**Article:**
Laurence, Sarah, BURTON, MIKE, Düring, Camilla et al. (2026) Longstanding mental representations of familiar faces. Cognition. 106555. ISSN: 0010-0277

https://doi.org/10.1016/j.cognition.2026.106555

Cognition 274 (2026) 106555

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

# Longstanding mental representations of familiar faces

Sarah Laurence [a,**], A. Mike Burton [b,c], Camilla Düring [d], Jennifer Pink [a], Lucy Wilson [e], Mila Mileva [e,*]

[a] School of Psychology & Counselling, The Open University, Milton Keynes, UK
[b] Department of Psychology, University of York, UK
[c] Faculty of Society and Design, Bond University, Australia
[d] University Hospitals Sussex NHS Foundation Trust, UK
[e] School of Psychology, University of Plymouth, UK

ARTICLE INFO

*Keywords:*
Face recognition
Face learning
Familiar faces
Face age

ABSTRACT

As we recognise people we know over many years, their faces can change, sometimes profoundly, and yet we continue to recognise them with ease. How do we update our representations over time? We present four pre-registered experiments to examine this. In Experiment 1, using likeness ratings and speeded name verification, fans of a long-running TV soap opera demonstrated that their representations of the characters' faces were weighted towards their most recent encounters – when the characters were oldest. While we initially hypothesised that this was due to recency, Experiment 2 showed this not to be the case. When new participants were taught these characters either in chronological or reverse-chronological order they all demonstrated representations weighted towards the characters at their oldest ages, regardless of the order in which they had encountered them. We ruled out potentially artefactual explanations using statistical analysis of the images themselves and, in Experiment 3, restricted learning sets. A further, final experiment showed that our results are unlikely to be fully explained by perceived distinctiveness of the stimuli. We conclude that the processes involved in developing representations for familiar people are more sophisticated than previously thought, incorporating real-world constraints, including natural chronology.

## 1. Introduction

Each encounter with a familiar person takes place at a particular point in time. Someone who has encountered the people shown in Fig. 1 frequently over the last 20 years will have recognised their identity at each encounter. They will also have seen the face gradually change over the 20-year period. While we already know quite a lot about the general changes in facial structure and/or skin texture and pigmentation that occur with aging (Burt & Perrett, 1995; Geng, Zhou and Smith-Miles, 2007; Pittenger & Shaw, 1975), we know relatively little about how these temporal changes might be incorporated into the stored representations of familiar individuals, whose aging will be particular to them. Are memory representations of familiar faces optimised to enable us to recognise someone's most recent/current appearance? Are they optimised to recognise an overall average of all the encounters to which someone has been exposed to? Here, we explore how familiar faces known for many years are represented.

### 1.1. Understanding familiarity: The role of variability

Research on face recognition has established that, compared to unfamiliar face recognition, people are more accurate at recognising images of familiar faces, even in highly variable, poor quality or distorted images (Gilad-Gutnick, Harmatz, Tsourides, Yovel and Sinha, 2018; Hole, George, Eaves and Rasek, 2002; Sandford & Rego, 2019). For example, Jenkins, White, Van Montfort and Burton (2011), asked participants to sort a stack of 40 face photographs (20 images of each of two Dutch celebrities) by identity such that all the photographs of the same person were grouped together. Most participants in the Netherlands, where the celebrities were famous, performed this task without error and sorted the images into two groups. Unfamiliar participants in the UK, however, tended to perform poorly and sorted the photographs into

* Corresponding author at: School of Psychology, Faculty of Health: Medicine, Dentistry and Human Sciences, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK.
** Corresponding author at: School of Psychology & Counselling, The Open University, Milton Keynes MK7 6AA, UK.
*E-mail addresses:* sarah.laurence@open.ac.uk (S. Laurence), mila.mileva@plymouth.ac.uk (M. Mileva).

https://doi.org/10.1016/j.cognition.2026.106555
Received 7 October 2025; Received in revised form 8 April 2026; Accepted 10 April 2026
Available online 21 April 2026

between 3 and 16 groups (median = 7.5). This study neatly demonstrates the stark difference between familiar and unfamiliar face recognition: familiarity allows perceivers to cohere together superficially different images of the same person, whereas unfamiliar viewers often mistake image differences with identity differences. This observation is consistent with studies using many different experimental approaches (e.g. Bruce, 1982; Hancock, Bruce and Burton, 2000; Longmore, Liu and Young, 2008).

It has been argued that stored visual representations of faces depend on the statistical properties of exposure (e.g., Burton, Kramer, Ritchie, & Jenkins, 2016). The nature of facial familiarity (i.e., encountering a person multiple times) means that the visual system has been exposed to variation in appearance during and across different encounters. This allows the observer to create a mental representation that captures stable identity information, while discarding non-informative changes in appearance or the environment (Bruce, 1994). The mental representation of an unfamiliar face that has only been encountered once will be based on the statistics of that single encounter, informative or otherwise. Support for this idea comes from studies of face learning, where exposure to many different images of the same person (i.e., greater within-person variability; Jenkins et al., 2011) facilitates learning relative to less variable learning stimuli (similar to the variability one might be exposed to in a single encounter). For example, Ritchie and Burton (2017) familiarised participants with multiple images of a face that were either captured from a single recorded event (low variability) or across multiple different events (high variability). The ability to recognise novel images of the learned identities was better after exposure to high variability.

Computational models of face recognition have also shown a familiarity advantage when the training set is comprised of variable images of faces. For example, by applying Principal Components Analysis (PCA) to multiple face images of an identity and then applying Linear Discriminant Analysis (LDA), Kramer, Young and Burton (2018) were able to simulate familiarity effects; the more images of an identity included in the training set, the better the model subsequently performed at recognising novel images of a particular identity. Similar results have also been found for Deep Convolutional Neural Network (DCNN) models of face recognition - increasing the number of highly variable ambient images that a DCNN was trained on provided an additional benefit for the identification of novel faces (Blauch, Behrmann and Plaut, 2021). These studies clearly acknowledge that facial appearance varies. However, while the studies mentioned previously do incorporate some minimal age differences in the images, the continual changes to facial appearance due to aging and how they are incorporated within our mental identity representations have rarely been studied explicitly.

### 1.2. Faces with longstanding representations

Representations of familiar faces are durable, even if a face has not been encountered for many years - Bruck, Cavanagh and Ceci (1991) found that people could match yearbook photos of their former classmates to recent photos taken 25 years later at a rate of 49%, which was better than matching rates for control participants who were unfamiliar with these people (33%). The ability to recognise a face many years later, while imperfect, is impressive given that these rates refer only to people who had not seen their classmates for at least 17 years. This is challenging to understand because faces change continually; people need not only to retain a face in their memory for 17 years, but also to recognise a face that looks very different compared to 17 years ago. Each time a face is encountered it is likely that the viewer will experience some variability (e.g., changes in expression or lighting), but it is only across encounters that we start to see more substantial changes in appearance. A face that is encountered frequently will show variation from day to day (e.g., changes in make-up/facial hair/fatigue-related appearance changes) and more substantial changes from year to year (e.g., structural and textural changes, e.g., Porcheron, Mauger and Russell, 2013; Tiddeman, Burt and Perrett, 2001).

In recent research, there has been a renewed interest in how the

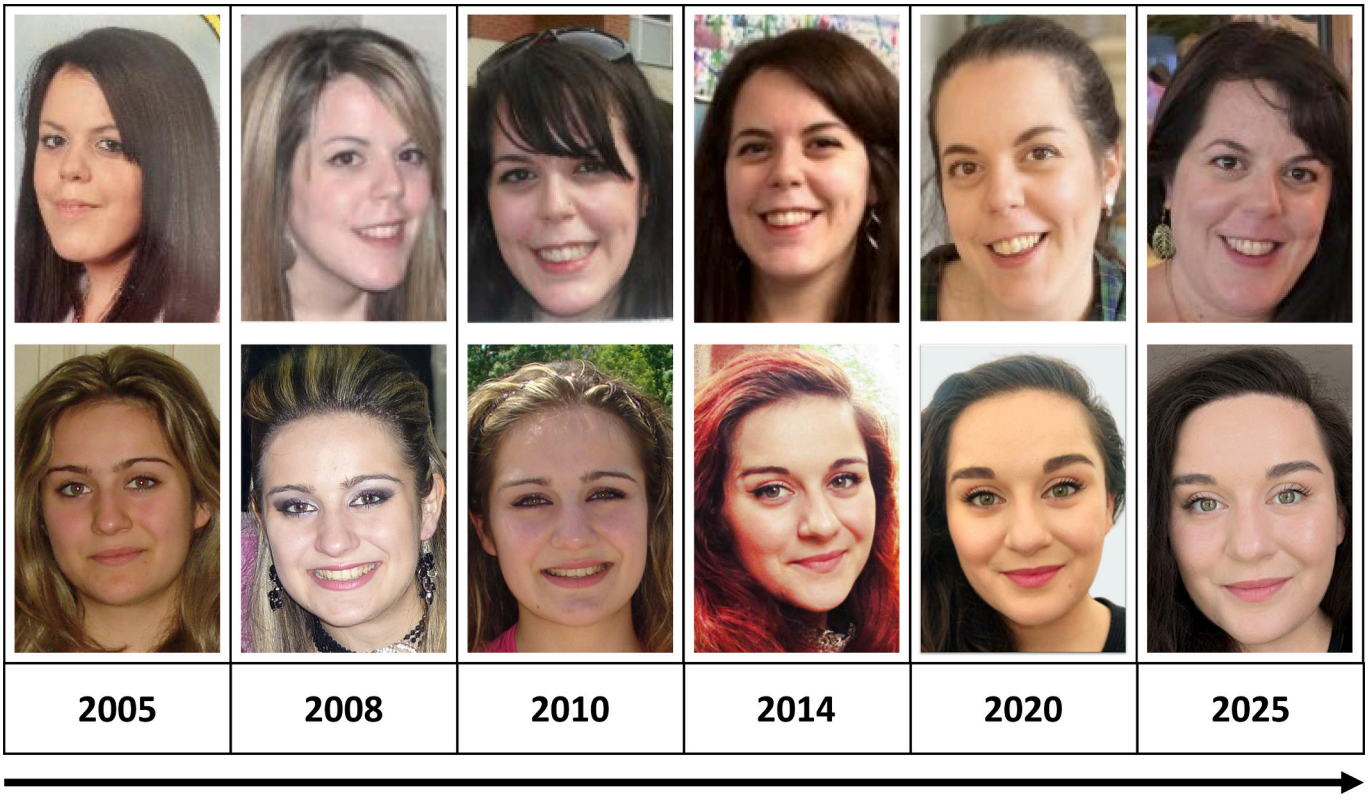


**Fig. 1.** Face images of two identities spanning a 20-year period.

human face recognition system might cope with differences due to aging. Mileva, Young, Jenkins and Burton (2020), for example, use a computational approach where two models are trained and tested on images taken over a 60-year period. One model was based on the statistical properties of images using PCA, akin to unfamiliar face recognition, whereas the other model incorporated additional LDA clustering based on the identity depicted in the images, allowing for abstraction similar to that in familiar observers. In one simulation, they ask whether after training the model on variable images from a constrained time period, it is possible to generalise to new time period. Here, images of the same identity from three different time periods were represented as three separate identities (e.g., 1960s Paul McCartney, 1980s Paul McCartney and 2000s Paul McCartney) and the types of errors from the two models were recorded. The familiar recognition model (PCA + LDA) substantially outperformed the unfamiliar (PCA only) model, with most errors being the classification of an image as belonging to the correct identity from a different time period (e.g., classifying a 1960s image of Paul McCartney as belonging to the 1980s Paul McCartney). There was a much higher proportion of real misidentification errors in the unfamiliar model, where an image showing one identity (1960s Paul McCartney) was classified as a completely different identity from the same or a different time period (1960s Frankie Avalon or 1980s Frankie Avalon). Subsequent simulations showed that the familiar model was able to generalise across time even when trained on images from a specific time period and that there were some idiosyncratic differences in how individual faces aged over time: some faces changed a lot over a 60-year period whereas others changed less over the same period. Altogether, this suggests that familiar faces with longstanding representations do not necessarily require multiple representations for each time point (e.g., a representation of Paul McCartney in the Beatles and a representation for today is not needed). This is further supported by behavioural findings, for example Laurence, Baker, Proietti and Mondloch (2022) who found priming and adaptation effects for familiar celebrity faces, such as Paul McCartney, were not affected by face age.

