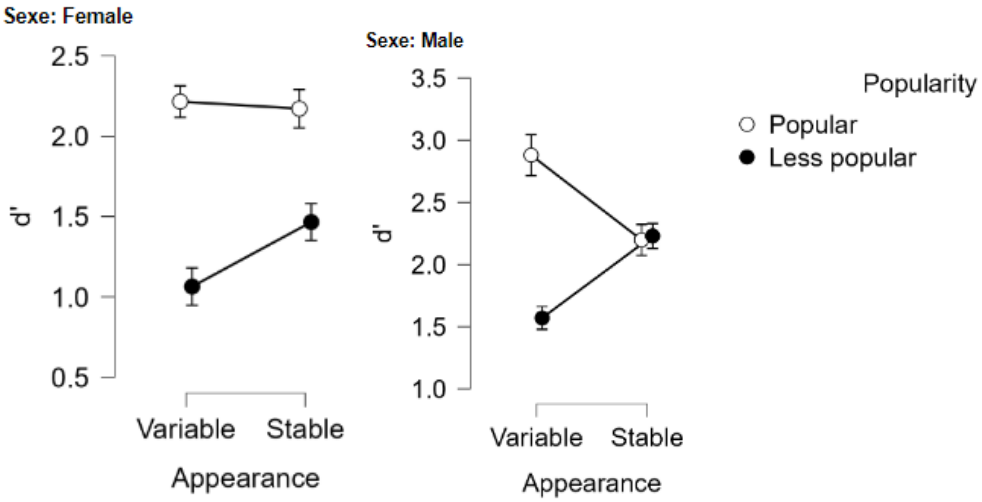

Figure S3. Experiment 3b - Intact images

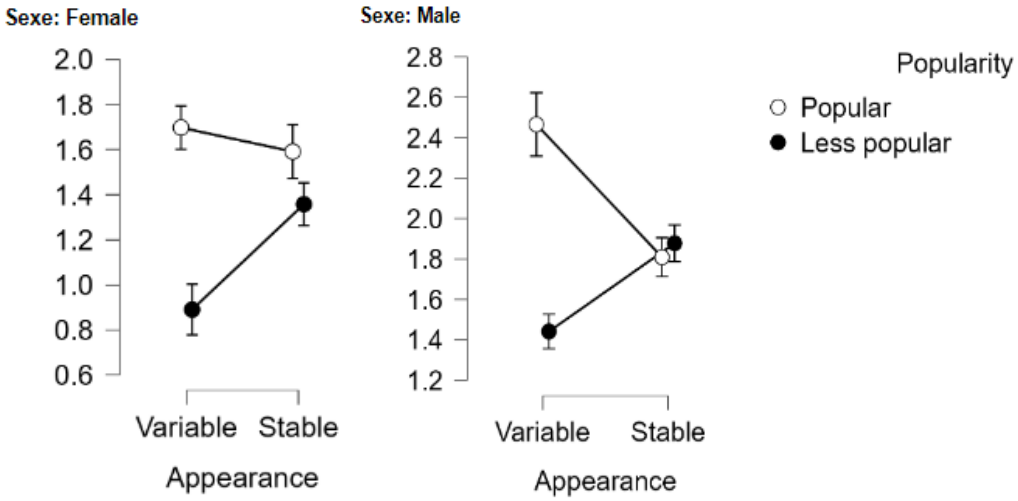

Figure S4. Experiment 1 - Cropped inner features

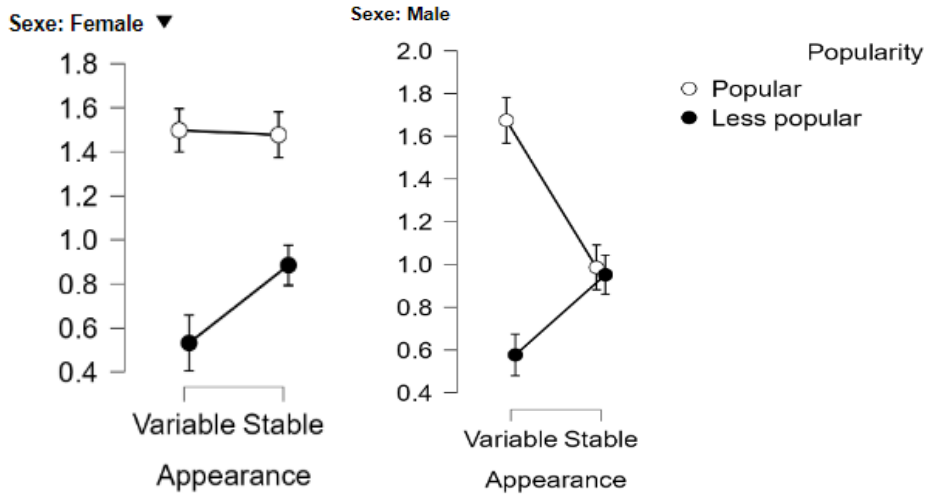

Figure S5. Experiment2 - Cropped inner features

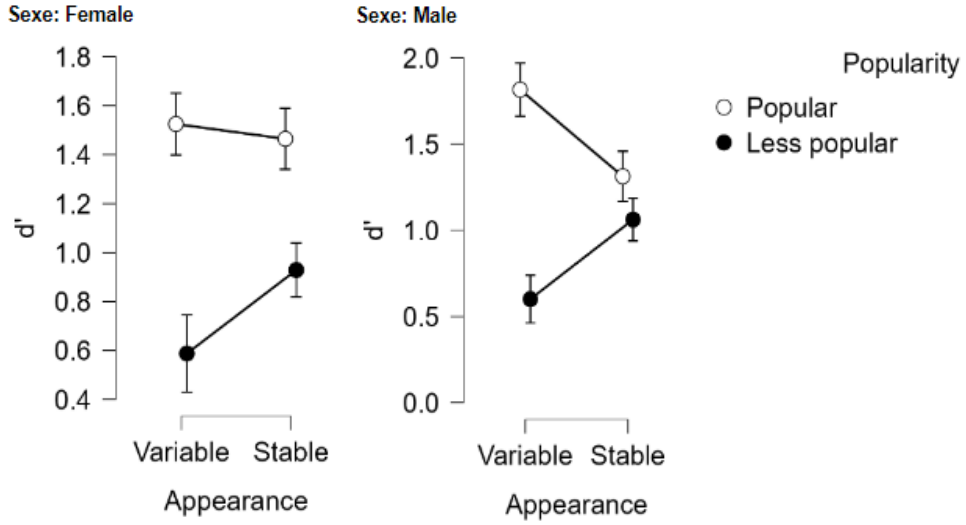

Figure S6. Experiment 3a - Atypical images

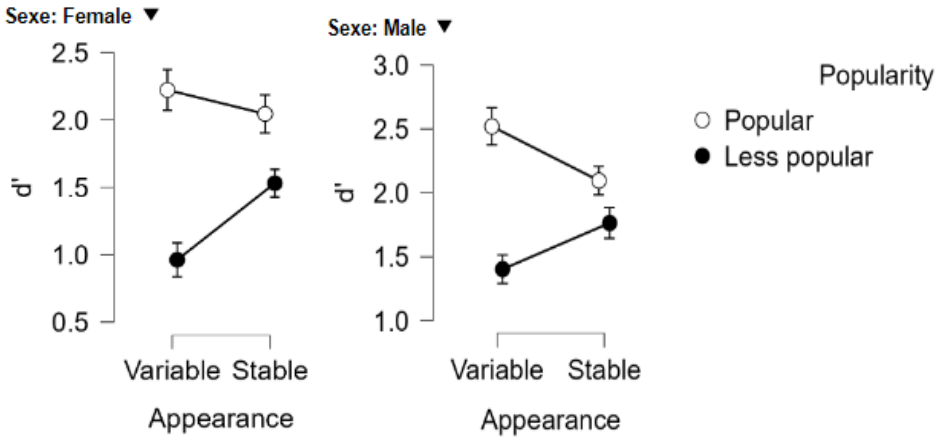

Figure S7. Experiment 3b - Atypical images

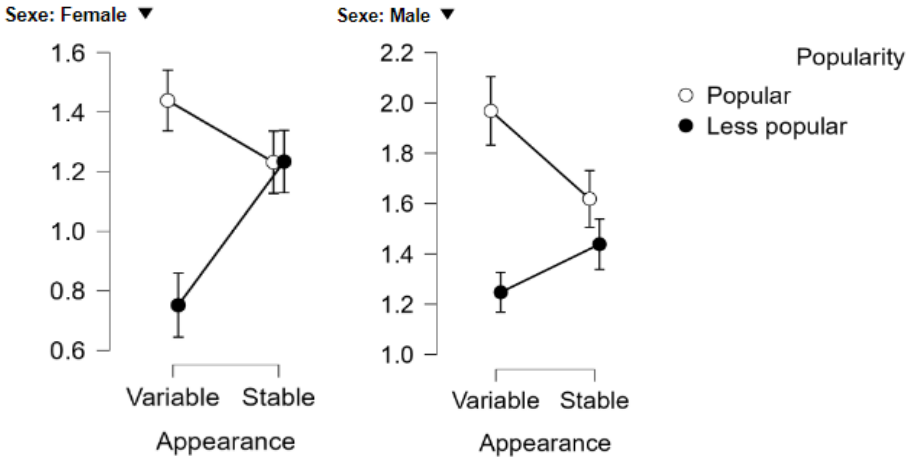

Table S1. List of actors used in the three experiments. Ages and StarMeter ranks were collected in August 2018.

	Sex	Age	StarMeter		Sex	Age	StarMeter
Popular – Variable				**Popular – Stable**			
Elisabeth Moss	F	35	32	Amanda Seyfried	F	32	21
Emilia Clarke	F	31	69	Amy Adams	F	43	29
Emily Blunt	F	35	83	Blake Lively	F	30	112
