# SALIENCY IN CONTEXT: THE EFFECT OF CONTEXT ON THE DIAGNOSTICITY OF FACIAL FEATURES

BY

MORGAN REEDY

A thesis

submitted to the Victoria University of Wellington

in fulfilment of the requirements for the degree of

Master of Science in

Cognitive and Behavioural Neuroscience

Victoria University of Wellington

2020

**Abstract**

How might faces we have learned be represented in our memory? Researchers believe that our memory for faces is based on building a robust averaged representation comprised of the stable aspects of the face (i.e., eyes, nose, mouth). However, anecdotal evidence suggests this one size fits all approach to face representations may not be correct. A new theory suggests our representation for faces is instead based on a dynamic weighting, wherein what is seen as most diagnostic during learning will be encoded to a greater extent than other features in the face. One factor that may be especially important for a weighted representation is the context in which a face is initially viewed. Dependent on the context of learning, certain features may appear more distinctive than others and therefore be deemed diagnostic and receive representational weight. The current study had participants learn four faces with one manipulated to appear distinctive in the experimental context by having a unique hair colour (Experiment 1), or eye colour (Experiment 2) compared to the other faces. Participants then completed a recognition task where the feature of interest (i.e., hair or eye colour) was either available or unavailable (i.e., bald and eye closed conditions) for recognition. Findings suggested recognition was disrupted when the diagnostic feature was unavailable compared to when that feature was available, across both distinctive and typical faces. Interestingly, Experiment 2 showed a distinctiveness performance advantage compared to Experiment 1, most likely because neighbouring features may be more diagnostic than others during recognition. In addition, further exploratory analysis showed the order of the test could further affect what was encoded.

**Acknowledgements**

Firstly, to my supervisors, Dr. Tirta Susilo and Dr. Christel Devue, thank you for your guidance throughout the past two years. You have been great mentors through my studies and have always had helpful insights whenever I have needed them. It has been an honour to work in your lab.  To Christel specifically, you have helped me grow so much as a student and researcher. I have gained so much from being your student. I am grateful for the knowledge, skills, and opportunities I have had because you have been my supervisor over the past two and a half years.

Secondly, I'd like to thank all the members of the Vision and Cognition lab for all their support, encouragement, and understanding throughout the past year. You have all been so helpful over the past two years in my development as a researcher. Special thanks to Hazel, Chelsea, Maya, and Max, I could not have done this without your encouraging words.

Thirdly, I'd like to thank my CBNS classmates. Special thanks to Andre and Stephanie for their encouraging words, helpful comments, and support throughout my Master's.

Fourthly, thank you to my family and friends. You have always been there for me when I have needed you and have been so supportive throughout my Master's.

Lastly, I would like to acknowledge the Victoria University of Wellington Master's by Thesis scholarship with which I would not have been able to financially support myself throughout my Master's.

# Table of Contents

## List of Tables

**List of Figures**

## Saliency in context: The effect of context on the diagnosticity of facial features

Face recognition is not an easy feat. Face recognition is challenging because faces have identical basic structures and can only be discriminated against by differences in features and facial configurations (Young & Burton, 2017). So, how are we able to expertly recognize individuals we are familiar with from strangers on the street?

### 1.1. Qualitatively Different processes? Familiar and Unfamiliar Face processing

For the past several decades, researchers have examined the difference between unfamiliar and familiar face processing. For the most part, researchers argue that familiar and unfamiliar face processing are qualitatively different processes with familiar face recognition being superior to unfamiliar face recognition (Bruce et al., 2001; Burton et al., 1999; Clutterbuck & Johnson, 2002; Klatzky & Forest, 1984).

Compared to unfamiliar faces, familiar face recognition is not as easily disrupted by changes in lighting, appearance, or image distortion (Andrews et al., 2015; Bruce et al., 2001; Burton et al., 1999; Hancock et al., 2000). Additionally, familiar face recognition is viewpoint independent (i.e., recognition is not reliant on viewpoint remaining the same between learning and test[1]). By contrast, unfamiliar faces are often not recognized when viewpoints change from learning to test (Longmore et al., 2008). Instead, unfamiliar face recognition may rely on a more simplistic image or pictorial processing (e.g., matching multiple images from viewpoint, lighting, hairstyle) than on more face-specific processing (e.g., image independent) (Megreya & Burton, 2006).

---

[1] Learning and test refer to stages in a recognition paradigm: usually with learning referring to a study phase where participants try to memorize a set of faces and with test referring to a second phase of the experiment where participants are tested either through a recognition, identification or other task to test their abilities. Learning and test will be used throughout the thesis to describe any experimental design that includes a learning phase (study of a set of faces) and a test phase (where memory for those faces is measured).

In addition, for familiar faces, different exemplars of a given face can be easily recognized compared to unfamiliar faces, even if quite physically different. For example, when participants are given a set of images and asked to divide the set of images into several faces, participants tend to mistake the images as featuring about eight different identities rather than the two identities displayed (Jenkins et al., 2011). However, when participants were familiar with the same two identities, they successfully sorted the images into two sets of identities (Jenkins et al., 2011). Additionally, when initially learning a face, if hair is changed or removed in between learning and test phases, it significantly impairs recognition performance for these faces (Bartel et al., 2018; Toseeb et al., 2012).

Of interest, an individuals’ ability to match unfamiliar faces does not predict their ability with familiar face matching (Megreya & Burton, 2006); suggesting unfamiliar faces may be processed and viewed differently than familiar faces for face matching tasks. Additionally, matching performance may be another indication of the qualitative differences between unfamiliar and familiar face processing. For example, the accuracy of familiar face matching differs in performance between upright and inverted images, whereas this is not the case with unfamiliar faces (Megreya & Burton, 2006). Since inversion disrupts holistic processing of faces (which has long been associated with face-specific processing); identification of a face should substantially decrease when a face is inverted.

The evidence for lack of an inversion effect with unfamiliar faces might suggest that the faces are not always processed similarly to familiar faces. Evidence suggests familiar face recognition relies on much more ‘holistic’ or configural processing (Maurer et al., 2002; Rossion, 2008) than unfamiliar face processing which may be more featural (Megreya & Burton, 2006). Unfamiliar face processing may also rely less on the spatial relations of features than familiar faces, for example, when sequentially matching faces (Lobmaier & Mast, 2007; Ramon, 2015). In sum, it seems while unfamiliar face processing may be more image dependent and feature-based; familiar face processing may be more holistic and is image independent. Explanations of these differences and how the transition occurs between the two processes are discussed below.

**1.2. Current Theories**

Current theories predict that, because variation (in lighting, viewpoint, appearance) harms recognition for unfamiliar but not familiar faces, then variation must be crucial in the development of face familiarisation and building robust face representations. Below we will discuss two key models that have focussed on variation, the pictorial coding model, and the averaging hypothesis model.

In the pictorial coding model, the recognition of familiar faces requires comparing the image of a face with previously-stored instances of that face from memory (Longmore et al., 2008). As you begin to learn a face, you store individual instances of that face in your memory. Over time with the build-up of multiple stored instances of a face, recognition becomes easier and more robust. Additionally, because multiple exemplars are available during recognition, your representation is more resistant to additional variation in newly encountered images of a face.

Another model, the averaging hypothesis, suggests exposure to variation (in viewpoint, appearance, context, etc.) over time facilitates learning (Burton et al., 2016; Jenkins et al., 2011; Jenkins & Burton, 2008; Murphy et al., 2015). Data suggests learning through multiple unique exemplars of a face aids recognition over multiple similar exemplars, including when learning contains 3-D structural information (Baker et al., 2017; Dowsett et al., 2016; Menon et al., 2015; Menon et al., 2018; Murphy et al., 2015; Ritchie & Burton, 2017; Robins et al., 2018). Viewing variation when learning would help individuals to focus on the invariant aspects of the face (e.g., internal features like eyes, nose, mouth) and ignore variable aspects (e.g., external or peripheral features like hair, clothing). Focusing on these invariant features then aids in building a robust averaged memory representation (Murphy et al., 2015). These representations may be based on the same internal features for each face. This averaged representation presumably will get better with the addition of variation as more exemplars leads to more refinement of the average. In addition, the representation will become more robust with the increased variation leading to greater strength (i.e. more efficiency) in the derived average (Jenkins & Burton, 2008).