While previous research suggests multiple representations of faces is not necessary to account for substantial age-related changes in appearance, the nature of the stimuli used in previous research might limit these conclusions. Celebrity faces, such as those used by Laurence et al. (2022), are often used as a proxy for real world familiar face recognition because they are familiar to many people and their images are easy to access. However, the nature of exposure to celebrities, such as Paul McCartney and Susan Sarandon, is likely to be different to personally familiar faces. Whereas exposure to a personally familiar face tends to follow a linear timeline, exposure to famous movie stars or singers is more unpredictable. For example, the next time you see Susan Sarandon she might be younger (e.g., if you decide to watch one of her classic movies such as *Thelma and Louise* or *The Rocky Horror Picture Show)* or older (e.g., if you might decide to watch her 2019 movie, *Blackbird*).

Interestingly, research has found an effect of face age on the perception of personally familiar faces, where a preference is observed for images showing how this familiar identity looks currently, at their oldest age. For example, Kurth, Moyse, Bahri, Salmon and Bastin (2015) found that older adults were faster to recognise recently taken photographs of personally familiar faces than photographs from the past. Similarly, Schneider and Carbon (2021) asked family members to rate images of a longstanding personally familiar face for 'prototypicality' and found that recently taken face photographs were rated as more prototypical than photographs taken many years ago. They also morphed together images from three different time periods to create three 'episodic prototypes' (youngster prototype, middle-age prototype, recent prototype) and one 'exhaustive prototype' created by morphing together all 20 images of each identity. As with the photographs, the 'recent prototype' was rated as the most prototypical. Based on these data, they propose an Episodic Prototype Model in which there are distinct prototypes for different episodes of life. An additional component of their model is the formation of new prototypes when appearance changes to such an extent that existing prototypes are no longer useful. While we agree that existing theories of face recognition do not account for temporal changes in facial appearance, there is limited evidence for distinct prototypes being required to support face recognition across decades (e.g., Laurence et al., 2022; Mileva et al., 2020).

### 1.3. The present research

Recent theories propose that each known face has its own representational identity-specific 'face space' (e.g., Kramer et al., 2018; Noyes et al., 2021). However, little is known about how variability is distributed within identity-specific face space for faces with longstanding representations. For example, a colleague, friend or family member who has been encountered frequently and consistently over a 20-year period will look different 20 years ago compared to the present day. Each encounter that takes place over that period will be recognised and update the stored mental representation of a face. Over time the representation will therefore be continuously updated. Here, we used face averages to explore how identity-specific face space accounts for long term representational updating. Previous research has suggested that face averages might form a useful model of our representations of familiar people, in the sense of morphing together multiple encounters into a single "central tendency" (Burton, Jenkins, Hancock and White, 2005). However, these models typically weigh all encounters equally. As an alternative, we might predict that representations could be updated giving greater weight to recent encounters – or perhaps particularly salient encounters, such as a first meeting, might be weighted more heavily.

To explore the possibility of more representational weight being given to recent appearance in identity-specific face space, we created face averages for several characters from the long-running UK soap opera *Coronation Street.* It was first aired in the 1960s with episodes broadcast multiple times a week and viewed by millions (see https://coronationstreet.fandom.com/wiki/Coronation_Street). The soap follows the lives of people who live in a fictional town in Greater Manchester, with many characters having been featured for many years (see https://coronationstreet.fandom.com/wiki/Longest_running_characters). Fans of the show often tune in every week and follow the storylines, and there are anecdotal reports of people having watched the show continuously for many years ('It's the closest thing my family has to religion': Guardian readers on Coronation Street, see https://www.theguardian.com/tv-and-radio/2024/apr/12/its-the-closest-thing-my-family-has-to-religion-guardian-readers-on-coronation-street). The nature of the viewing habits of fans of the show mean that it is a good candidate for studying face representations as they develop over time.

In order to test how well familiar or familiarised perceivers generalise their representations across time, we collected images of longstanding Coronation Street characters spanning the last 20 years (2003−2023). We then used these images to create three types of identity averages by morphing different image sets together – 1) a 2000s-weighted average, where images taken approximately 20 years ago were weighted more heavily, 2) a 2020s-weighted average, where images taken more recently were weighted more heavily, and 3) an exhaustive average where all images of a certain identity were morphed together, with equal weighting for each image.

Across four pre-registered experiments, we explored whether representations of familiar faces are weighted towards recent encounters. In Experiment 1, we examined whether people who have watched Coronation Street continuously for the last 20 years have representations of the characters that are weighted towards recent encounters. Participants rated images and averages (weighted and non-weighted) for likeness and completed a speeded name-verification task. We predicted that if the central tendency of identity-specific face space is closer to recent facial appearance, then people would rate the 2020s-weighted averages and the 2020s instances as a better likeness and would recognise them faster when preceded by a name.

We further explored the role of recency in Experiment 2 where participants, unfamiliar with Coronation Street, were familiarised with characters from the show. Crucially, some participants learned these characters in chronological order, while others learned them in reverse-chronological order, which allowed us to disentangle the effect of recency (i.e., whether observers would prioritise their most recent encounters with the characters) and the effect of aging (i.e., whether observers would prioritise encounters where characters appeared at their oldest age, regardless of exposure recency).

Experiment 3 and 4 were conducted to examine alternative explanations for the results of Experiments 1 and 2. In Experiment 3, we examined whether 2000s-weighted averages are rated as a better likeness and recognised faster for people who have only encountered someone when they were younger. Finally, in Experiment 4, we explored whether the 2020s-weighted averages and 2020s instances were perceived as more distinctive than 2000s-weighted averages and instances.

## 2. Experiment 1

In order to determine whether familiar face representations are weighted towards recent encounters, we collected 21 images for each of 15 Coronation Street characters. These images spanned a 20-year period between 2003 and 2023 and included 7 images showing how the character looked earlier in this period (between 2003 and 2009), 7 images showing how the character looked in the middle of this period (between 2010 and 2016) and 7 images showing the character's most recent appearance at the end of the time period (between 2017 and 2023). These images were used to create three face averages – a 2000s-weighted average that captured the variability in all 21 images but prioritised the way the character looked earlier on in the time period, a 2020s-weighted average that captured the variability in all 21 images, but prioritised the way the character looked more recently within the time period and a non-weighted exhaustive average that incorporated the variability in all images without any prioritisation. We then asked highly familiar participants to rate some image instances and the three average images for likeness and to complete a speeded name-verification task with those same images. We pre-registered the following hypotheses: 1) we predicted that 2020s-weighted averages will be recognised more quickly and will represent a better likeness than non-weighted exhaustive averages and 2) we predicted that there will be a difference between likeness ratings and reaction times for specific image instances and average images. The direction of this effect is difficult to determine given the conflicting findings in the existing research on this topic (e.g., Burton et al., 2005; Ritchie, Kramer, Mileva, Sandford, & Burton, 2021; White, Burton, Jenkins and Kemp, 2014).

### 2.1. Method

#### 2.1.1. Participants

Participants were recruited via the following: 1) adverts posted on Coronation Street fan Facebook pages, 2) a Facebook advertising campaign, 3) an advert posted to staff at The Open University ("Ask the Community"), and 4) The Collaboration Laboratory Sona Systems site at The Open University. Note that our recruitment strategy differs from the pre-registration (adverts posted on Coronation Street fan Facebook pages and adverts in local newspapers) due to difficulties we encountered in recruiting enough participants from social media. We therefore needed to diversify our recruitment strategy. Participants were only eligible to take part if they met the following criteria: 1) they had watched Coronation Street continually for the last 20 years; 2) they watched Coronation Street at least once a week on average; 3) no cognitive impairments and normal (or corrected-to-normal) vision; 4) over 30 years of age; 5) could take part on a laptop or PC. A total of 93 participants met these criteria and were included in the final analyses ($M_{age}$ = 49.3 years; $SD_{age}$ = 11.6 years; 83 female, 8 male, 2 undisclosed). An a priori power analysis for a 2 × 3 within-subjects design using the Superpower package in R (Lakens & Caldwell, 2021) revealed a required sample of 90 participants to detect an effect size f of 0.204 for the interaction (equivalent to a partial eta squared of 0.04) with 80% power. All participants had the opportunity to enter a prize draw to win 20 x £25 Amazon vouchers. Ethical approval was granted by the Human Research Ethics Committee at The Open University, UK.

#### 2.1.2. Materials

We identified 15 Coronation Street characters who had appeared on the show continually for the previous 20 years (between 2003 and 2023; see the Appendix for a list of names). For each of the characters, we collected 21 face images taken between 2003 and 2023. Face images were the highest quality screenshots we were able to obtain showing a frontal view of the face taken from episodes of Coronation Street that could be dated. For each character, we aimed to obtain face images for each year between 2003 and 2023. In practice, however, this was not always possible (e.g., we were unable to obtain an image of Eileen Grimshaw in the year 2004 as there was not a high-quality clip that could be dated available online). All screenshots were cropped to show just the head. Images were resized to 380 × 570 pixels and were shown in colour using a lossless image format (bitmap).

In order to use these images to create identity averages, they were first normalised by aligning 82 fiducial points across each face (e.g., corners of the mouth, tip of the nose, etc.). This was performed using a semi-automatic process where five of these points were aligned manually and the remaining points were estimated automatically (see Kramer, Young, et al., 2017 for additional details). After normalising, the InterFace software (Kramer, Jenkins, & Burton, 2017) was used to create the three types of averages by morphing different sets of images together. For each average image, the normalised faces were morphed to the average shape of the specific identity. Exhaustive averages were created by morphing all 21 images together with each image having an equal weighting (i.e., each image was included once within the average image). 2000s- and 2020s-weighted averages were created from the same 21 images, but some images were included in the average more than once to increase the weighting applied to these images. Both types of averages were created from 78 images, with five specific images included multiple times depending on the type of average. For the 2000s-weighted averages, the first five earliest images of each character (where they appear the youngest) were repeated 32, 16, 8, 4, and 2 times respectively (e.g., the earliest image was repeated 32 times, the next earliest image was repeated 16 times and so on). For the 2020s-weighted averages, the last five images taken the most recently (those showing each character at their oldest age) were repeated 32, 16, 8, 4, and 2 times respectively (e.g., the most recent image was repeated 32 times, the next most recent image was repeated 16 times and so on). For both types of averages, the remaining 16 images were included only once. See Fig. 2 for a visual example of the weighting procedure as well as the resulting images.

In addition to these averages, three instance images were also selected for each identity – one taken at the start of the specified time period (2003–2005), one from the mid-point of the time period (2012–2015) and one from the end of the time period (2022−2023). The instance images were cropped to remove the background, so their appearance matched that of the averages.

#### 2.1.3. Procedure

The experiment was built and administered using Gorilla Experiment Builder (https://app.gorilla.sc/). Participants initially completed a Coronation Street face recognition test. The test comprised of four rows of faces, each containing four images: one face in each row depicted a Coronation Street character and three depicted characters from other UK soaps. Participants were asked to determine which of the four faces in each row belonged to Coronation Street characters. Participants were only able to proceed to the rest of the experiment if they were able to