Emma Roberts	F	27	307	Cate Blanchett	F	49	154
Jennifer Lawrence	F	27	115	Christina Hendricks	F	43	67
Julia Roberts	F	50	347	Hailee Steinfield	F	21	185
Keira Knightley	F	33	217	Kristen Bell	F	37	204
Kristen Stewart	F	28	194	Margot Robbie	F	28	39
Lily James	F	29	1	Megan Fox	F	32	282
Meryl Streep	F	69	89	Natalie Portman	F	37	140
Scarlett Johansson	F	33	124	Rachel Weisz	F	48	391
Zoey Deutch	F	23	91	Sandra Bullock	F	53	87
Brad Pitt	M	54	93	Adam Sandler	M	51	174
Bradley Cooper	M	43	385	Chris Pratt	M	39	108
Chris Pine	M	37	215	Harrison Ford	M	76	364
Christian Bale	M	44	222	Joseph Gordon-Levitt	M	37	407
Jake Gyllenhaal	M	37	201	Matt Damon	M	47	241
Jared Leto	M	46	336	Max Minghella	M	32	179
Joaquin Phoenix	M	43	123	Taron Egerton	M	28	99
Johnny Depp	M	55	78	Timothée Chalamet	M	22	141
Matthew McConaughey	M	48	269	Tom Cruise	M	56	5