The averaging hypothesis may explain the representation of many faces in our memory; however, it does seem that the memory representations of at least a subset of faces are not based on these inner invariant features. For example, when researchers removed ‘iconic’ features from celebrity images (e.g., Cindy Crawford’s mole, the Pope’s hat, Andy Warhol’s hair), recognition rates substantially dropped. Recognition performance decreased even when the feature only covered a small portion of the actual image (for instance; .08% for Cindy Crawford’s mole) (Carbon, 2008). Additionally, in Ellis and colleagues’ (1979) seminal study on external and internal feature recognition, the majority of the celebrity faces within the set were recognized better by internal features. However, two celebrities were as well recognized by external features as internal features alone. Interestingly, even familiar face recognition can be disrupted by superficial changes in appearance (Devue et al., 2018; Sinha & Poggio, 1996). If for familiar faces, we mostly relied on inner features for recognition, then these changes to external or superficial features should not impair our recognition. It seems that external features may play a more significant role in our face memory representations than the current theories propose, even if these features may potentially change and be less stable than other invariant internal features. While the averaging hypothesis can explain how changes in viewpoint, lighting, expression, and tilt information can be ignored when familiarisation occurs; it does not account for the role of external features on familiar face recognition. It may be that during initial learning, external information along with internal information is encoded based more on what features are seen as most diagnostic for that face rather than averaging out all information that is variable when viewing a face (Bartel et al., 2018; Ellis et al., 1979).

### 1.3. A new theory – Weighted Representation Model

Although the averaging hypothesis may explain representations for many faces; it does not account for the behavioural differences found in the studies mentioned above. A new theory suggests that rather than a common representation of the same set of features (as in the averaging hypothesis), our face representations may rely on a more dynamic representational weighting, based on how diagnostic facial features or combinations of features are for an individual face during learning and encoding (Devue, in prep).

To determine what aspects are most diagnostic for representations, two factors are important: 1) Stability of facial features: the variance in changeable aspects of a particular face (e.g., hair, facial hair, etc.) relative to invariant aspects (e.g., eyes, nose, mouth) and 2) Distinctiveness: the salience[2] of one or several facial elements within the face.

Stability relies on multiple encounters to determine what aspects remain stable and what aspects vary over time within an individual. By contrast, distinctiveness relies on comparing features with previously experienced faces. Features that are viewed as unusual from one's previous experiences will be viewed as salient and attract more attention than less unique features (Theeuwes, 1992). As distinctiveness is based on personal experience with faces, individuals may differ in what features are found to be most diagnostic (see also Abudarham & Yovel, 2016).

[2] While the previous literature has been unclear with its definition of saliency (other than that of a face standing out in comparison with other faces), we discuss saliency and distinctiveness at the featural level and interchangeably. In other words, a feature would be salient if it is unique or stands out in comparison with the same features in other faces.

Within the representation, salient features will have increased weighting to the detriment of less salient features[3]. Previous research shows recognition deficits when small aspects of celebrities are moved (i.e., Cindy Crawford's mole; Carbon, 2008). Additionally, for faces with a distinctive feature, there should be more weighting towards the unique feature itself compared to other less salient aspects of the face. As there will be more encoding and a more refined representation for that distinct feature (based on the greater weighting of that feature), then the representation of the face will be less holistic (and much more feature-based) than for other faces low in distinctiveness.

In addition to factors like stability and distinctiveness, parsimony potentially contributes to our weighted representations. Parsimony refers to our cognitive mechanisms attempting to be as cost-efficient as possible by using as few resources as possible for processing (Basset et al., 2009). To ensure parsimony, our representations must be as cost-efficient as possible by encoding only the most necessary information for recognition to occur (Devue, in prep).

To be cost-efficient, coarse diagnostic features (i.e., hair colour, hairstyle, gross inner facial configurations) will be encoded before fine-grained information (i.e., details of individual features) if it is enough for recognition to occur (Devue, in prep). Coarse-to-fine representations are present in other perceptual processes and therefore may be used for faces as well (see Morrison & Schyns, 2001; Oliva, 2005). With a coarse-to-fine strategy, you may expect, over time, for the representation to focus more on encoding fine-grain information for a face if it is helpful for recognition. However, for some faces, coarse information may be enough for recognition, and therefore over time, the representation will remain coarse and not undergo as much refinement.

How the coarse-to-fine strategy is implemented is dependent on levels of distinctiveness and stability. The type of initial representation is dependent on the distinctiveness of a face. If a face appears to have a salient feature when first encountered, that salient feature will most likely be encoded. Over multiple encounters, stability begins to play a role. If a person's appearance changes and a feature that was once salient is not always available (e.g., large ears now covered by a long hair-cut), then what is encoded in the face may change. For faces where there is a salient feature but whose appearance changes, the representation of that face will change over time to a much more refined representation based on the finer details of the stable aspects of that face. Comparatively, faces with a salient feature that is quite stable over time will have a coarser representation of the face based on the singular salient feature rather than a representation of the finer details of the face. For faces that are not distinctive, the representation is dependent on how stable the face is over multiple encounters. If the less distinctive person is stable over multiple encounters, a coarser representation (i.e., without the fine details of the face) can be enough for recognition. Whereas if the less distinctive person is highly variable in appearance, the representation must become finer-grained and rely on refining the representation to features that remain stable over time and remain diagnostic (similarly to an averaged representation). In each of these cases, representations may change over time as multiple encounters lead the learner to realize what is and is not diagnostic about the face.

This refinement over time may explain why, depending on the individual face, there are differences in the literature about behavioural performance on recognition tasks. For example, the literature has discussed the influence of distinctiveness on recognition and recollection of face identities without a consensus on why it shows differing patterns of performance. Additionally, distinctiveness as a concept has been treated as homogenous with no consensus on what key features or aspects lead to it, keeping debates about what distinctiveness is unresolved.

With the weighted representation theory, we noted that feature diagnosticity impacts how our representation may be weighted. Additionally, distinctiveness and stability could influence what features we find diagnostic; however, can some features just be more diagnostic than others? Is it possible that the saliency of facial features is dependent on the features' properties in and of themselves?

### 1.4. Saliency of Features

It is apparent from the previous literature that individual faces are not always recognized based on the same set of features (Carbon, 2008; Ellis et al., 1979). Indeed, as proposed by the weighted representation model, the encoding of a face may be more dependent on what is distinctive about that specific face rather than what is distinctive about all faces in general. To be able to recognise a face, it may be important to note what is most diagnostic in that face when first encountered. There may be several factors that determine what we find diagnostic and therefore use for recognition. For example, diagnosticity may come from the initial saliency of a feature (i.e., eyes may be viewed as more informative because they differ more uniquely than other features), our overall recognition abilities, or the learning conditions that we view faces within. All of these factors may play a role in our weighting of facial information.

When it comes to the saliency of features, researchers are undecided about which features are the most informative (see Shepherd, Davies, & Ellis, 1981). This may be because of the experimental design or how participants are asked to assess feature saliency that can change how salient a feature seems. For example, when participants are asked to assess the saliency of individual facial features, eyes are often rated as the most salient cue that participants use for recognition (Laughery et al., 1971), and overall eyes are considered by many as the most important feature for recognition compared to other features (Fraser et al., 1990).

However, when participants are asked to describe a face, hair is often the most frequently cited descriptor ahead of eyes, nose, eyebrows, face shape, chin, and other features (Ellis et al., 1980). The features most frequently described when giving verbal descriptors also do not change even when a delay is added, and more memory components are required for the description (Ellis et al., 1980). These studies suggest feature informativeness can change depending on how or what is asked about features.