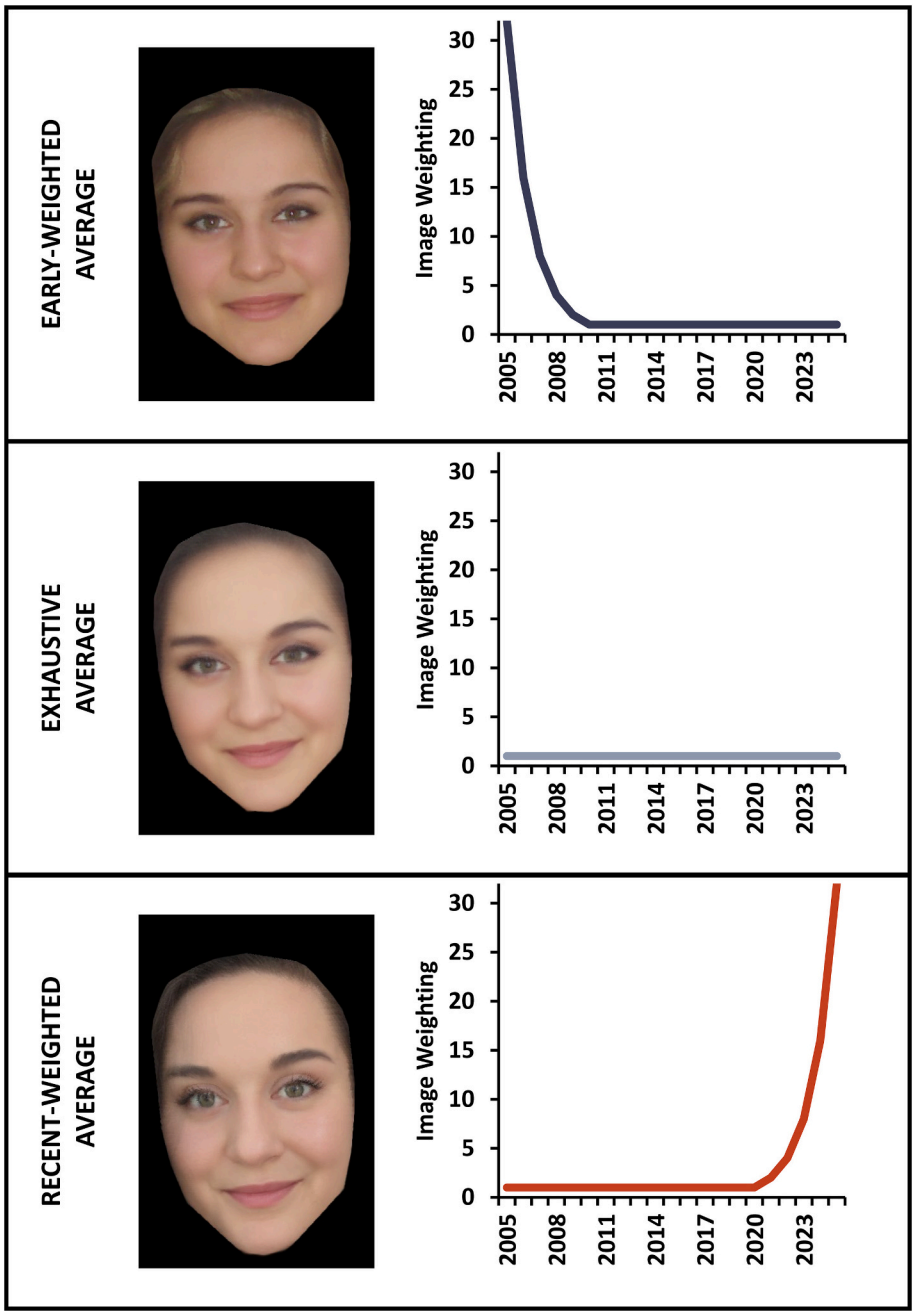


**Fig. 2.** Weightings used to create the three types of averages used in the study.

correctly identify all four Coronation Street characters. They were given two attempts and no feedback on which answers were correct/incorrect. Participants who passed the recognition test then completed a questionnaire in which they answered questions about their Coronation Street viewing habits to ensure they had watched Coronation Street regularly for the last 20 years. They also provided some demographic information. Participants who did not meet the inclusion criteria specified in the Participants section were unable to progress any further in the experiment.

Participants then completed two tasks, a likeness rating task and a speeded name verification task, similar to that used by Ritchie et al. (2018). The order in which participants completed the tasks was counterbalanced across participants (half of the participants completed the likeness rating task first and then the speeded name verification task and the other half completed the same tasks in the reverse order). For the likeness task, participants viewed three averages (2000s-weighted, 2020s-weighted and exhaustive) and three instances (2000s instance, 2010s instance, 2020s instance) for each of the 15 Coronation Street characters in a random order and were asked to indicate how good a likeness of that character each image was. The likeness scale that participants used to make their ratings ranged from 1 to 7 (1 = very bad likeness, 7 = very good likeness).

For the speeded name verification task, the name of a Coronation Street character was presented on the screen for 1500 ms, immediately followed by an image of a same-sex face that remained on the screen until the participant made a response. On half of the trials, the identity of the face matched the name. On the other half of the trials, the face belonged to a different Coronation Street character. Participants were asked to indicate as quickly and as accurately as possible via a keyboard response whether the face image showed the same person as the name or not (e.g., d = match, j = mismatch). For each character, there were 6 match trials in which the name and the face belonged to the same

person. There were also 6 mismatch trials for each identity (e.g., the name Gail Platt followed by an image of Audrey Roberts). The images shown on match and mismatch trials were the same images as in the likeness task. Each image was shown once in a match trial and once in a mismatch trial, resulting in a total of 180 trials which were presented in a random order for each participant. Each trial was preceded by a fixation cross for 500 ms.

Participants completed 4 practice trials (using the names/images of Homer Simpson and King Charles III) before completing the speeded name verification task proper.

#### 2.1.4. Transparency and openness

All experiments reported in the manuscript were pre-registered on the OSF (Experiment 1: https://osf.io/7ghsm; Experiment 2: https://osf.io/t6uqg; Experiment 3: https://osf.io/dz3p2; Experiment 4: https://osf.io/9mxjy). Data from all four experiments can be accessed from the project's OSF page (https://osf.io/e8vd4). The face average images used in all experiments can also be found on the same project page, however, instance images and videos used for training are protected by copyright and therefore cannot be shared.

### 2.2. Results

#### 2.2.1. Likeness ratings

Data from 2 participants were excluded as they provided the same rating for all images. Fig. 3 shows mean ratings across conditions. A 2 × 3 within-subjects ANOVA with factors: image type (instance vs average) and time point (2000s/2000s-weighted vs 2010s/exhaustive vs 2020s/2020s-weighted) showed a significant main effect of image type ($F(1, 92) = 68.97$, $p < .001$, $\eta^2_G = .43$), with instances rated as a better likeness than averages. The main effect of time point was also significant ($F(2, 184) = 126.90$, $p < .001$, $\eta^2_G = .58$) and so was the interaction between image type and time point ($F(2, 184) = 47.91$, $p < .001$, $\eta^2_G = .34$). The significant main effects and interaction were followed up with Tukey-corrected post hoc tests. Instance images from the 2020s were perceived as being a better likeness than 2010s images ($t(92) = 6.17$, $p < .001$) and 2010s images were perceived as being a better likeness than 2000s images ($t(92) = 11.78$, $p < .001$). Average images followed the same pattern, however the differences between the three time points were less pronounced than those observed with the instances. 2020s-weighted averages were perceived as being a better likeness than exhaustive averages ($t(92) = 5.41$, $p < .001$) and exhaustive averages were perceived as being a better likeness than 2000s-weighted averages ($t(92) = 6.31$, $p < .001$). Moreover, while there were no significant differences in likeness between 2000s instances and 2000s-weighted averages ($t(92) = 1.83$, $p = .455$), 2010s and 2020s instances were perceived as a better likeness than exhaustive ($t(92) = 8.96$, $p < .001$) and 2020s-weighted ($t(92) = 8.52$, $p < .001$) averages respectively.

#### 2.2.2. Name verification

The average response time was calculated for correct name verification match trials only, with specific trials where response times were more than 2SDs away from the participant's average response time excluded.

Fig. 4 shows mean RTs across conditions. A 2 × 3 within-subjects ANOVA showed a significant effect of time point ($F(2, 188) = 3.37$, $p = .036$, $\eta^2_G = .002$). The main effect of image type was not significant ($F(1, 94) = 0.19$, $p = .661$, $\eta^2_G < .001$) and neither was the interaction between time point and image type ($F(2, 188) = 1.60$, $p = .206$, $\eta^2_G = .001$). Simple main effects showed that 2000s images/2000s-weighted averages were recognised significantly slower than 2010s images/exhaustive averages ($t(94) = 2.34$, $p = .021$) and 2020s images/2020s-weighted averages ($t(94) = 2.14$, $p = .035$). There was no difference in response time for 2010s/exhaustive and 2020s/2020s-weighted images ($t(94) = 0.23$, $p = .818$).

## 3. Experiment 2

The results from Experiment 1 appear to provide evidence for a recency effect, with more representational weight given to the most recently encountered facial appearance. However, the results could also be explained by an age effect. That is, rather than representations being weighted towards recent appearance, they could be weighted towards older age appearance (the recent images were also older in age). For example, research has shown face age can affect face recognition (e.g., the own-age bias, see Rhodes & Anastasi, 2012; Wiese, Komes and Schweinberger, 2013 for reviews). Therefore, in Experiment 2, we aim to disentangle the effects of recency and face age.

Here, we used a learning study that took place over a few days during which unfamiliar participants learned Coronation Street characters from videos taken over the last 20 years. Half the participants were presented with the videos in chronological order (e.g., videos from 2003 to 2007 on day 1, videos from 2011 to 2015 on day 4, videos from 2019 to 2023 on day 7). The other half of participants were presented with the videos in reverse-chronological order. All participants then completed the likeness task and name-verification task on day 8. This allowed us to determine whether the results in Experiment 1 were due to a recency effect, similar to that observed in research on memory, where items encountered at the end of a list are more successfully recalled, even after a substantial delay (Baddeley & Hitch, 1977; Glenberg, Bradley, Kraus and Renzaglia, 1983; Greene, 1986a, 1986b). We predicted that averages weighted towards the most recent encounters will be recognised more quickly and will represent a better likeness than averages weighted towards the information encountered first during the learning phase. Thus, participants who learn the characters in chronological order might prefer 2020s-weighted averages, whereas participants who learn the characters in reverse-chronological order might prefer 2000s-weighted averages.

Past research comparing averages and instances has produced mixed results (e.g., Burton et al., 2005; Noyes et al., 2021; Ritchie et al., 2018) and the present data does not lend support to theories that propose familiar faces are represented as averages (e.g., Burton, Jenkins and Schweinberger, 2011). In Experiment 1, we found little difference between averages and instances, despite our pre-registered predictions (we had initially expected there would be a difference). We therefore did not to include a comparison of averages and instances in Experiment 2.

### 3.1. Method

#### 3.1.1. Participants

Participants were recruited from Prolific using the following screeners: 1) age between 30 and 65, 2) approval rating 96–100, 3) first language English, 4) normal or corrected to normal vision, 5) primary language English, and 6) location USA. Our final sample consisted of 51 people - 26 participants completed the chronological presentation condition ($M_{age} = 53.1$; $SD_{age} = 9.3$; 12 female, 14 male) and 25 completed the reverse-chronological presentation condition (M age = 42.9; SD age = 10.6; 12 female, 12 male, 1 nonbinary). An a priori power analysis using G*Power revealed that a minimum sample size of 42 participants was required to detect a medium effect (Faul, Erdfelder, Lang and Buchner, 2007). Participants were paid £9 upon completion of the final part of the experiment.

#### 3.1.2. Materials

Five characters (out of 15) were used as learning stimuli in Experiment 2. Previous research by Mileva et al. (2020) showed that some faces change more across time than others. In order to have the best chance of finding an effect of recency or age, we opted to use characters that changed the most across the 20-year time span. We therefore asked a group of 50 participants recruited via Prolific to rate the averages used in Experiment 1 for similarity. For each of the 15 Coronation Street characters, participants were presented with face pairs and were asked

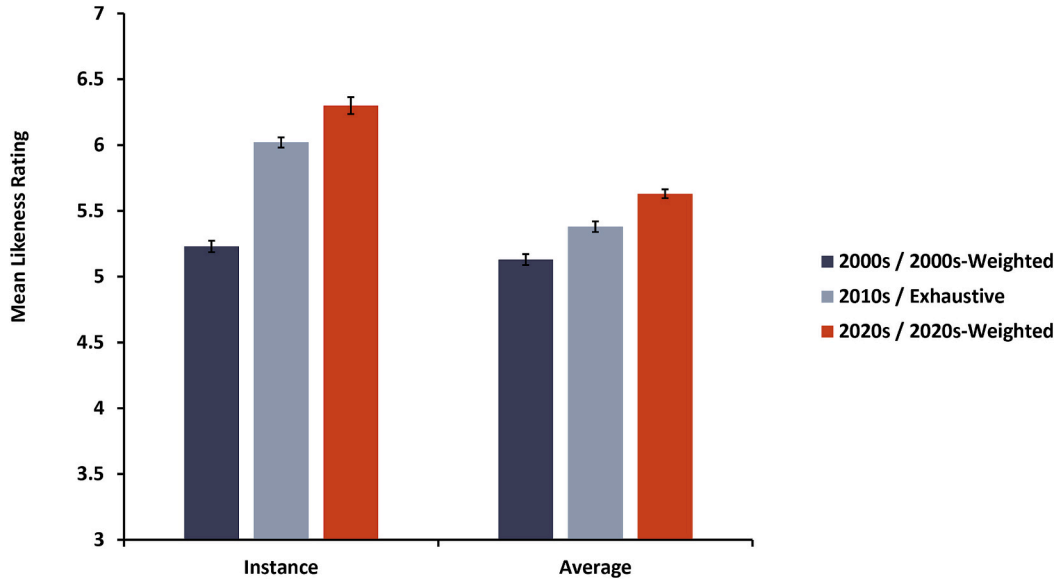


**Fig. 3.** Mean likeness ratings across time point and image type. Error bars show within-subjects standard error (Cousineau, 2005).