Ryan Reynolds	M	41	90	Tom Hanks	M	62	117
Steve Carell	M	55	190	Tye Sheridan	M	21	157
Zac Efron	M	30	220	Vin Diesel	M	50	340
Mean		39.7	170.5			40.6	168.5
SD		11.3	104.5			13.6	116.4
Less popular – Variable				**Less popular – Stable**			
Alison Sudol	F	33	1108	Andie MacDowell	F	60	1384
Alyssa Milano	F	45	1006	Ashley Benson	F	28	1078
Brittany Snow	F	32	1255	Crystal Reed	F	33	1099
Clemence Poesy	F	35	1085	Danielle Campbell	F	23	1115
Embeth Davidtz	F	52	1227	Drea de Matteo	F	46	1231
Jena Malone	F	33	1009	Gwyneth Paltrow	F	45	1097
Kate Walsh	F	50	1300	Hilary Swank	F	44	1275
Natalia Tena	F	33	1452	Katherine McNamara	F	22	1015
Scout Taylor-Compton	F	29	1086	Lisa Kudrow	F	54	1264
Sigourney Weaver	F	68	1082	Maggie Gyllenhaal	F	40	1267
Valorie Curry	F	32	1090	Melissa Rauch	F	38	1178
Zoe McLellan	F	43	1468	Tilda Swinton	F	57	1068