Additionally, the perceived saliency of features does not always indicate the actual informativeness of a feature. For example, researchers have found participants' subjective rating of feature saliency, does not always line up with what features are informative when completing an identification task (Friedman et al., 1971).

What features are most salient and informative is also dependent on the experimental design. In recognition paradigms, features in the upper half of the face tend to be better recognized alone than the lower half, and this remains when only features (viewed alone and not within a face) from the top or bottom half are viewed as well (Fisher & Cox, 1975). When only individual features are used in recognition tasks, participants tend to recognize the eye region more readily for familiar faces than other features (Goldstein & Mackenberg, 1966). However, other studies find features other than the eyes may also be similarly good cues to identification (such as the mouth) than others (such as the chin, or nose) (Ruiz-Soler & Beltran, 2012). In paradigms where only certain facial information is given, certain features can become more informative than in the other recognition paradigms mentioned. For example, nose information may not be diagnostic on its own, but when added to eyes, this information increases the effectiveness of recognition compared to the omission of the nose region in a full-face image of familiar faces (Fisher & Cox, 1973).

More recently, Abudarham and Yovel (2019) had participants identify faces, and block by block, the faces' features would be individually modified (through small increments of physical change) until identification could be made. The researchers found that identification of an unfamiliar face was more greatly affected by changes in what they termed High perceptual sensitivity features (e.g., lip thickness, hair, eye colour, eye shape, and eyebrow thickness) than Low perceptual sensitivity features (e.g., mouth size, eye distance, face proportion, skin colour, and nose shape). Interestingly, when they replicated this research using familiar faces, the same sets of features were most important for identification, matching, and recognition (Abudarham et al., 2019). Abudarham et al.'s (2019) findings indicate that certain features of the face may be more diagnostic for recognition than others and that diagnosticity may not be modulated by levels of familiarity contrary to the averaging hypothesis in which you would expect that as a face becomes more familiar, the representation would focus on the inner features of the face and disregard information about other features.

In addition to experimental paradigms and initial feature saliency, feature saliency is also dependent on individual differences in recognition abilities. The identification of features is influenced by development with more reliance on eye regions for recognition as age increases and those with developmental deficits such as in Autism using the mouth region more readily for recognition (Langdell, 1978). However, to the average person, certain features may seem to be more salient and be viewed as more useful cues for recognition than other features (which may seem less informative like the nose) (Seamon et al., 1978).

Therefore, it seems multiple aspects can affect how salient a feature is considered. Factors such as experimental design, individual preference, and perceived saliency influence how features are viewed and used. What makes a feature diagnostic is, therefore, dependent on these factors. Also, distinctiveness and stability will play a role in how diagnostic a feature is considered. Stability relies on long term viewing of a face, whereas saliency can determine the distinctiveness of a face from initial viewing. But how do we actually characterize distinctiveness? What properties of a face are distinctive, and how might that affect our weighted representation?

### 1.5. Distinctiveness as a concept

Within the weighted representation theory, distinctiveness is a key factor for representations; however, the literature is unclear on what distinctiveness entails. Several theories have tried to explain the theoretical background for distinctiveness, and these theories will be explored below. However, distinctiveness may be more influenced by factors outside of just inherent distinctiveness in a face. Instead, context may play an important role in how we determine distinctiveness.

Throughout the literature, researchers have found an advantage for the recognition of more distinctive faces over those that are viewed as more typical (see Valentine, 1991). This distinctiveness advantage is shown for both unfamiliar and familiar faces (Bartlett et al., 1984; Light et al., 1979; Valentine & Bruce, 1986a, b) and even occur with delayed testing of up to 6 weeks (Metzger, 2006).

Valentine and Bruce (1986b) wanted to know if distinctiveness could be explained by familiarity using a speeded recognition task. Participants unfamiliar with the set of faces were asked to rate the faces on levels of distinctiveness (as defined by a face that would stand out in a crowd). Afterwards, a second group previously familiar with the faces were presented with a speeded familiarity decision task; both distinctiveness and familiarity separately predicted reaction times as decreased reaction times were associated with increased distinctiveness or familiarity. However, this distinctiveness advantage is not found in all face processing mechanisms. While there is an advantage in tasks that require learning and then recognizing a face, face categorization tasks do not show the same advantage. For example, a typicality advantage for speeded classification is found when the task is face categorization (i.e., participants must distinguish which is a face between jumbled faces and intact faces) (Valentine & Bruce, 1986a). Why does this recognition advantage for distinctive faces exist and what mechanisms might be involved?

### 1.5.1. Bruce and Young's PDP Model

One explanation for the recognition distinctiveness advantage may come from an information-processing model of face recognition first outlined by Bruce and Young (1986). Bruce and Young (1986) argued there might be face recognition units or FRUs. Each familiar face has an individual FRU, and when a face is encountered, each FRU works by signalling the resemblance between a stimulus and the stored representation of a face (similar to template matching). FRUs mediate the structural encoding of the physical appearance of a face while it coordinates with person identity nodes (PINs) which hold the semantic information to the identity. When a face is seen, a structural code is derived and matched to previously-stored representations. If this newly derived face matches previously-stored representations, then the FRU will fire and allow PINs to send semantic information about a particular face. Young and Ellis (1989) proposed that the distinctiveness advantage is due to the signal-to-noise activation ratio of FRUs. They proposed that typical faces which by definition are similar to many other faces should activate many FRUs while distinctive faces should activate fewer FRUs because there will be fewer physically similar familiar faces competing for activation. Distinctive faces would have a higher signal-to-noise ratio allowing for faster and more accurate recognition than typical faces and thus accounting for the recognition advantage found in behavioural data.

### 1.5.2. Multidimensional Face space

Additionally, Valentine (1991) argued that each representation of a face that we have learned would be placed into a multidimensional space (termed 'Face space'). One version of

Face space (norm-based coding) relies on each face being placed within the space based on how physically similar it is to an averaged or normed face. This average face is meant to be a prototypical example of a face based on prior experience with faces. Face space can account for race, age, gender, and distinctiveness effects found in face recognition tasks (see Valentine et al., 2016). According to this principle of norm-based coding, when we view a face, we compare it to our norm and how far the distance is from the norm face based on its physical characteristics. If a face is distinctive, it will sit farther out in face space, whereas more typical faces will be closer to the norm and will, therefore, have many other typical faces neighbouring it. For distinctive faces, there will be fewer faces surrounding it and leading to better recognition (similarly to FRUs) and a distinctiveness advantage much like the literature shows.

### 1.5.3. Context-free Familiarity

Other researchers have argued the distinctiveness advantage is due to the nature of familiarity and memorability for distinctive faces. Early on in the research, distinctive faces were thought to have an advantage because of context-free familiarity (Vokey & Read, 1992). This means that, in the case of distinctive faces, because they generally stand out compared to faces you have seen in the past, you will be less likely to mistake having seen those faces. Contrarily, for a typical face, it brings up feelings of familiarity (due to the nature of it being a face more like previously experienced faces) and therefore increases false positives.

The theories mentioned above could all partly account for the distinctiveness advantage found in recognition tasks. However, researchers have also argued this advantage is not to do with an underlying distinctiveness but to experimental designs determining how ‘distinctive’ a face is. Instead, distinctiveness may not even need to be reliant on internal facial information (i.e., facial features) but on more peripheral information (e.g., hair, the addition of glasses, etc.) dependent on experimental context or to specific features rather than overall intrinsic distinctiveness.