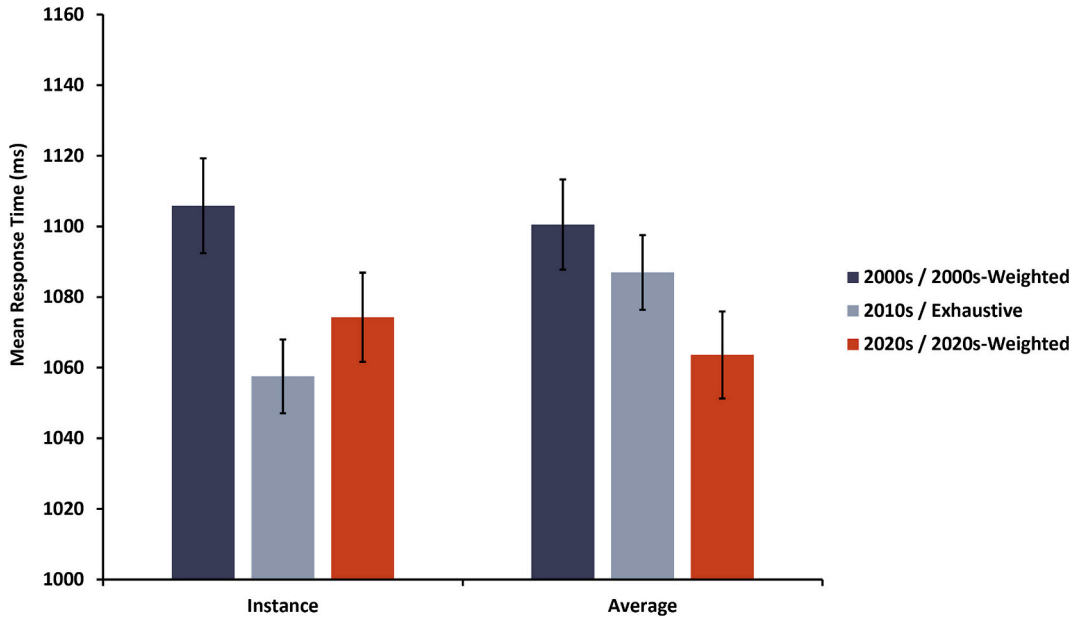


**Fig. 4.** Mean name verification response times across time point and image type. Error bars show within-subjects standard error (Cousineau, 2005).

to give a rating on how similar the images were to one another using a scale from 1 (not similar at all) to 9 (extremely similar). Participants rated a total of 30 trials - 2 per character, with one pairing the 2000s-weighted average with the exhaustive average and another pairing the 2020s-weighted average with the exhaustive average. The five characters with the lowest similarity ratings were used in Experiment 2 (David Platt, Maria Connor, Kevin Webster, Sally Webster, and Chesney Brown).

For each of the five characters, we created 15, 20-s video clips from episodes of Coronation Street that aired between 2003 and 2023. Five clips were taken from each of the time windows: 2003–2007, 2011–2015, 2019–2023. Footage was taken from YouTube videos that met the following criteria: 1) the date of the episode was stated, 2) the video showed a clear view of the character's face, and 3) there were no other faces in shot. In addition to the videos, the same three types of averages – an exhaustive, a 2000s-weighted and a 2020s-weighted average was used for each of the five characters. These were the same images as used in Experiment 1.

### 3.1.3. *Procedure*

The study consisted of three learning sessions, followed by a test. Participants were asked to complete a learning session on days 1, 4, and 7, with a test session on the final 8th day. In each learning session, participants were briefly familiarised with the five characters. They were first presented with the name of the character (e.g., 'You will now see a video of Chesney Brown'), followed by a short video clip showing the character. The name of the character was displayed on the screen while the video was playing. The video was played once only and to ensure participants were paying full attention, they were asked to click on the video in order to play it whenever they were ready. At the end of the video clip, participants were presented with the names of all characters used in the study and had to select the character they had just

seen. This procedure was repeated for each of the five characters, with participants receiving feedback on every trial. At the end of each familiarisation session, participants were presented with 20 new video clips (4 for each character) in a random order and were asked to select the correct name of the character from a list of all five. Participants received feedback on every trial, with the correct names being shown when errors had been made.

Participants were assigned to one of two conditions corresponding to the order in which the videos were presented – one group saw the videos in chronological order (i. e., videos from 2003 to 2007 first) while the other group watched the videos in reverse-chronological order (i.e., videos from 2023 to 2019 first).

In the test session, participants were asked to complete two tasks – likeness ratings and a speeded name verification. For the likeness rating task, participants were presented with a total of 15 average images (5 characters × 3 types of average weighting) in a random order and were asked to rate to extent to which each image was a good likeness of the character using a scale from 1 (very bad likeness) to 7 (very good likeness). The name verification task followed the same procedure as in Experiment 1, with the only difference being the reduced number of trials – 30 (5 characters × 3 types of average weighting x 2 trial types – match or mismatch). Unlike in Experiment 1, only average images, and no instance images, were used here. The order of the likeness rating and name verification tasks was randomised individually for each participant. At the end of the test session, participants were also asked to report whether they were familiar with any of the characters prior to participating in study and whether they have ever watched Coronation Street.

### 3.2. Results

None of the participants reported being previously familiar with the characters shown. One participant reported having watched Coronation Street in the past, but was not familiar with the characters. Therefore, this participant was included in the subsequent analysis. Accuracy in the final third familiarisation task was high, with an average of 98.1% ($SD$ = 3.7%). There were two participants who made an error on more than two trials – one in each of the two presentation order conditions. These participants were, therefore, removed from any subsequent analyses.

#### 3.2.1. Likeness ratings

Mean likeness ratings were calculated for each participant, separately for 2000s-weighted, exhaustive and 2020s-weighted averages (see Fig. 5). Data were analysed with a 2 × 3 mixed factorial ANOVA with factors: average type (2000s-weighted, exhaustive, or 2020s-weighted), manipulated within-subjects and learning direction (chronological or reverse-chronological), manipulated between-subjects. Significant main effects and interaction were followed up with Tukey-corrected post hoc tests. This analysis showed a significant main effect of average type ($F(2, 98) = 18.5$, $p < .001$, $\eta^2_G = .052$), with 2000s-weighted averages being rated as a significantly worse likeness than both exhaustive $t(49) = 3.61$, $p = .002$) and 2020s-weighted $t(49) = 5.93$, $p < .001$) averages. There were no significant differences between exhaustive and 2020s-weighted averages ($t(49) = 2.31$, $p = .064$). The main effect of learning direction was not significant ($F(1, 49) = 0.30$, $p = .587$, $\eta^2_G = .005$). However, the interaction between average type and learning condition was significant ($F(2, 98) = 3.18$, $p = .046$, $\eta^2_G = .009$). Likeness ratings showed the same pattern for both learning conditions, with 2000s-weighted averages receiving the lowest ratings, followed by exhaustive averages and finally, 2020s-weighted averages received the highest likeness ratings. In the chronological order condition, the differences between the three types of averages were not significant ($t_{max} = 2.61$, $p_{min} = .115$). In the reverse-chronological order condition, 2000s-weighted averages were rated as a significantly worse likeness than exhaustive averages ($t(49) = 3.85$, $p = .004$) and 2020s-weighted averages ($t(49) = 5.75$, $p < .001$). There were no significant differences in likeness ratings between exhaustive averages and 2020s-weighted averages ($t(49) = 1.82$, $p = .464$).

#### 3.2.2. Name verification RTs

Overall accuracy on match trials was high, with an average of 93.1% ($SD = 10.2\%$). There were two participants with an accuracy lower than 70% - these participants were excluded from the subsequent analysis.[1] Fig. 6 shows the mean response times across average type and learning condition. A 2 × 3 mixed factorial ANOVA showed a significant main effect of average type ($F(2, 94) = 19.80$, $p < .001$, $\eta^2_G = .060$). 2000s-weighted averages were recognised significantly slower than both exhaustive ($t(47) = 5.54$, $p < .001$) and 2020s-weighted ($t(47) = 4.93$, $p < .001$) averages. There were no significant differences in recognition response times between exhaustive and 2020s-weighted averages ($t(47) = 0.75$, $p = .738$). The main effect of learning direction was not significant ($F(1, 47) = 0.57$, $p = .454$, $\eta^2_G = .010$) and neither was the interaction between average type and learning direction ($F(2, 94) = 1.45$, $p = .239$, $\eta^2_G = .005$).

### 3.3. Interim discussion

Experiment 1 shows a consistent data pattern across both the likeness rating and the name verification tasks where images that capture the character's most current appearance (2020s instance images and 2020s-weighted average images) more closely resemble the mental representation of that character. This suggests that our mental representations of familiar identities may be weighted towards recent encounters. However, the results of Experiment 2 force a re-examination of that conclusion. While we predicted a recency effect, based on research in memory and learning (e.g., Baddeley & Hitch, 1977), our facial representations seem not to be dominated by our most recent encounters with someone, but by the oldest age at which we have seen them. When shown the same person at older and younger ages, participants seem to develop a representation based on the older age, regardless of the order in which they have encountered this person.

These findings imply a more complex process of learning familiar identities than simple recency. Instead, they suggest that abstracting and updating mental representations of faces might recruit more sophisticated processes that incorporate the passing of time and the related consequences to human facial appearance.

Given these surprising results, it is important to consider whether they might be driven by some stimulus-specific properties of the images used in our experiments. One possibility is that the experimental images might differ in ways that favour more recent images, showing the people in older age. For example, image quality in broadcast TV has improved over the past 20 years, with technical advances, and this may somehow lead to confounds in our stimuli between aging characters and recency of image sampling. If the 2020s images, taken from modern TV, were more discriminable (for example) this might lead to preferential likeness ratings.

To address these issues, we measured the physical similarity of the images included in the three types of averages. This was done by subjecting all 315 images (21 per character) to PCA, which is often used to measure the dimensions that capture the variability contained in a certain set of face images. For this analysis, face shape was already extracted as part of the average generation process by aligning the images to a standard face template consisting of 82 fiducial points (see Materials in Experiment 1). All images were then standardised by first calculating the average shape across the entire image set and morphing the images to it (Burton, Miller, Bruce, Hancock and Henderson, 2001;

[1] These were two additional excluded participants to the exclusions reported for the likeness rating analysis. For improved consistency between the two types of tasks, the likeness analysis was repeated with data only from those participants included in the name verification analysis. This resulted in the same pattern of results reported in the original analysis.