Anson Mount	M	45	1055	Billy Magnussen	M	33	1175
Ben Barnes	M	36	1060	Clive Owen	M	53	1205
Boyd Holbrook	M	36	1100	David Schwimmer	M	51	1310
David Thewlis	M	55	1224	Jamie Bell	M	32	1207
Dermot Mulroney	M	54	1480	Jared Padalecki	M	35	1236
Elijah Wood	M	37	1289	Liam Hemsworth	M	28	1470
Eric Bana	M	49	1463	Owen Wilson	M	49	1146
Garrett Hedlund	M	33	1405	Patrick Dempsey	M	52	1242
Jerry O'Connell	M	44	1045	Richard Gere	M	68	1025
Jesse Eisenberg	M	34	1385	Rupert Grint	M	29	1361
Matthew Lillard	M	48	1076	Stanley Tucci	M	57	1230
Matthew Perry	M	48	1246	Vince Vaughn	M	48	1112
Mean		41.8	1208.2			42.7	1199.6
SD		9.8	162.4			12.6	114.4

Table S2. Results of Experiment 1. Mean proportion of "familiar" responses—hits for the four types of actors and false alarms for unfamiliar faces—and corresponding reaction times (RT) in milliseconds. Standard deviation are in italics.

Popularity	**Popular**		**Less popular**		**Unfamiliar**
Appearance	Variable	Stable	Variable	Stable	
'Familiar' responses	.676	.579	.355	.476	.199
	.199	*.203*	*.161*	*.169*	*.145*
Reaction times	1108	1172	1291	1201	1152
	263	*303*	*312*	*310*	*306*

Table S3. Results of Experiment 2. Mean proportion of "familiar" responses—hits for the four types of actors and false alarms for unfamiliar faces—and corresponding reaction times (RT). Standard deviation are in italics.

	Popularity	Popular		Less popular		Unfamiliar
	Appearance	Variable	Stable	Variable	Stable	
	Image condition					
'Familiar' responses	Inner features	.691	.610	.361	.483	.189
		.215	*.210*	*.194*	*.192*	*.140*
	Headshot	.816	.724	.397	.617	.071
		.113	*.131*	*.149*	*.142*	*.070*
RTs (ms)	Inner features	1121	1173	1312	1305	1119
		486	*323*	*468*	*586*	*347*
	Headshot	957	978.5	1087	1059	986
		221	*204*	*237.5*	*266*	*188*

Table S4. Results of Experiment 3. Mean proportion of "familiar" responses—hits for the four types of actors and false alarms for unfamiliar faces—and corresponding reaction times (RT). Standard deviation are in italics.

	Popularity	Popular		Less popular		Unfamiliar
	Appearance	Variable	Stable	Variable	Stable	
Experiment 3a						
	Image condition (between-subject)					
'Familiar'	Atypical	.752	.678	.398	.55	.104
responses		*.177*	*.179*	*.175*	*.189*	*.105*
	Typical	.786	.702	.419	.6	.089
		.138	*.145*	*.172*	*.175*	*.092*
RTs (ms)	Atypical	1030	1073	1196	1123	1085
		262	*256*	*405*	*311*	*257*
	Typical	907	980	1093	1028	1022
		150	*219*	*231*	*263*	*256*
Experiment 3b						
	Image condition (within-subject)					
'Familiar'	Atypical	.597	.523	.378	.491	
responses		*.212*	*.18*	*.166*	*.21*	.138
	Typical	.695	.607	*.434*	.577	*.162*
		.214	*.188*	*.184*	*.185*	
RTs (ms)	Atypical	1099	1084	1092	1173	
		295	*241*	*249*	*644*	1007
	Typical	1036	1008	1133	1030	*263*
		290	*238*	*344*	*260*	