### 1.5.4. Distinctiveness in Experimental Designs

Researchers have tried to equate how various aspects of a face (rather than the entire face seen holistically) may influence or encourage viewing a face as distinctive. For example, Metzger and Bridges (2004) investigated the effect of eyeglasses on recognition and showed effects similar to the distinctiveness advantage (compared to non-spectacled typical faces). Participants learned a series of spectacled faces that had previously been rated on distinctiveness. Half the faces were distinctive (spectacled), and half were typical (non-spectacled) with foils for distinctive faces also having spectacles and typical foils having no spectacles. Results showed better hit rate performance (correctly responding ‘yes’ to a learned face) for distinctive faces but found higher false alarms (incorrectly responding ‘yes’ to a face as being familiar when it was not learned) for distinctive faces as well. Metzger and Bridges (2004) argue face-space would predict the found increase in hit rate for distinctive faces. However, face-space would not have predicted higher false alarms for distinctive faces (as shown in the results) because face-space would view each face as not having many neighbouring face representations, and therefore false alarms should be lower (Valentine, 1991). However, Metzger and Bridges (2004) did not control for each face’s overall underlying physical distinctiveness. The authors further argue that future research should consider that some faces have distinguishing features (such as spectacles, scars, facial hair, etc.) that may influence distinctiveness and patterns of recognition.

Furthermore, other researchers have argued that the distinctiveness advantage is due more to experimental design than to cognitive mechanisms (Hosie & Milne, 1996). For example, Davidenko and Ramscar (2005) studied the influence of distractors on the distinctiveness advantage. Using silhouette stimuli and controlling for foil stimuli to be equally similar to typical and distinctive faces, they found that the distinctiveness advantage vanished once distractor stimuli were matched to distinctive faces (Davidenko & Ramscar, 2005). Indeed, other researchers have argued that the distinctiveness advantage is primarily a result of the learning context rather than due only to the distinctiveness of a face. By conducting several experiments manipulating distractor sets in classic distinctiveness recognition paradigms, they found that having equal numbers of distinctive distractors and typical distractors greatly reduced the distinctiveness effect. Additionally, they noted a change in criterion when distinctive distractors were added at the test phase (Hosie & Milne, 1996). The effect of distinctiveness and its apparent advantage may have more to do with learning context (i.e., the properties of where/how the face is learned and recognition is tested) than initially thought.

Others have argued that typical and distinctiveness ratings themselves differ across studies and that can influence the findings and theoretical implications of distinctiveness. Wickham et al. (2000) have argued that our ratings of distinctiveness may not be an accurate indicator of the actual physical level of distinctiveness of a face. Running several different rating experiments, they found that how the ratings were set up (for example on a standard 1-7 Likert scale with 1 being typical) can skew participants' scores so that rather than viewing faces as either typical or distinctive, it can lead all faces to be positively skewed (i.e., so that most faces are shifted to the right so that they are all seen as distinctive) and have quite small differentiations in distinctiveness ratings.

#### 1.5.5. Context-dependent Distinctiveness

According to the current literature, the distinctiveness advantage results from the difficulty in discriminating typical faces due to their physical likeness and similarity (which some researchers have argued is due to face space; Valentine, 1991). However, Hosie and Milne (1996) argued that as most research has used mixed list designs for distinctive faces (i.e., where participants see a distinctive face intermixed with more typical faces), that the distinctiveness advantage is enhanced by the context of learning and test. Because the distinctive face is placed in a group, the group context makes the distinctive face seem more salient than if it was viewed on its own. Vokey and Read (1992) had previously argued that the advantage found for distinctive faces is based on memorability and that memorability leads to a greater encoding of details related to the face. Memorability is related to the specific context in which you learn a face, suggesting the context of learning might enhance distinctiveness effects.

Moreover, a distinctive face may seem even more unique within a group of more typical faces due to the Restorff effect (Hosie & Milne, 1996). The Restorff effect posits that when one object is unique within a learning context, it will stand out and therefore receive additional processing (von Restorff, 1933, as cited in Hosie & Milne, 1996). Hosie and Milne (1996) attempted to examine if this might occur with distinctive faces as well. They found that participants had a distinctiveness advantage, whereby recognition was better for distinctive faces over typical faces when both were included during the learning and recognition phases (most prominently in lower false alarm rates for distinctive faces).

Interestingly, typical face recognition did not significantly differ between typical face only lists and mixed distinctive/typical face lists. The Restorff effect would predict that typical faces should have worse recognition when distinctive faces are on the same list than when they are alone (as distinctive faces would pull attention and additional processing towards itself and away from typical faces; von Restorff, 1933, as cited in Hosie & Milne, 1996). So, the Restorff effect could not account for the fact that typical face recognition was not hurt by the mixed distinctive/typical lists. Instead, Hosie and Milne (1996) argued that the distinctiveness advantage was in part due to primary distinctiveness; the context was making the distinctive faces more unique based on the other faces within a set. Moreover, contextual information influences both recognition accuracy and response biases, with distinctive faces having a more conservative response criterion than more typical faces (Hosie & Milne, 1996).

In sum, the lack of clear definition for distinctiveness and the inconsistencies in the literature suggest context when viewing and learning a face may play a key role in determining distinctiveness. This is supported by the distinctiveness effects within the literature being based on learning distinctive faces within a group of typical faces. Context seems especially important to determine how diagnostic features of a face are and how distinctive a face may be. In the frame of the weighted representation model, the context in which you learn a face may determine what features are viewed as most diagnostic and therefore, how the initial weighted representation is made.

Faces are rarely viewed in isolation in the real world, and so contextual cues (i.e., the environment in which you see a person; such as seeing your co-worker at the office) and information could influence what facial information is encoded and used within our representations. Moreover, we often have to learn several faces simultaneously, like when we start a new position at a new company. Since we view multiple faces at once during an experiment, group contexts may be especially important for how our representations are formed.

It has been shown that the physical similarity between faces can influence how we recognize them. For example, when participants are implicitly asked to compare two physically similar faces during learning (by being asked to identify handedness of the faces), their discrimination performance significantly increases at test compared to less similar faces (Mundy et al., 2007). In addition, recognition performance is better when participants viewed similar faces simultaneously compared to successively one after the other (Mundy et al., 2007). The advantage of simultaneously viewing similar faces has been likened to other perceptual learning mechanisms. Gibson (1969) originally posited that the comparison of two physically similar stimuli allows for better stimulus differentiation processes to be activated, leading to attention being drawn to unique features rather than more common features. Mundy et al. (2007) argued that their findings provide evidence that as participants simultaneously viewed the faces, they began to ignore (or habituate to) the common features and focus on the unique features (as Gibson, 1969 had originally posited). Additionally, Mundy and colleagues (2007) argue that while the habituation may seem like a short-term process, it will lead to long-term effects on the attentional and representational weighting of the common and unique features of faces leading to better representations of these faces. Researchers have argued that by viewing physically similar objects, discrimination of said objects improves. From the evidence listed above, discrimination of two similar faces most likely allows for better recognition, and in other learning contexts, recognition should also be improved by viewing faces together.

Although there seems to be evidence for the physical similarity of faces to aid recognition, some studies find no advantage for discrimination when test arrays include physically similar faces (Jones et al., 2015), suggesting that task constraints may affect this discrimination advantage. Others have argued that experimental design plays a larger part in the discrimination advantage than initially thought. As previously described above, Mundy et al., (2007) argued stimulus presentation rather than actual physical similarity of the faces leads to better perceptual learning and discrimination.

Additionally, some researchers have found an advantage when learning a face, where a memory representation is built based on the group context—viewing a set of unique identities aids in building an averaged representation of the set that influences future recognition of an identity (Matthew et al., 2018a). This phenomenon is known as ensemble encoding. Interestingly, trying to use newly learned similar comparators (i.e., two faces you have just learned within a recognition task) to try to aid discrimination of novel face stimuli (that also share physical similarity to the faces you have learned) does not seem to work. Indeed, there is no transfer of improved discrimination (to similar stimuli) between newly learned faces and novel unlearned faces (Jones et al., 2015). In sum, based on the averaging hypothesis, you might expect that recognition is hurt by viewing faces in a group if these faces are similar in appearance (as the representations would be based on the same set of features), yet there is evidence that viewing faces within a group is helpful. It may be that the weighted representation model can account for this difference.