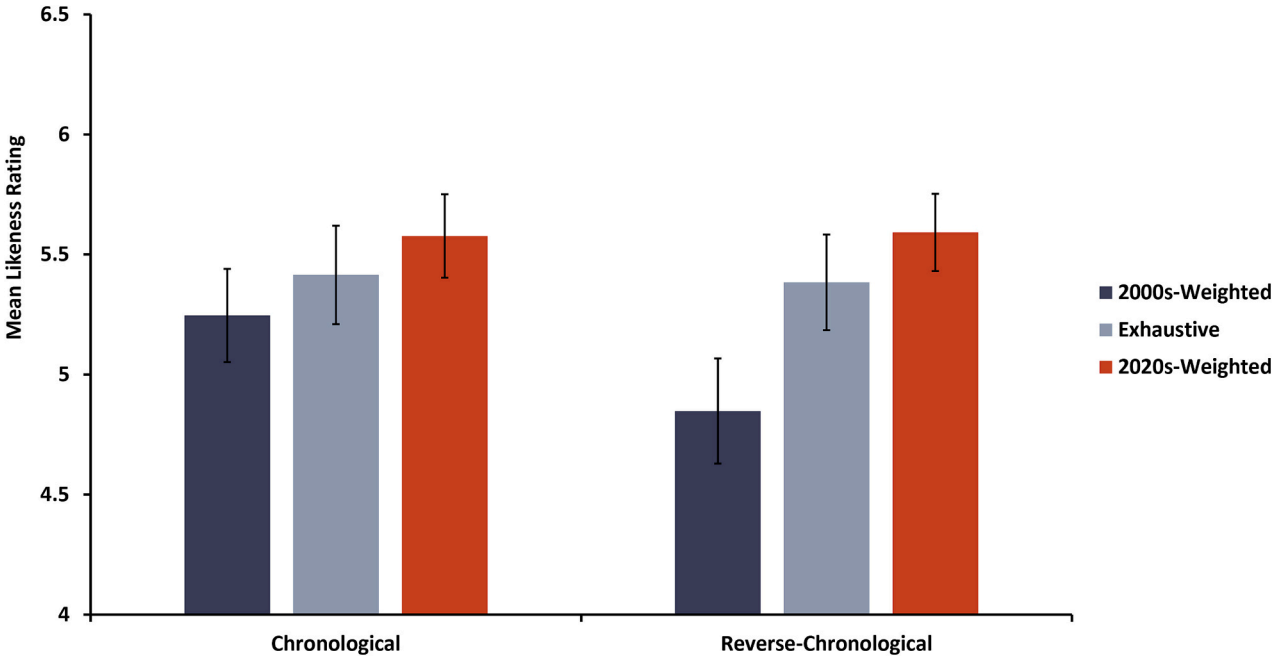


**Fig. 5.** Mean likeness ratings across time point and learning direction. Error bars show standard error.

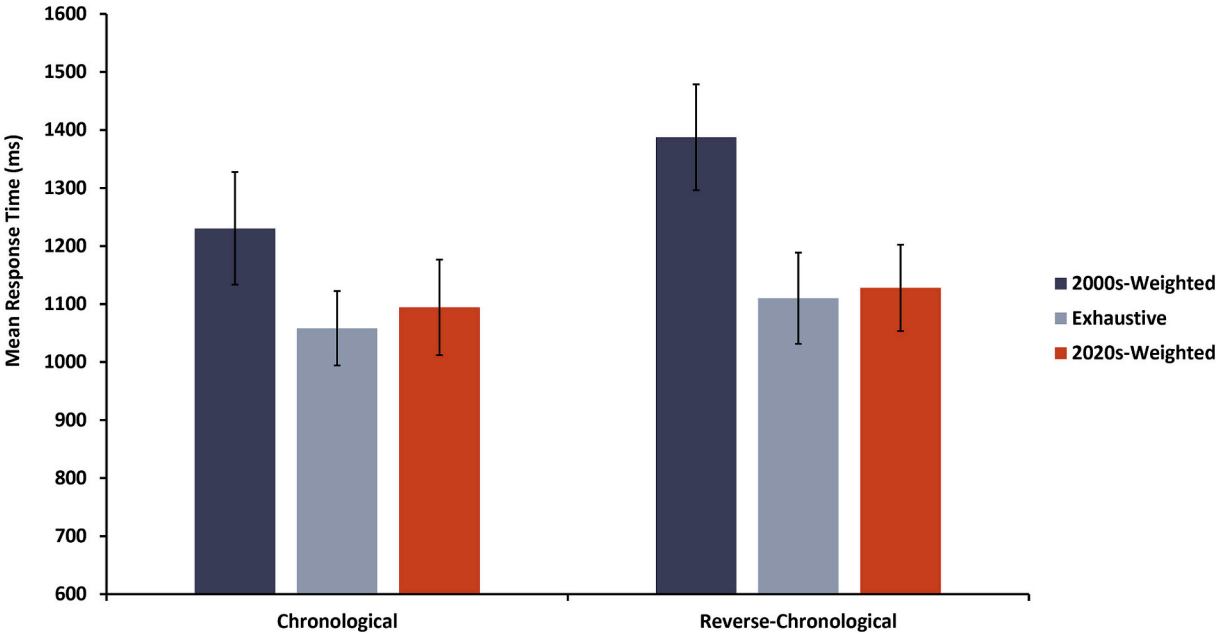


**Fig. 6.** Mean name verification response times across average type and learning direction. Error bars show standard error.

Craw, 1995). This resulted in shape standardised face textures consisting of pixel intensities across 3 colour layers (RGB). PCA was applied to these normalised textures (Rogers et al., 2022), with each dimension capturing a certain pattern of variability within the set. Fig. 7 shows the variability captured by the first 5 dimensions. The first dimension seems to be coding for general differences in lighting across the entire face, while the second seems to code for variability in lighting from the right to the left side of the face. Further dimensions capture differences in hair colour as well as presence or absence of facial hair. By definition, the initial dimensions capture the largest amount of variability, with less and less variability explained by every next dimension.

Only the first 70 dimensions explaining 94% of the image set variance were retained for the subsequent analyses. Therefore, following PCA, images were represented in a 70-dimensional face space with the location of each image within the face space coded as a unique set of 70 reconstruction coefficients of mean zero. We then used the location of each image within this space as a measure of image similarity – highly similar images will lie closer to each other in this space, whereas images that are dissimilar will be located further apart from one another. To measure the similarity of the three types of averages used in Experiments 1 and 2, all 45 averages (3 types × 15 characters) were projected into the same 70-dimensional space, allowing us to establish their location in the space relative to the other images.

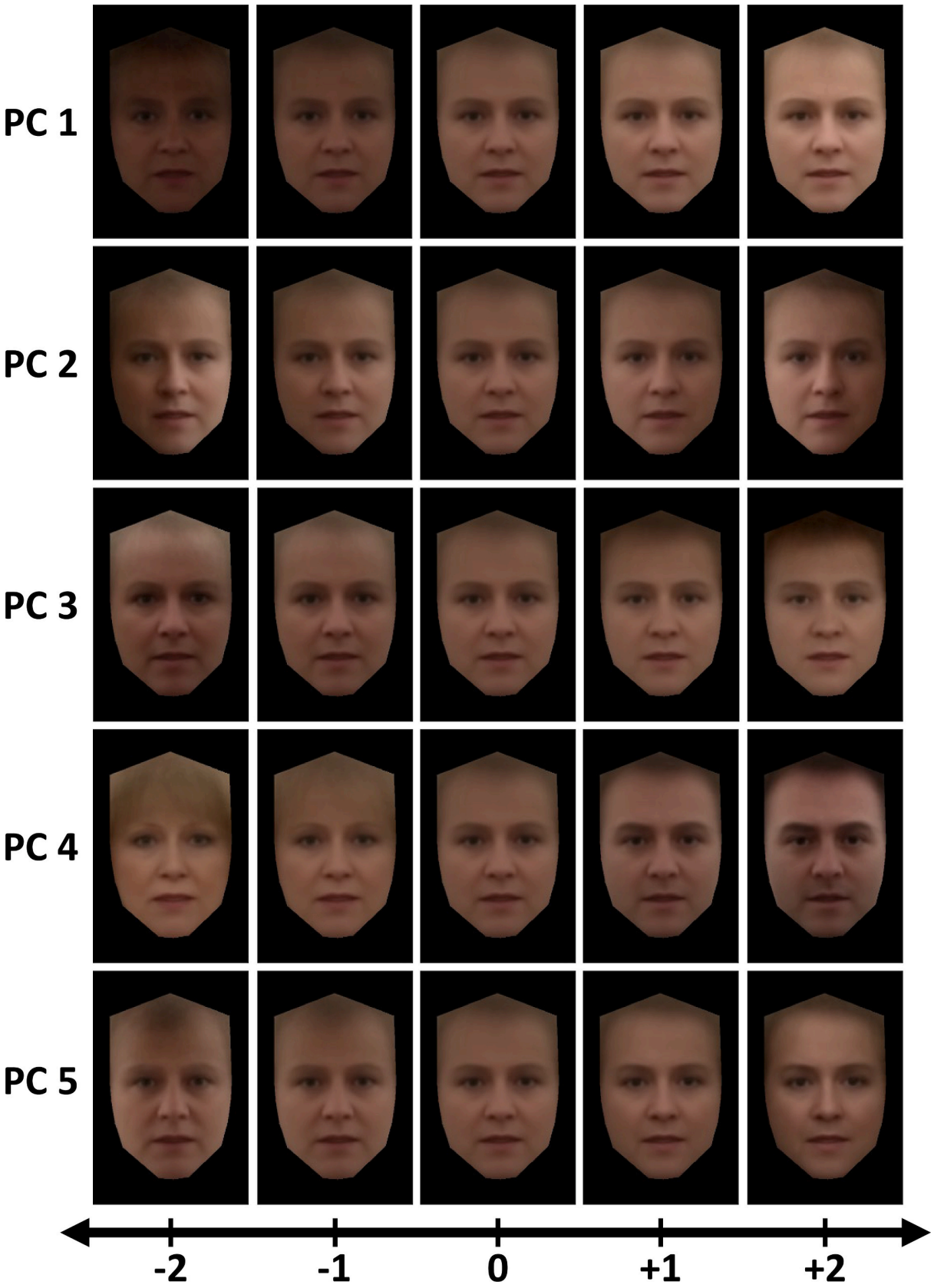


**Fig. 7.** First five principal components capturing the variability in the image set.

For each character, we then calculated the Euclidean distance between each pair of 2000s images, each pair of 2010s images and each pair of 2020s images.[2] There were 21 pairs for each character. We also calculated the Euclidean distance between each 2000s and 2010s, 2000s and 2020s, and 2010s and 2020s images – 49 pairs of images per character. Finally, we calculated the Euclidean distance between each 2000s/2010s/2020s image and the three types of average (2000s-weighted, exhaustive, and 2020s-weighted) – 7 pairs per character. Fig. 8 shows the average distances across all 15 characters. Larger values here indicate higher dissimilarity (i.e., images located further apart from one another), whereas lower values show higher similarities (i.e., images located closer to one another).

No substantial or systematic differences were found in the similarity of 2000s, 2010s and 2020s images. The images taken earlier on seemed to be, on average, further apart from one another and recent images seem to be the closest – so there is no hint that advances in TV technology have led to a level of facial discriminability that could explain our results. In fact, all differences are rather small, suggesting that the pattern of results observed in these studies is unlikely to be driven by differences in the levels of variability captured by the images. In terms of the similarity between instance and average images, it is reassuring that 2000s-weighted averages are closest to the images taken early on, while 2020s-weighted averages are closest to the most recent images. It is also interesting to note that while computationally exhaustive averages were located closest to all images (across the three time points), perceptually participants in both experiments reported that the 2020s-weighted averages best resembled the characters, with higher likeness ratings and shorter name verification RTs.

Additionally, in Experiment 1, a difference between instances and averages was only observed for the likeness task, with 2020s and 2010s instances rated as a better likeness than 2020s-weighted and non-weighted (exhaustive) averages. There was no difference in reaction time between averages and instances in the name-verification task. Therefore, we report no clear benefits for averages over instance images or vice versa. It should also be noted that due to the way they are created, averages often have a softer and more blurred appearance than many instance images, and so it is likely that any potential differences between these two types of images are due not only to the physical information captured by them but also by these lower-level differences in the properties of the images. Relatedly, similar superficial differences might be responsible for 2000s-weighted averages being less preferred than both 2020s-weighted and exhaustive averages. This might in turn artificially inflate the benefits from these latter types of images reported in Experiments 1 and 2. In other words, the results of Experiment 1 and 2 might be due to a stimulus effect, with participants preferring or being biased towards the more recent images relative to the 2000s-weighed images. Therefore, in our third experiment, to rule out a possible stimulus effect, participants were familiarised with the same 5 characters from Experiment 2 only using videos showing them at the start of the 20-year period (2003–2004). This allowed us to ensure that learning is possible with these images and that there were no unintentional reasons why these images were the least preferred in the previous studies in terms of both likeness ratings and name verification response times.

## 4. Experiment 3

In Experiment 3, participants new to Coronation Street were familiarised with the same 5 characters used in Experiment 2. Unlike in the previous experiment, however, these participants only watched videos from 2003 to 2004 so were only familiarised with the way these characters looked at the start of our 20-year period. Unlike the previous experiments, here we do not aim to test generalisation across time per se, but to rule out any potential inadvertent biases in stimulus selection. We predict that averages weighted towards these early (2000s) encounters will be recognised more quickly and will represent a better likeness than averages weighted towards recent (2020s) instances. On the other hand, if participants in these circumstances continued to prefer images of the people shown at older age, this would suggest an explanation in which the recent averages were preferred under all circumstances, due to some inherent characteristics of the stimuli we have used in these experiments.

### 4.1. Method

#### 4.1.1. Participants

A total of 42 participants were recruited from Prolific using the same criteria as in Experiment 2 (23 male, 19 female, $M_{age} = 51.3$, $SD_{age} = 10.3$). Participants who had participated in Experiment 2 were excluded from this experiment. An a priori power analysis using G*Power revealed that a minimum sample size of 28 participants was required to detect a medium effect (Faul et al., 2007). Participants were paid £4 upon completion of the final part of the experiment.

#### 4.1.2. Materials

A set of 30 20-s short video clips was created, with 6 videos for each of the five characters used in Experiment 2. These videos were created using the same approach and procedure, showing Coronation Street excerpts from the year that had been weighted most in the average – 2003/2004). The same 15 average images used in Experiment 2 were used here – 3 per character, an exhaustive average, a 2000s-weighted average and a 2020s-weighted average. As part of a distractor task used between the learning and test sessions, a scene from Where's Waldo was downloaded using a Google Image Search.