**1.6. Current Study**

Some of the literature shows an advantage of group learning on later recognition. It may be that group learning encourages attention to aspects of a face that differ from those around them and further enhance encoding and later recognition of a face. Gibson (1969) first hypothesized that simultaneous learning of two stimuli aided perceptual learning because it allowed for differentiation between the stimuli, drawing attention towards the features that uniquely differed between quite similar stimuli and decreased attention towards commonly shared features. It is possible that if a feature in one face is very different from the same feature in the other faces during learning in a group context, that that specific feature will be salient, receive increased attention and be used rather than be ignored for recognition.

Based on the literature above, it seems as though both the saliency of features and the nature of an individual face are factors that affect our encoding and later recognition of faces. The weighted representation model would predict that both the distinctiveness and the stability of a face will influence what aspects of that face will be encoded and used for later recognition of a face. As we have discussed previously, distinctiveness relies on assessing the differences between faces (in comparing what aspects make a face distinctive compared to another). Therefore, distinctiveness might be characterised by how we assess faces in comparison to faces in groups.

As the weighted representation model relies on the parsimonious encoding of a face, the model would hypothesize that for distinctive faces, recognition would be enhanced compared to more typical faces, but that distinctive faces' representations would be based on specific features. Due to the enhanced encoding of these specific features, other features within the face that was less salient would be less well encoded (i.e., the distinctive features would be encoded to the detriment of less salient/distinctive features). For distinctive faces, representations will be refined for the distinct feature while the other less distinct features will have only coarse encoding. For more typical faces, the representation will rely on less salient features and maintain a more refined encoding of the overall face.

In the current study, we have specifically characterized distinctiveness using group context as the factor that leads a face to become distinctive. Context plays a role in what we find diagnostic or what strategies we use when trying to learn a face. Therefore, depending on the context that we learn a face in, some features may be especially salient on one face compared to others around it, even if the feature is not salient in and of itself. For example, when one person in a room has short hair compared to the rest, hair length may become a useful recognition cue. Because of this salient feature, the face may appear distinctive, and therefore encoding and later recognition of that face may rely on that feature more than they would in other contexts.

Therefore, we assumed that context could be used to artificially manipulate the distinctiveness of a face that is not in itself 'distinctive'. We investigated how artificially making one face distinctive compared to others impacts what aspects are encoded and whether the assumptions of the weighted representation model about the nature of representations can be supported.

Over two experiments, participants learned a set of faces in which one face was made distinctive either through manipulation of hair or eye colour compared to the rest of the faces in the group (e.g., brown-haired in a group of blonde-haired individuals). Importantly, in this experiment, distinctiveness was thus contextual (i.e., based on learning context) and did not rely on natural distinctiveness of features which can be ill-defined and not easily controlled for. Recognition performance was assessed through a recognition task where images of the same faces either included or excluded the distinctive feature (through addition or removal of the feature).

We predicted that, if in the learning context, one face is characterized as distinctive in one feature compared to the rest (e.g., one blonde-haired individual in a block of three brown-haired individuals) that feature, not in itself distinctive, may increase recognition performance when it is present at test because it was diagnostic at time of encoding. For faces without a distinctive feature (typical faces), faces will be less well recognized overall than the distinctive face when the distinctive feature is present. In other words, we expected a distinctiveness advantage for distinctive faces (leading to a performance disadvantage for typical faces).

More specifically, if the learning context affects the way we form our initial memory representation of a face and that we rely on salient features, then the distinctive face should be less well recognised than typical faces when the salient feature is not present. As typical faces will not have one salient feature that is diagnostic of identity, they will be better recognized than distinctive faces at test when the salient feature of interest is absent because their representations will be based on a more refined representation of the rest of the face.

Most importantly, if the memory representation is reliant on encoding the salient feature for the distinctive face, then when that feature is not available during the recognition test, recognition rates should be substantially lower than if the feature was available. This difference between salient feature present and salient feature absent images should be much larger for distinctive faces than for typical ones.

## 2. General Methods

### 2.1. Stimuli

To create face stimuli, we used the MIT-CBCL face recognition database (Weyrauch et al., 2004). One face from the database set was chosen to be used as a template face for the later creation of artificial faces. The image showed a male Caucasian face (approximately aged between 25-30) with a neutral expression and frontal viewpoint.

After selecting the template face, the image was manipulated in FaceGen SI modeller 3.5 (Singular Inversions, 2019) to create a set of 8 faces (four learned and four unlearned faces). Each of the eight faces created from the template face was numerically as dissimilar in physical features to the others. For each face, 81 images were created differing in viewpoint (E.g. centre view, right 45 degrees from centre, and left 45 degrees from centre), tilt, and lighting. These 81 images were divided into three different types of external features; 27 images were of the faces bald, 27 with blonde hair, and 27 with brown hair. For the hair images, one hairstyle option was selected from the FaceGen catalogue of features. This hairstyle had the option of being blonde or brown in hair colour. The hairstyle remained the same across both blonde and brown hair images. All images had a grey background, and all faces maintained a neutral expression throughout the images. All images were scaled to a 455x745 pixels ratio on XnView image editor.

### 2.2. Rating Study

In the first study devoted to ratings, the individual distinctiveness ratings of the images were verified to account for individual physical distinctiveness of each face. We also collected group distinctiveness ratings to identify how distinctive each face would be perceived within the group context by participants. Finally, paired ratings were collected to select target/foil pairs for later Experiments.

For the rating studies, only bald images were used to test how distinctive the faces' facial features were on their own. For each of the faces, eight individual images were selected from the bald image condition (frontal view, with 0.45 above lighting and frontal tilt, grey background). These images were compiled to create a rectangular image with two rows of four faces on GIMP photo editor. All group images were resized to a 791x650 ratio in GIMP.

**2.2.1. Participants and Exclusions**

We recruited 69 first-year psychology students (44 women and 25 men; *Mean age* = 19.5 ± 3.1 years, range 18-37) from Victoria University of Wellington. Participants were excluded if they failed 2 out of 4 attention checks within the study. After exclusions, we had 46 participants left (30 women and 16 men) with an age range of 18-23 (*Mean age* = 18.94 ± 1.39). Informed consent was gained, and all participants were given course credits for participation. This study was approved by the Victoria School of Psychology Human Ethics Committee.

**2.2.2. Procedure**

Participants completed the study on their own personal computers, online, through the testable platform (www.testable.org). They were told that they would complete a rating task that included four phases of ratings.

Participants were told they would rate a set of faces on several attributes. Before the ratings began, they viewed each face individually on the screen for 3 s each, separated by a 500 ms interstimulus interval (ITI). This was done so participants could acquaint themselves with the individual faces before viewing them as a group (potentially changing how they view the faces within the context of the group). For each face, only one frontal viewpoint bald image was shown, and so there were eight trials altogether. Each face image was randomly intermixed within the block.

### A. Individual Distinctiveness Ratings

In the first series of ratings, participants rated each face (on a 7-point Likert scale) on its perceived distinctiveness. Specifically, participants were asked to rate each face on how likely they thought the face was to stand out in a crowd (with 1 being “Not distinctive at all” and 7 being “Very distinctive”). Each face appeared on screen until participants answered by sliding a response bar from between 1-7. A 500 ms ITI followed the rating before the next image of a face appeared; faces were presented in random order within the block. The goal of this rating was to measure how distinctive participants thought each individual face was before viewing the faces together and potentially comparing distinctiveness based on the group context.

### B. Group Distinctiveness

Participants were then instructed to view a group image of the same set of faces. For each trial, the group image was displayed with one of the 8 faces being contained in a black rectangle (8 trials total). Participants were instructed on each trial to rate the face that was contained within the rectangle. Participants again rated distinctiveness as in the single image ratings (i.e., rating each face by how much it would stand out in a crowd from 1-7). Each face was rated once with each image being randomized within the block (8 trials altogether). Images remained on the screen until participants had made a response. After an image had been rated, it was followed by a 500 ms ITI before the next rating. The goal of this rating was to measure how distinctive the participants thought the faces were within the context of the group.

After the initial group image ratings, participants were instructed to now rate each face based on how physically similar its facial features were to those of the other faces within the group. This was done to verify that each face was comparably physically dissimilar as the Facegen computation had intended. Again, for each trial (8 trials in total) the face to be rated was highlighted with a black rectangle. Images were randomized within a block with each face being rated once. After each rating was a 500 ms ITI before moving onto the next rating image, for each rating, participants responded on a 7-point Likert scale (1 being “not physically similar at all”, and 7 being “very physically similar”).

For each of the two group ratings, there were two orders that faces were positioned in the group images, and these two group images were counterbalanced across participants so to avoid a position confound, wherein participants' ratings were influenced by where the faces were presented on screen.

**C. Paired Ratings**

For the final set of ratings, participants viewed the identities in pairs. Participants were instructed to rate the two faces based on how physically similar they found the two identities’ facial features to be to each other. Two images would simultaneously appear on screen until the participant made a response, and after each pair of images, a 500 ms ITI occurred before the next trial. Each of the 8 faces was paired with all the others, giving 56 trials. Again, participants rated pairs of images using a 7-point Likert scale with the images being presented until a response was made. The goal of this rating was to compare which pairs of faces were most similar to select the target and foil pairs for the experimental recognition task.

Within each of the ratings blocks, were a set of attention checks (4 altogether) which appeared randomly intermixed within the block. An image appeared on screen with a number at the centre of a circle, and participants were instructed to slide the rating sliders to the corresponding number. These attention checks were to assure participants were paying attention while rating.

### 2.2.3. Descriptive Statistics

From the rating task, descriptive statistics were used to determine groups in which to assign individual faces for the learning task (learned vs unlearned) and establish pairings between target and foils in the recognition task. To equate mean distinctiveness and similarity of individual faces to others in each group and pairings, we used the mean ratings of Table 1 and 2. Four of the faces were placed together as the learned faces based on mean rating scores (see Figure 1). The remaining four faces were grouped as the unlearned faces (see Figure 2). Both learned, and unlearned face groups had similar mean distinctiveness ratings to make sure one group was not more distinctive than the other. After the face stimuli were categorized as learned or unlearned, one unlearned face was picked to be paired with each learned face based on the similarity pair ratings. The mean ratings were used to make sure that each target and foil pair had on average the same similarity rating (compared to the group) to equate difficulty between different versions of the task we will use during the recognition test.

*Figure 1*. The four faces selected for the learned stimuli.

*Figure 2.* The four faces selected for the unlearned stimuli.

Table 1. *Mean, SD, distinctiveness Ratings for Individual Faces*

| Face ID | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Distinctive Individual | 3.48 (*1.26*) | 3.35 (*1.35*) | 4.00 (*1.28*) | 3.35 (*1.30*) | 3.50 (*1.46*) | 3.78 (*1.50*) | 3.65 (*1.30*) | 4.09 (*1.43*) |
| Dist. Group | 3.22(1.13) | 3.11(1.39) | 3.96(1.43) | 3.46(1.49) | 4.09(1.26) | 4.04(1.40) | 3.76(1.45) | 4.33(1.58) |
| Sim. Group | 4.11(1.29) | 4.13(1.42) | 3.87(1.41) | 4.46(1.21) | 4.15(1.48) | 3.24(1.34) | 4.52(1.36) | 3.33(1.66) |

Table 2. *Paired Similarity Ratings*

| Face ID | N | Mean | Median | SD |
|---|---|---|---|---|
| AB | 46 | 5.51 | 5.75 | 1.25 |
| AC | 46 | 4.24 | 3.00 | 1.39 |
| AD | 46 | 3.96 | 4.00 | 1.43 |
| AE | 46 | 3.46 | 3.00 | 1.49 |

| | | | | |
|---|---|---|---|---|
| AF | 46 | 4.09 | 4.00 | 1.26 |
| AG | 46 | 4.04 | 4.00 | 1.40 |
| AH | 46 | 3.76 | 4.00 | 1.45 |
| BC | 46 | 4.46 | 4.50 | 1.58 |
| BD | 46 | 4.43 | 4.50 | 1.44 |
| BE | 46 | 3.90 | 4.00 | 1.70 |
| BF | 46 | 4.15 | 4.50 | 1.51 |
| BG | 46 | 4.23 | 4.50 | 1.53 |
| BH | 46 | 4.79 | 5.00 | 1.49 |
| CD | 46 | 4.98 | 5.00 | 1.27 |
| CE | 46 | 4.41 | 4.75 | 1.48 |
| CF | 46 | 4.00 | 4.25 | 1.53 |
| CG | 46 | 4.27 | 4.50 | 1.54 |
| CH | 46 | 4.27 | 4.50 | 1.39 |
| DE | 46 | 5.23 | 5.50 | 1.54 |
| DF | 46 | 3.79 | 4.00 | 1.54 |
| DG | 46 | 4.77 | 5.00 | 1.39 |
| DH | 46 | 4.64 | 4.75 | 1.46 |

## 3. Experiment 1a

The goal of Experiment 1a is to investigate the effect of context on what features are encoded and how the weighted representation for a face is created. Participants will view a series of faces with one face being distinctive in hair colour (in either a brown or blonde-haired condition). Participants will then complete a recognition task in which the diagnostic feature (hair) will be either available during recognition (in hair present trials) or unavailable (in hair absent trials). We predict recognition performance will be better for distinctive faces than typical faces when hair is available in the recognition task. In addition, we predict recognition will be worse when hair information is not available than when it is available. In addition, when hair information is not available, distinctive face recognition will be more disrupted than typical face recognition from this lack of hair for recognition.

### 3.1. Methods

#### 3.3.1. Participants and exclusions

For our experiment, we wanted to recruit 10 people per version (eight versions total) and replaced those that did not pass our attention checks, leading to the recruitment of 87 participants (*Mean age* = 19.52 ± 2.30, range 18-31, female 73, male 13, 1 gender unspecified). With seven participants excluded (based on too fast RTs and attention check failures), leaving 80 participants (67 female, 12 male, and 1 gender unspecified; *Mean age* = 19.41 ± 2.27, range 18-31 years) from Victoria University of Wellington. Participants were excluded if they failed 2 out of 4 attention checks (within the testing phase of the experiment). Participants were also excluded if their response times were too fast in either the CFMT or testing phase of the task (under 500 ms).

Informed consent was obtained prior to participating. Participants were given course credit for their participation. This study was approved by the School of Psychology Human Ethics Committee.

**3.1.2. Stimuli**

This time, four of the faces from the rating study were labelled and used as learned identities based on the distinctiveness ratings, while the other four remaining faces were kept as unlearned foils. The learned faces were chosen based on the distinctiveness ratings of the rating study so that each of the chosen faces was rated similarly on levels of distinctiveness (from both ratings of distinctiveness viewing faces alone and viewing faces within a group context) and as to equate physical distinctiveness across the four learned faces. The unlearned foil faces were also chosen based on levels of distinctiveness (based on the rating study). To avoid an uncontrolled distinctiveness advantage for learned faces (i.e., based on other facial features than those we manipulated), the unlearned faces were chosen based on similar levels of distinctiveness in the rating study to the learned faces.  Each of the unlearned faces was matched to one learned face as a foil based on the paired similarity ratings. The foils were chosen to ensure that all target/foil pairings were similarly matched in difficulty (to allow the distinctiveness manipulation to be seen).