#### 4.1.3. Procedure

Participants completed the study in one sitting. First, they completed a single learning session that followed the same procedure as the learning sessions used for Experiment 2. Unlike in Experiment 2, participants only saw videos of characters taken from a single year. Participants watched 5 videos in the familiarisation stage (1 per character) and 25 videos in the main learning stage (5 per character). After learning the characters, participants were asked to complete a short distractor task where they saw a Where's Waldo scene and had to answer as many questions about the scene as possible for 5 min (e.g., “How many beach balls are there” or “How many women are wearing a red bathing suit?”). Following the distractor task, participants were asked to complete the same name verification and likeness rating tasks as in Experiment 2 and were also asked whether any of characters were familiar to them prior to taking part in the study and whether they had watched Coronation Street in the past.

### 4.2. Results and discussion

Participants' accuracy was high following the familiarisation procedure with a mean of 95% ($SD = 0.11$). As outlined in the study preregistration, participants who made more than 2 errors were excluded from subsequent analyses ($N = 5$). Another eight participants were excluded as they had an average accuracy of less than 70% on the match trials in the name verification task. Fig. 9 shows mean likeness ratings and Fig. 10 shows average RTs across the three time points.

[2] This image set contains images taken between 2017 and 2025.

| A | 2000s Images | 2010s Images | 2020s Images |
|---|---|---|---|
| 2000s Images | 11.7 | | |
| 2010s Images | 11.3 | 10.4 | |
| 2020s Images | 11.5 | 10.5 | 10.1 |

| B | 2000s-Weighted Averages | Exhaustive Averages | 2020s-Weighted Averages |
|---|---|---|---|
| 2000s Images | 8.1 | 8.1 | 9.3 |
| 2010s Images | 8.6 | 7.2 | 8.2 |
| 2020s Images | 8.7 | 7.1 | 7.0 |

**Fig. 8.** Mean Euclidean distances between images used in experiments 1 and 2. A shows distances between 2000s, 2010s, and 2020s images and B shows distances between all images and the three types of averages.

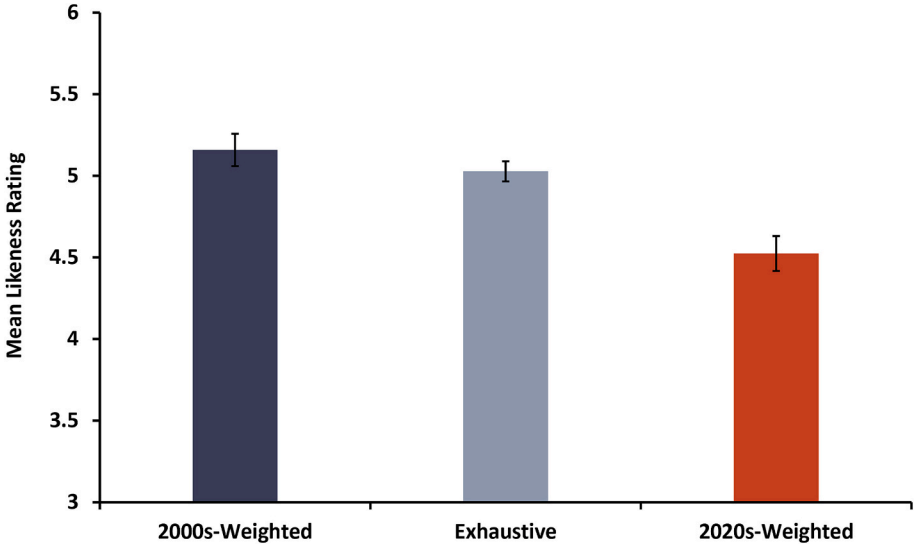


**Fig. 9.** Mean likeness ratings across time point. Error bars show within-subjects standard error (Cousineau, 2005).

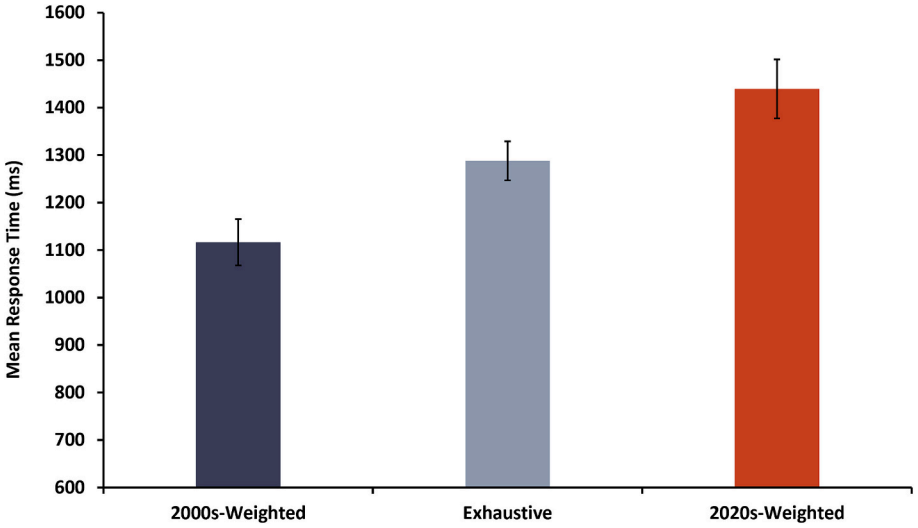


**Fig. 10.** Mean name verification response times across time point. Error bars show within-subjects standard error (Cousineau, 2005).

#### 4.2.1. Likeness ratings[3]

A one-way repeated measures ANOVA showed a significant main effect of time point ($F(2, 56) = 8.98$, $p < .001$, $\eta^2_G = .24$). Post hoc tests with a Tukey correction showed that 2020s-weighted averages were rated as a significantly poorer likeness than exhaustive and 2000s-weighted averages, $t(28) = 3.51$, $p = .004$) and $t(28) = 3.22$, $p = .009$, respectively. 2000s-weighted averages and exhaustive averages were perceived similar in likeness, $t(28) = 1.05$, $p = .555$.

#### 4.2.2. Name verification accuracy & RTs

Overall accuracy on the name verification task was high ($M = 84.3\%$, $SD = 17\%$). A one-way repeated measures ANOVA showed a significant main effect of time point ($F(2, 54) = 12.54$, $p < .001$, $\eta^2_G = .26$). 2020s-weighted averages ($M = 72.1\%$, $SD = 17.5\%$) were recognised significantly less accurately than exhaustive ($M = 89\%$, $SD = 12.7\%$) and 2000s-weighted averages ($M = 91.4\%$, $SD = 13.8\%$), $t(27) = 3.96$, $p = .001$) and $t(28) = 4.25$, $p < .001$, respectively. While higher numerically for 2000s-weighted averages, there were no significant differences in accuracy between 2000s-weighted and exhaustive averages ($t(27) = 0.57$, $p = .837$).

For response times, a one-way repeated measures ANOVA showed a significant main effect of time point ($F(2, 54) = 6.12$, $p = .004$, $\eta^2_G = .04$). Post hoc comparisons with a Tukey correction showed that 2000s-weighted averages were recognised significantly faster than both exhaustive and 2020s-weighted averages, $t(27) = 2.53$, $p = .045$) and $t(27) = 2.99$, $p = .016$, respectively. There were no significant differences in recognition response times for exhaustive and 2020s-weighted averages $t(27) = 1.56$, $p = .278$).

These results show a clear advantage for younger-aged averages, i.e. those derived from images taken early in the period of exposure to Coronation Street actors that we have studied here. 2000s-weighted averages were rated as having higher likeness than averages of the same actors taken later in the period, and these same 2000s-weighted averages were recognised faster in the name verification task. This effectively eliminates one alternative explanation for our results. If the 2020s-weighted average (taken when the person was older) had been preferred even when participants were exposed only to more youthful instances, then we could have explained these results in terms of some general preference for the 2020s-weighted averages we have used. But this is not the case – when exposed only to youthful examples of someone, a youthful average is preferred.

## 5. Experiment 4

Taken together, the PCA analysis and the results of Experiment 3 rule out alternative explanations for the results of Experiments 1 and 2. Participants in Experiment 3 rated the 2000s-weighted averages as a better likeness and recognised them faster in the name verification task. In addition, PCA suggests little difference in discriminability of the stimuli across the different time points.

The results of Experiments 1 and 2 are challenging for accounts of face learning based on simple statistical abstraction (Burton et al., 2016; Matthews & Mondloch, 2021; Murphy, Ipser, Gaigg and Cook, 2015). In particular, the idea that face representations are based on a simple agglomeration of all one's encounters seems inadequate (Burton et al., 2005). Instead, these data suggest that the abstraction process is informed and constrained by natural aspects of the world.

We propose two possible explanations. The first is natural chronology and our knowledge that facial aging in the real world is directional (young to old). It is possible that we "know" that faces age and therefore assume that the older-looking a person appears, the more likely that represents their current appearance. This is consistent with Schneider and Carbon (2021) who report that regardless of whether participants learned younger or older photographs of faces, the older photographs were subsequently rated as more prototypical than the younger. Nevertheless, other research has found little evidence of an effect of aging direction when generalising across changes in age (Sexton, Mileva, Hole, Strathie and Laurence, 2023).

An alternative explanation draws on the idea that faces become more distinctive as they age. For example, O'Toole, Vetter, Volz and Salter (1997) applied a caricature algorithm to faces which exaggerated the difference between an individual face and an average face (the average was based on a large number of faces). O'Toole et al. observed a linear relationship between perceived age and distance from the average; as faces became increasingly caricatured, they also appeared to have aged. Deffenbacher, Vetter, Johanson and O'Toole (1998) subsequently showed that distinctiveness ratings also showed a linear relationship with caricature and that caricatured faces were remembered better after a brief learning phase. Rather than an effect of aging, the results of Experiments 1 and 2 could therefore be explained by distinctiveness, with older appearance being more distinctive and better remembered.

Experiment 4 was designed to explore the relationship between distinctiveness and age for the Coronation Street stimuli used in the present research. A new sample of unfamiliar participants were asked to rate the images used in Experiment 1 and 2 for distinctiveness. If a relationship exists, then we expect that the older images (e.g., 2020-weighted averages) used in Experiment 1 and 2 would be perceived as more distinctive than the younger images (e.g., 2000-weighed images).

### 5.1. Method

#### 5.1.1. Participants

Participants were recruited from Prolific using the same screeners as in Experiment 2. We also excluded participants who had completed previous experiments from taking part. Our final sample consisted of 92 people ($M_{age} = 39.5$; $SD_{age} = 11.8$; 42 female, 48 male, 2 other). An a priori power analysis for a 2 × 3 within-subjects design using the Superpower package in R (Lakens & Caldwell, 2021) revealed a required sample of 90 participants with 80% power. In order to reach this sample size, we initially recruited 114 participants. Out of these, 16 were excluded because they reported being previously familiar with the characters and the Coronation Street show, 3 were excluded because they failed to respond to the attention and comprehension checks accurately and another 3 participants were excluded because they reported having technical difficulties with the study.

#### 5.1.2. Materials

The stimuli were the same as in Experiment 1. For each of the 15 Coronation Street characters, there were three averages and three instances.

#### 5.1.3. Procedure

Participants were instructed to rate each face on how distinctive it was on a scale from 1 (extremely typical) to 9 (extremely distinctive). The instructions explained that a distinctive face is one that is unusual, and would stand out in a crowd of more typical faces. Participants were then presented with 90 face images, with each face remaining on the screen until a decision was made. Each participant was also asked to complete three attention check trials where instead of a face image, they were presented with a number between 1 and 9 on the screen and they were instructed to select that number on the rating scale. Upon completion of the ratings task, participants were also asked whether any of characters were familiar to them prior to taking part in the study and whether they had watched Coronation Street in the past.

[3] Given that participants did not receive as much exposure to the tested identities as participants did in Experiments 1 and 2, it is possible that in some trials, participants were unable to recognise the depicted identity. It such cases, their likeness ratings might not be as meaningful and this should be taken into consideration when interpreting the results of the likeness ratings analysis.