For Experiment 1a, the four learned faces were given hair. Each of the four faces was viewed as distinctive, namely the one with distinctive hair colour (either brown or blonde hair), in one out of four versions of the task to account for any existing uncontrolled physical distinctiveness of its other facial features. To account for the fact that one hair colour may be relatively more distinctive within a group than the other (i.e., blonde hair may stand out more among other brown-haired faces than brown among blond), each individual face was distinctive in both a blonde and a brown hair version. In total, eight versions of the task were created (four blonde and four brown-haired versions). Participants were randomly assigned to one of the eight versions (10 participants per version).

For the learned faces, 324 images were created (81 per identity). The 324 images were split into three conditions (108 blonde hair images, 108 brown hair images, and 108 bald images, 27 images per identity per condition) for use in the learning and recognition phase. For the unlearned faces, 324 images (with 108 images per hair condition) were created for later use in the recognition phase. All of these images included variations in lighting, viewpoint, and tilt.

*Learning Phase*. For the learning phase, 18 images from the two hair conditions (nine blonde and nine brown-haired images) were chosen for each identity. Between them, the 18 images combined all three types of variation on three dimensions (viewpoint [frontal, left 45 degrees from centre and right 45 degrees from centre], tilt [frontal, up, and down], and lighting [above .45, left .45, and right .45]) in a Latin square fashion to make a set of images with several different combinations of their variations within the learning set. Across the two hair conditions, the same variation combinations were used (i.e., blonde-haired images had the same amount of viewpoint, lighting, and tilt variation as brown hair images) so that the only difference between the images participants saw was the hair colour. All images were resized to a 366x600 ratio to ensure the faces could be holistically viewed with their respective external feature.

In addition, to make one face within the set seem distinctive, group images were created showcasing that one face differed in hair colour compared to the rest (see Figure 3). An additional set of 8 group images was created for each of the 8 versions of the experiment. To create the group images, one frontal viewpoint (frontal tilt and 0.45 lighting from above) image of each face was selected and added to an array. Each group image contained the 4 learned faces in a horizontal row. All group images were resized to a 1055x435 ratio.

*Figure 3.* Example of stimuli from the learning phase.

*Recognition Phase*. In the recognition task, to account for participants possibly using general image processing rather than face specific processing, we used the remaining images, which are those that were not included in the learning phase. The remaining 36 images from the blonde and brown hair conditions were used here (18 per condition). In addition, 18 images with no hair (the bald condition) were included to test recognition abilities without the distinctive feature. All three image sets had the same levels of variation in tilt, viewpoint, and lighting to each other so that the only differing aspect was the presence of hair and its colour. All images were resized to the 366x600 ratio as in the learning phase.

The Cambridge Face Memory Test (CFMT; Duchaine and Nakayama, 2006) was used to measure individual face recognition abilities. The CFMT is a standardized test where participants view a set of 6 faces differing in viewpoint (frontal, left and right profile) and are instructed to study the faces. During test, an array of 3 faces (one target face and two foils) is shown, and participants must choose which face they had previously learned. Over the trials (72 trials altogether), test images increasingly differ from the learned images (i.e., the addition of digital noise, lighting changes, head orientation differences, and changes in external features).

**3.1.3. Procedure**

Participants completed the experiment through the testable platform (www.testable.org) on their own personal computers. Participants were randomly placed into one of the 8 versions of the experiment.

*Learning Phase*. At the start of the experiment, participants were instructed to learn a set of 4 faces that they would have to recognise later. Participants first viewed a group image showing the distinctive face alongside the three typical faces for 15 s. This was done to make one face distinctive by drawing participants' attention to its unique hair colour compared to the others.

Afterwards, each of the 4 faces was learned individually in separate blocks. In each block, a face was learned via 9 different images. Images were presented for 3 s before a 500 ms ITI followed by the next image. Within the learning phase, the order of each identity block was randomized, and within each of the identity blocks, 9 images were randomly intermixed.

*Recognition Phase*. After participants completed the learning phase, they completed a recognition task. In the recognition task, participants viewed two separate blocks with hair present (144 trials) and hair absent conditions (144 trials). Within the two conditions, the remaining 18 images from the original set of 27 were used for each individual identity. In the hair present condition there were 36 distinctive hair colour trials (18 images for one learned distinctive face and 18 images for one unlearned distinctive foil) and 108 typical hair colour trials (54 images total for the three learned typical faces and 54 images for the three unlearned typical foils). The hair absent condition used the same identities and had the same number of trials per face.

Participants viewed an image centrally on screen and were asked to respond with a keypress of 1 for 'yes' if it was a face they had previously studied in the learning phase or a keypress of 2 for ‘no’ if it was a face they had not studied in the learning phase. Each image appeared on screen for 3 s before disappearing and being replaced by text prompting participants to respond. The order of test blocks was randomly determined for individual participants via the counterbalance function in Testable. Before the start of each recognition block, a group image (with hair) was displayed centrally for 15 s before the block began. This was to remind participants that one face was distinctive based on hair colour. Within each block, two attention checks were randomly dispersed to assess if participants were paying attention to the task or were randomly pressing buttons. The attention checks were the same as previously mentioned in the rating study.

*CFMT*. After completing the recognition task, participants completed the CFMT.

### 3.1.4. Design and Analyses plans

Prior to data collection, the plans for the analyses were pre-registered on Open Science Framework for Experiment 1. This plan and detailed predictions can be found in [https://osf.io/sy72m].

To assess the impact of context on the saliency of external features we analysed recognition performance in a 2 (Distinctiveness: distinctive vs typical faces) x 2 (Image type: hair absent vs hair present) repeated measures ANOVA on hit rates and false alarms.

Additionally, we conducted paired samples t-tests comparing hit rates for hair absent trials versus hair present trials of distinctive and typical faces. We conducted the same set of paired samples t-tests mentioned above on false alarms.

We also conducted descriptive analyses on median reaction times (RTs) to check for speed-accuracy trade-offs.

For exploratory purposes, we conducted a 2 (Distinctiveness: distinctive vs typical faces) x 2 (Image type: hair present vs hair absent) within-subjects repeated measures ANOVA on d’ (sensitivity) and criterion c. Sensitivity is a measure of a participants’ ability to discriminate between learned target faces and unlearned distractor faces that incorporates hit rates (correct recognition of a target face) and false alarm rates (incorrect recognition of a distractor face). C is a measure of response bias, it is calculated based on participants' tendency to be liberal (tendency to respond 'yes' this is a previously learned face) or conservative (tendency to respond 'no' this is a novel face).

We did not include d' and C in the main thesis because we thought the effect of context on hits and false alarms might mean that the patterns of performance between the two would lead (i.e., higher hit rate accompanied by higher FA rate for distinctive faces compared to typical faces) to cancel out differences between faces. These measures are nonetheless visible in the appendices (see Appendix A). Moreover, the previous literature often separates hit rates from false alarms due to the differing effects of distinctiveness on target and foil faces (see Bartlett et al., 1984; Davidenko and Ramscar, 2005; Valentine 1986a).

In addition, as part of our pre-planned analyses, we conducted Pearson's correlations between CFMT scores and performance in each condition of the recognition task (distinctive hair present and absent trials, typical hair present, and absent trials for hit rate and false alarm rate). These were not included in the main thesis (but can be viewed in Appendix B) as they were for future exploratory purposes.

## 3.2. Results

### 3.2.1. Pre-planned Analyses

*Hit rate*. Results are presented in Figure 4. There was a significant main effect of image type on hit rate, $F(1, 79) = 28.54$, $p < .001$, $\eta^2 = .265$, whereby participants had higher hit rates for images with hair than without hair, but no main effect of distinctiveness, $F(1, 79) = .62$, $p = .435$, $\eta^2 = .002$. A significant interaction was found between image type and distinctiveness, $F(1, 79) = 12.35$, $p < .001$, $\eta^2 = .135$. Follow-up paired samples t-tests showed participants did not significantly differ in performance between distinctive ($M = .65$

± .31) and typical faces (*M* = .61 ± .21) for hair present trials, *t*(79) = 1.18, *p* = .242, *d* = .132. By contrast, for hair absent trials, participants were significantly better for typical (*M* = .56 ± .56) than for distinctive faces (*M* = .47 ± .30), *t*(79) = 2.58, *p* = .012, *d* = .288.