### 5.2. Results & discussion

Fig. 11 shows the mean distinctiveness ratings for each image type and time point. Distinctiveness data were analysed with a 2 × 3 within-subjects ANOVA with factors: image type (instance vs average) and time point (2000s/2000s-weighted vs 2010s/exhaustive vs 2020s/2020s-weighted). The main effects of image type ($F(1,91) = 40.82$, $p < .001$, $\eta^2_G = .035$) and time point ($F(2,182) = 7.14$, $p = .001$, $\eta^2_G = .004$) were both significant and so was the interaction between them ($F(2,182) = 4.11$, $p = .018$, $\eta^2_G = .002$). Tukey-corrected post hoc comparisons showed no significant differences in the perceived distinctiveness of 2000s-weighted, exhaustive and 2020s-weighted averages ($t_{max} = 2.19$, $p_{min} = .252$). Both 2000s instances and 2010s instances were perceived as being significantly less distinctive than 2020s instances, $t(91) = 3.07$, $p = .032$ and $t(91) = 3.48$, $p = .01$ respectively. Instances from the 2000s and 2010s did not differ significantly in their perceived distinctiveness, $t(91) = 0.91$, $p = .942$.

We also calculated the correlation between the average distinctiveness rating attributed to each image and the corresponding average likeness rating attributed to the same image using the data from highly familiar participants (Experiment 1). This analysis showed a significant positive correlation between ratings of distinctiveness and likeness ($r(88) = .40$, $p < .001$. This correlation, however, was mostly driven by ratings attributed to average images ($r(43) = .39$, $p = .008$), while the correlation for instances was not significant ($r(43) = .27$, $p = .070$.

The small differences in distinctiveness ratings could potentially account for the instance data from Experiment 1, where 2020s instances were perceived as a significantly better likeness than 2010s which were then perceived as a significantly better likeness than 2000s images. However, perceptions of distinctiveness do not seem to explain any of the differences observed for averages in Experiments 1 and 2, where 2020s-weighted averages were generally perceived as a better likeness and were recognised faster than the other types of averages. Similarly, while distinctiveness and likeness seem to be significantly related when attributed to average images, this does not hold true for instances. This, again, indicates that perceptions of distinctiveness might only partially account for the higher likeness ratings observed in images showing the most recent appearance of the TV show characters.

It is interesting to note that ratings of distinctiveness seem to reflect differences in image quality and image blurriness in particular. More recently-taken instances (such as those taken in the 2020s) inadvertently have better, sharper quality than instances taken early on due to changes in the quality of TV broadcasting across time. This fits well with the findings that more recently-taken instances were perceived as being more distinctive than those taken earlier on. Similarly, due to the way they were created, 2000s-weighted and 2020s-weighted averages are likely to appear sharper than exhaustive averages. This is because some instances were weighted more heavily in these averages, making them appear sharper, whereas each instance received the same weighting for the exhaustive averages which could lead to them appearing blurrier than the other averages. Such results, therefore, further strengthen the idea that perceivers seem to prioritise information about natural chronology when forming or updating their representations of known identities.

## 6. General discussion

The results of these studies consistently show that the representations we use to recognise familiar faces derive from sophisticated processes which abstract over encounters. Across four pre-registered experiments, we have shown that a simple average of all the encounters we have ever made with a face is insufficient to form an optimal representation of that person. Instead, patterns inherent in the aging process itself seem to be recruited. We expand on these ideas below.

In Experiment 1, we observed an effect that seems straightforwardly intuitive – recognition of people we have known over 20 years is weighted towards recent encounters. Under most circumstances, in daily life, recency of encounter with someone corresponds to that person's aging – each time we meet, we are all a little older. As psychologists trying to understand our flexible representations, we could account for changes in terms of recency effects – which are typically strong in both short term and long term memory (Baddeley & Hitch, 1977; Glenberg et al., 1983; Greene, 1986a, 1986b). However, modern media affords an opportunity to test this account. For actors and other celebrities with a media presence spanning many years, it is possible to sample our encounters in non-chronological manner.

In Experiment 2, we taught participants to recognise new people, constraining their encounters with these people either to follow or reverse natural chronology. So, some participants experienced people when they were at their youngest first, while others experienced these people at their oldest first. All participants saw the same set of faces – in clips from their acting careers – the only thing that differed was the

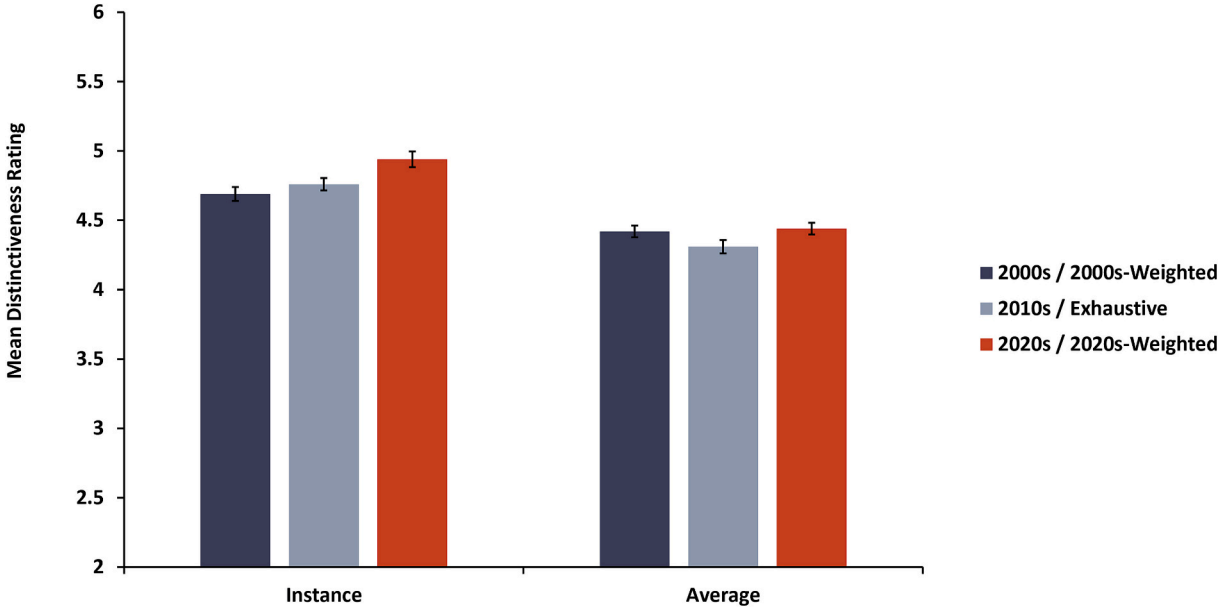


**Fig. 11.** Mean distinctiveness ratings across time point. Error bars show within-subjects standard error (Cousineau, 2005).

sequence in which they saw them. Here we observed a very striking result: regardless of order encountered, people's representations of these learned faces were dominated by the people at their oldest. Contrary to our initial predictions, recency of encounter did not explain the results. Instead, it seems that our ability to form abstract representations of people is informed by natural chronology – if we have encountered people spanning many years, we tend to represent them most in their later years.

This observation is rather surprising, and for this reason, we examined alternative explanations. It is possible that the particular stimuli we have used give rise to this effect for artefactual reasons – i.e. perhaps people tend to 'prefer' images based on more recent TV clips from the soap opera for some reason reflecting superficial changes in the images. This could come about, for example, because TV technology has changed over the years that we have sampled.

We first examined the stimuli themselves. One way the technical differences could support the results here is if people in recent images are more discriminable than images from an earlier age of TV. Resolution has certainly improved over the years, and perhaps this makes individuals more discriminable, and so easier to learn. However, our modelling showed no evidence supporting this explanation. Working in a PCA person-space, we showed that, if anything, there was a slight tendency for earlier images to be more discriminable, and so this account based on image quality cannot explain our results.

We next tested people's response to our test images to rule out a general preference for the 2020s-weighted images. In Experiment 3, we taught participants the same faces used in Experiments 1 and 2, but this time using only early (2000s) examples, when the actors were at their youngest. Following learning, we asked whether perceivers would nevertheless prefer representations based on recent images – but this was not the case, when exposed to a confined range, participants developed representations based on that range (also see Devue, Wride and Grimshaw, 2019). So, there seems no inherent stimulus-driven reason that participants in earlier experiments had preferred the 2020s-weighted representations – this just seems to be a natural consequence of exposure to someone new over a range of ages.

Finally, in a fourth experiment, we examined the role of perceived distinctiveness and whether it can account for the pattern of results we have observed in the first two experiments. That is, are 2020s instances and 2020s-weighted averages perceived as better representations of the characters because face distinctiveness increases as we age (e.g., O'Toole et al., 1997)? While more recently-taken 2020s instances were perceived as more distinctive than 2000s and 2010s images, no significant differences in distinctiveness perception were found for averages. Perceived distinctiveness was also only significantly related to perceived likeness when these were attributed to average images, but not instances. While these results suggest that increased distinctiveness as we age might play some role in the preference for the most up-to-date appearance of known identities, it is unlikely that distinctiveness alone can account for this pattern of results. This further helps strengthen the idea that there are some underlying mechanisms that prioritise information about the way someone looks right now, at their oldest. Nevertheless, given that there is some indication for the potential role of distinctiveness, it might be worthwhile for future work to consider this further, especially with distinctiveness ratings from highly familiar observers.

### 6.1. Likeness and face averages

It might be important to consider what exactly might be captured by a likeness rating. Similarly to previous work, here we think of likeness as a measure of how well a particular image looks like the person it is depicting. We further assume that this judgement is made by comparing a specific image of a familiar person to our stored representation of this person, where a higher likeness rating indicates that the image more closely approximates our stored representation of this person. Such judgements have been shown to be important in familiar face recognition. Hay, Young and Ellis (1991), for example, report that the most frequent explanation people used when they failed to recognise familiar identities is that the image presented was a bad likeness of that person. What is more, images deemed to be a good likeness are often recognised faster and higher levels of familiarity lead to increased likeness ratings (i.e. as we get familiar with a person, the range of images that we deem to resemble this person well expands, Ritchie et al., 2018). This suggests that perceptions of likeness could be idiosyncratic and might depend on the specific personal experience of each individual observer. While the idiosyncratic nature of likeness perception might pose some challenges for the results reported in Experiment 1, where the exact exposure of each participant to the characters was not measured, this was resolved in Experiment 2 where each participant was exposed to the same visual information. Most importantly, both experiments reveal a bias towards images showing the oldest appearance of the characters, regardless of the order this information was presented in or the overall levels of exposure and familiarity.

Given previous work has implicated a benefit of face averages for face recognition (Burton et al., 2005), it is somewhat surprising that average images received lower ratings of likeness compared to instances. A similar pattern has also been reported by Ritchie et al. (2018) and could suggest that a likeness rating might not be a direct measure of resemblance to our mental representations. Recently, Balas, Sandford and Ritchie (2023) showed that there is no strong one-to-one mapping between ratings of likeness attributed to an image and the similarity of this image to an average of the depicted identity.

One, more trivial explanation for why averages might be perceived as a poorer likeness might come from the superficial differences between instance and average images imposed by the current state of the technologies used to create average images. Averages often have a softer, blurrier appearance that lack the sharp textural details present in the images used for their creation. This might not only result in averages being perceived as less distinctive, as we show in Experiment 4, but also that they might be perceived to show a younger version of the target identity. Wrinkle lines might appear less pronounced and skin texture might appear softer and more even. Previously, George and Hole (1995) have shown that image manipulation that removes all skin texture details led to an underestimation of over 20 years by some participants and Burt and Perrett (1995) have also shown that composite images, similar to the average images used here, were perceived as younger than the images used to create them, a pattern that became stronger with older faces. As our data suggest that observers value highly information about the most current appearance of known identities, this might explain why the image averaging technique might inadvertently make them less optimal for recognition compared to instances.

Another, more theoretical, explanation concerns the nature of our mental representations of familiar identities. Current theories and face models favour the idea that an average (prototypical) representation is abstracted from all our encounters with a particular identity over the idea that each encounter is individually stored (e.g., Bruce & Young, 1986; Burton et al., 2011). Lower likeness ratings of averages, shown here as well as previously (Ritchie et al., 2018), do not necessarily oppose these theories but they do show that the way averaging has been operationalised thus far might be too simplistic. For example, our data show that people seem to be particularly attuned to pick up and incorporate cues to natural chronology when updating their representations of familiar identities. Even with the weighting applied here, instances were perceived to be a better likeness and were recognised faster – this could be due to the superficial differences to image quality discussed earlier.

There are precedents to the idea that abstraction of representations is not based on a simple global average of encounters. For example, Kramer, Jenkins, Young, et al. (2017), used soap operas to teach participants new faces upright, inverted or contrast reversed. A simple statistical model might predict that faces would be best learned in the

format encountered, and that test performance would follow learning format. But this was not the case – faces were only learnable in their natural, upright orientation, and not when shown in unusual formats. Furthermore, this was a *visual* effect – participants could learn about the soap opera characters, and answer questions about the plot regardless of how it had been presented, they were just unable to learn the faces in non-natural presentations.

The present findings cannot discount an account where information about each instance is stored though, again, they suggest that if this is the case, instances likely follow a more complex hierarchical structure (e.g., more recently taken instances are prioritised over past ones). Recently, Balas et al. (2023) also raised the possibility of multiple prototypes/averages being formed for a single familiar identity – an account that they describe as a hybrid between norm- and instance-based encoding. The data we present here could be consistent with this model if these prototypes follow a specific hierarchy that could account for the strong bias towards someone's current appearance. Discriminating between these different encoding strategies is beyond the scope of the present work, however, we show that regardless of the exact mechanisms, the structure or operationalisation of our mental presentation/s of familiar identities is more complex than previously thought.

Finally, it is also possible that merely asking participants to rate face images for likeness might be interpreted as asking them to indicate how closely an image resembles their current appearance rather than the observer's mental representation of this person. Such an interpretation aligns well with our findings as 2020s instances and 2020s-weighted averages show the most accurate depictions of the way characters look at present. In fact, some of the comments made by participants who reported certain images being a bad likeness in Hay et al. (1991) referred to age and current appearance (e.g., "she's much younger in that photograph" or "her hairstyle's changed since"). This further strengthens the idea of a different interpretation of likeness by participants as well as our finding of the strong interest in the most up-to-date appearance of familiar identities. However, even if this interpretation could have artificially increased likeness ratings for the most recently taken instances and the averages weighted towards them, it is reassuring that these images were also recognised quicker in the name verification task.

### 6.2. Methodological constraints

The approach taken in the present research aimed to explore long-standing representations of faces using behavioural methods. While our stimuli incorporated natural changes in age, participants in Experiment 2 were exposed to learning stimuli over the course of a few days rather than over a period of 20 years. This contrasts with Experiment 1 where the characters had been encountered over changes in chronological age that were equivalent to the changes in chronological age of the participant. What is more, it is possible the strength of a perceiver's memory plays an important part that cannot be accurately represented in a lab learning study, compared to the more realistic and natural learning over a longer period of time. Importantly, despite differences in learning duration (days vs. years) the pattern of results across Experiment 1 and 2 is consistent.

It should also be noted that some of the characters were children/adolescents in the images from the 2000s. While the visual effects of aging are more pronounced between childhood and young adulthood than between young adulthood and middle adulthood, it is also the case that some faces change more with age over the adult lifespan (see Mileva et al., 2020). We capitalised on this in the design of Experiment 2, selecting characters whose appearance changed the most with age (based on pilot data) to examine the effect of aging when learning new faces. Experiment 2 therefore included 2 out of 5 characters who were children/adolescents in the 2000s images. While it could be argued that own-age biases in face recognition (see Rhodes & Anastasi, 2012) may have influenced our results, participants were highly accurate at learning the faces (98.1% accuracy in the third learning phase). Any effect of an own-age bias on our results is therefore likely to be minimal.

### 6.3. Conclusion

The results described here represent an ongoing challenge to understand the process of representation formation in face recognition. We learn new faces throughout our lives, and familiarity confers many advantages on face processing (Johnston & Edmonds, 2009). It is therefore of fundamental importance to understand the processes by which a face moves from being unfamiliar to familiar. Here, we have demonstrated that this process may be more sophisticated than previously imagined.

## Author note

The data underlying the reported analyses as well as the preregistrations for all four experiments can be accessed on the OSF: https://osf.io/e8vd4

## CRediT authorship contribution statement

**Sarah Laurence:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **A. Mike Burton:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Camilla Düring:** Software, Resources, Project administration, Methodology, Investigation. **Jennifer Pink:** Writing – review & editing, Software, Resources, Project administration. **Lucy Wilson:** Writing – review & editing, Software, Resources, Project administration, Investigation. **Mila Mileva:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Acknowledgements

MM's contribution to this work was supported by a British Academy Postdoctoral Fellowship (PF20\100034). SL's contribution to this work was supported by an Economic and Social Research Council New Investigator grant (ES/R005788/2) and funding from the Open Psychology Research Centre. Thanks to Ben Herbert-Düring for support with programming Experiment 1.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2026.106555.

## References

Baddeley, A. D., & Hitch, G. J. (1977). Recency reexamined. In S. Dornic (Ed.), *Attention and performance VI* (pp. 647–667). Routledge.

Balas, B., Sandford, A., & Ritchie, K. (2023). Not the norm: Face likeness is not the same as similarity to familiar face prototypes. *i-Perception, 14*(3), Article 20416695231171355. https://doi.org/10.1177/20416695231171355

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2021). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition, 208*, Article 104341. https://doi.org/10.1016/j.cognition.2020.104341

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*(1), 105–116. https://doi.org/10.1111/j.2044-8295.1982.tb01795.x

Bruce, V. (1994). Stability from variation: The case of face recognition the MD Vernon memorial lecture. *The Quarterly Journal of Experimental Psychology, 47*(1), 5–28. https://doi.org/10.1080/14640749408401141

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruck, M., Cavanagh, P., & Ceci, S. J. (1991). Fortysomething: Recognizing faces at one's 25th Reunion. *Memory & Cognition, 19*(3), 221–228. https://doi.org/10.3758/BF03211146

Burt, D. M., & Perrett, D. I. (1995). Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the*

*Royal Society of London, Series B: Biological Sciences, 259*(1355), 137–143. https://doi.org/10.1098/rspb.1995.0021
Burton, A. M., Jenkins, R., Hancock, P. J., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256–284. https://doi.org/10.1016/j.cogpsych.2005.06.003
Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology, 102*(4), 943–958. https://doi.org/10.1111/j.2044-8295.2011.02039.x
Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*(1), 202–223. https://doi.org/10.1111/cogs.12231
Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: A comparison across image formats. *Vision Research, 41*(24), 3185–3195. https://doi.org/10.1016/S0042-6989(01)00186-9
Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to loftus and massons method. *Tutorial in Quantitative Methods for Psychology, 1*(1), 42–45.
Craw, I. (1995). *Cognitive and computational aspects of face recognition* (pp. 201–221). https://doi.org/10.4324/9780203428979
Deffenbacher, K. A., Vetter, T., Johanson, J., & O'Toole, A. J. (1998). Facial aging, attractiveness, and distinctiveness. *Perception, 27*(10), 1233–1243. https://doi.org/10.1068/p271233
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
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.
Geng, X., Zhou, Z. H., & Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(12), 2234–2240. https://doi.org/10.1109/TPAMI.2007.70733
George, P. A., & Hole, G. J. (1995). Factors influencing the accuracy of age estimates of unfamiliar faces. *Perception, 24*(9), 1059–1073. https://doi.org/10.1068/p241059
Gilad-Gutnick, S., Harmatz, E. S., Tsourides, K., Yovel, G., & Sinha, P. (2018). Recognizing facial slivers. *Journal of Cognitive Neuroscience, 30*(7), 951–962. https://doi.org/10.1162/jocn_a_01265
Glenberg, A. M., Bradley, M. M., Kraus, T. A., & Renzaglia, G. J. (1983). Studies of the long-term recency effect: Support for a contextually guided retrieval hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(2), 231–255. https://doi.org/10.1037/0278-7393.9.2.231
Greene, R. L. (1986a). A common basis for recency effects in immediate and delayed recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(3), 413–418.
Greene, R. L. (1986b). Sources of recency effects in free recall. *Psychological Bulletin, 99*(2), 221–228. https://doi.org/10.1037/0033-2909.99.2.221
Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*(9), 330–337. https://doi.org/10.1016/S1364-6613(00)01519-9
Hay, D. C., Young, A. W., & Ellis, A. W. (1991). Routes through the face recognition system. *The Quarterly Journal of Experimental Psychology Section A, 43*(4), 761–791. https://doi.org/10.1080/14640749108400957
Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception, 31*(10), 1221–1240. https://doi.org/10.1068/p3252
Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001
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., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods, 49*(6), 2002–2011. https://doi.org/10.3758/s13428-016-0837-7
Kramer, R. S. S., Jenkins, R., Young, A. W., & Burton, A. M. (2017). Natural variability is essential to learning new faces. *Visual Cognition, 25*(4–6), 470–476. https://doi.org/10.1080/13506285.2016.1242522
Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition, 172*, 46–58. https://doi.org/10.1016/j.cognition.2017.12.005
Kramer, R. S. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review, 124*(2), 115–129. https://doi.org/10.1037/rev0000048
Kurth, S., Moyse, E., Bahri, M. A., Salmon, E., & Bastin, C. (2015). Recognition of personally familiar faces and functional connectivity in Alzheimer's disease. *Cortex, 67*, 59–73. https://doi.org/10.1016/j.cortex.2015.03.013
Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1), Article 2515245920951503. https://doi.org/10.1177/2515245920951503
Laurence, S., Baker, K. A., Proietti, V. M., & Mondloch, C. J. (2022). What happens to our representation of identity as familiar faces age? Evidence from priming and identity aftereffects. *British Journal of Psychology, 113*(3), 677–695. https://doi.org/10.1111/bjop.12560
Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34*(1), 77–100. https://doi.org/10.1037/0096-1523.34.1.77
Matthews, C. M., & Mondloch, C. J. (2021). Learning and recognizing facial identity in variable images: New insights from older adults. *Visual Cognition, 29*(10), 708–731. https://doi.org/10.1080/13506285.2021.2002994
Mileva, M., Young, A. W., Jenkins, R., & Burton, A. M. (2020). Facial identity across the lifespan. *Cognitive Psychology, 116*, Article 101260. https://doi.org/10.1016/j.cogpsych.2019.101260
Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577. https://doi.org/10.1037/xhp0000049
Noyes, E., Parde, C. J., Colón, Y. I., Hill, M. Q., Castillo, C. D., Jenkins, R., & O'Toole, A. J. (2021). Seeing through disguise: Getting to know you with a deep convolutional neural network. *Cognition, 211*, Article 104611. https://doi.org/10.1016/j.cognition.2021.104611
O'Toole, A. J., Vetter, T., Volz, H., & Salter, E. M. (1997). Three-dimensional caricatures of human heads: Distinctiveness and the perception of facial age. *Perception, 26*(6), 719–732. https://doi.org/10.1068/p260719
Pittenger, J. B., & Shaw, R. E. (1975). Aging faces as viscal-elastic events: Implications for a theory of nonrigid shape perception. *Journal of Experimental Psychology: Human Perception and Performance, 1*(4), 374–382. https://doi.org/10.1037/0096-1523.1.4.374
Porcheron, A., Mauger, E., & Russell, R. (2013). Aspects of facial contrast decrease with age and are cues for age perception. *PLoS One, 8*(3), Article e57985. https://doi.org/10.1371/journal.pone.0057985
Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin, 138*(1), 146–174. https://doi.org/10.1037/a0025750
Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70*(5), 897–905. https://doi.org/10.1080/17470218.2015.1136656
Ritchie, K. L., Kramer, R. S. S., & Burton, A. M. (2018). What makes a face photo a 'good likeness'? *Cognition, 170*, 1–8. https://doi.org/10.1016/j.cognition.2017.09.001
Ritchie, K. L., Kramer, R. S. S., Mileva, M., Sandford, A., & Burton, A. M. (2021). Multiple-image arrays in face matching tasks with and without memory. *Cognition, 211*, Article 104632. https://doi.org/10.1016/j.cognition.2021.104632
Rogers, D., Baseler, H., Young, A. W., Jenkins, R., & Andrews, T. J. (2022). The roles of shape and texture in the recognition of familiar faces. *Vision Research, 194*, Article 108013. https://doi.org/10.1016/j.visres.2022.108013
Sandford, A., & Rego, S. (2019). Recognition of deformed familiar faces: Contrast negation and nonglobal stretching. *Perception, 48*(10), 992–1012. https://doi.org/10.1177/0301006619872059
Schneider, T. M., & Carbon, C. C. (2021). The episodic prototypes model (EPM): On the nature and genesis of facial representations. *i-Perception, 12*(5), Article 20416695211054105. https://doi.org/10.1177/20416695211054105
Sexton, L., Mileva, M., Hole, G., Strathie, A., & Laurence, S. (2023). Recognizing newly learned faces across changes in age. *Visual Cognition, 31*(8), 617–632. https://doi.org/10.1080/13506285.2024.2315813
Tiddeman, B., Burt, M., & Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications, 21*(5), 42–50. https://doi.org/10.1109/38.946630
White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied, 20*(2), 166–173. https://doi.org/10.1037/xap0000009
Wiese, H., Komes, J., & Schweinberger, S. R. (2013). Ageing faces in ageing minds: A review on the own-age bias in face recognition. *Visual Cognition, 21*(9–10), 1337–1363. https://doi.org/10.1080/13506285.2013.823139