Follow-up paired samples t-tests for distinctive faces showed better performance on hair present trials than hair absent trials, *t*(79) = 5.00, *p* < .001, *d* = .559, suggesting participants better remembered the distinctive faces when they viewed them with hair than without hair. Similarly, for typical faces, participants' performance was better for hair present than hair absent trials, *t*(79) = 2.64, *p* = .01, *d* = .295, suggesting participants better remembered typical faces when hair images are shown compared to when bald images were shown. As predicted, the difference between hair present and hair absent trials was greater for distinctive faces than for typical faces, as evidenced by the weaker effect size for typical faces than for distinctive faces and the interaction between distinctive and typical faces. This suggests that participants relied on hair for recognition of distinctive faces and were hurt to a greater extent when hair was not available than for typical faces. For typical faces, participants seemed to base their recognition more on internal features of the faces for recognition to occur but were still hurt by the exclusion of hair.

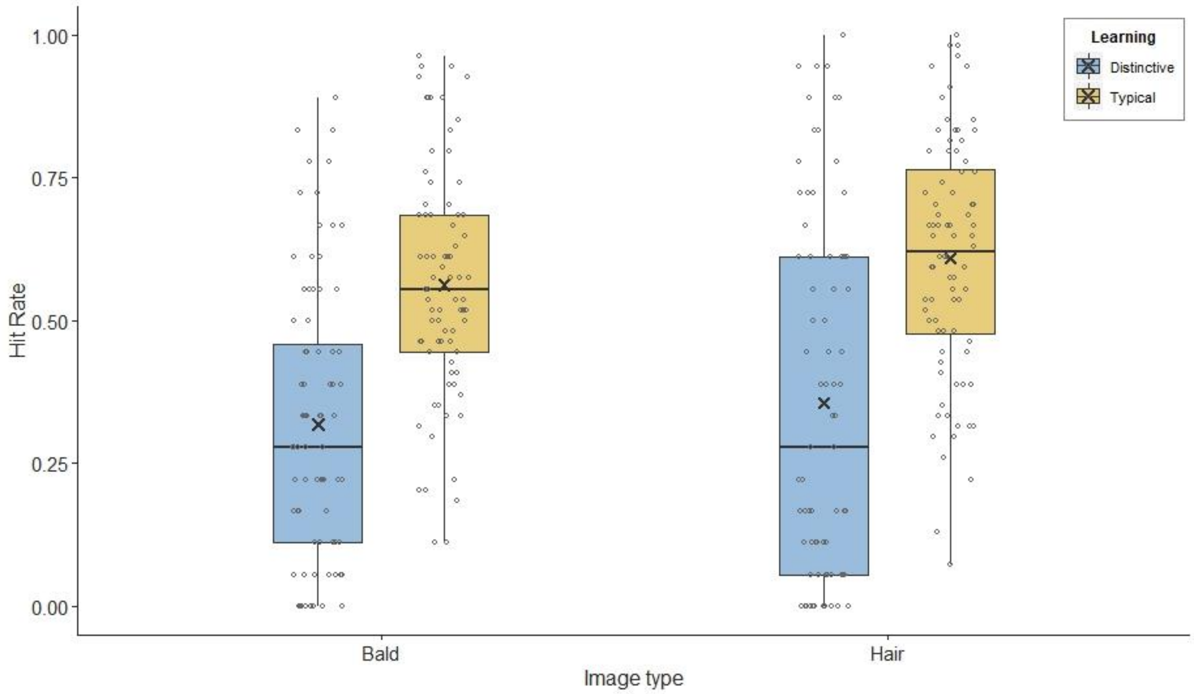


*Figure 4.* Hit rates for Experiment 1a as a function of bald distinctive/typical face conditions and hair distinctive/typical face conditions. Mean values are represented by the cross. The line within the boxplot represents the median while the boxplot itself represents the interquartile range. Dots show individual data points.

*False Alarms*. Results are displayed in Figure 5. No main effect of distinctiveness was found, $F(1, 79) = 2.58$, $p = .112$, $\eta^2 = .008$. No significant main effect of image type was found, $F(1, 79) = 1.09$, $p = .300$, $\eta^2 = .002$. There was also no significant interaction between distinctiveness and image type, $F(1, 79) = 0.95$, $p = .332$, $\eta^2 = .002$.

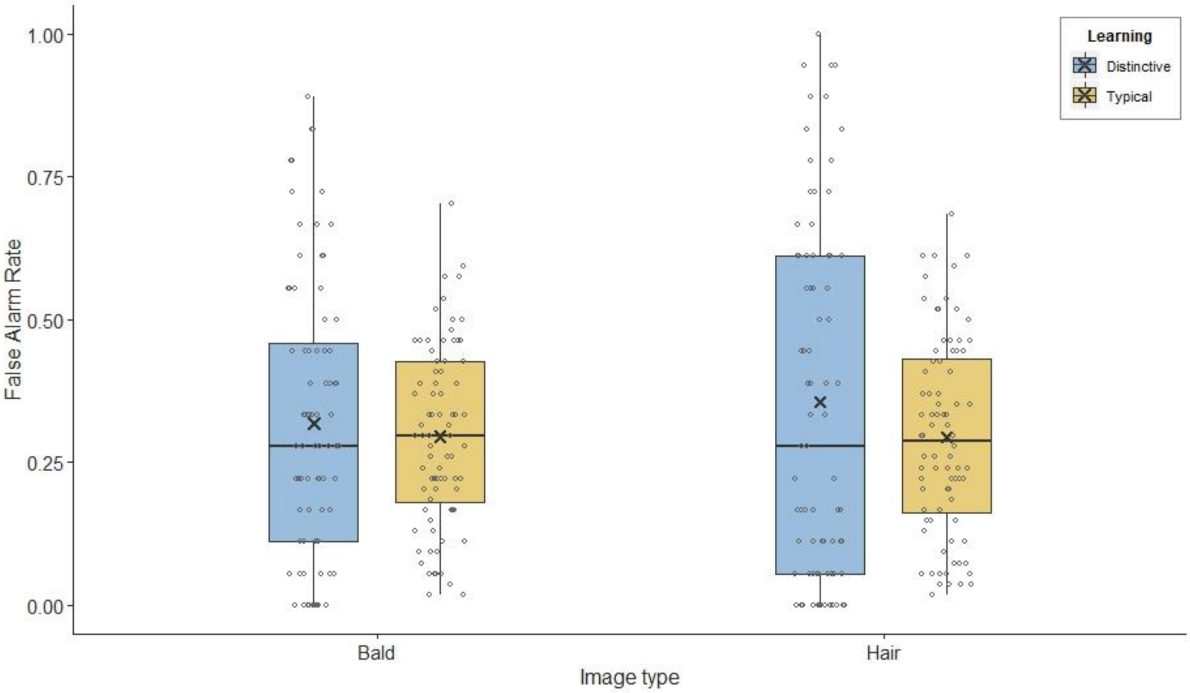


*Figure 5.* False Alarm rates for Experiment 1a as a function of bald distinctive/typical face conditions and hair distinctive/typical face conditions. Mean values are represented by the cross. The line within the boxplot represents the median while the boxplot itself represents the interquartile range. Dots show individual data points.

*Reaction times*. Median RTs are displayed in Table 3. They indicated that there were no speed-accuracy trade-offs (range = 856-949 ms).

Table 3. *Reaction times corresponding to hit rates and false alarms in Experiment 1a*

| **Median(*SD*)** | **Hit Rate** | **False Alarm Rate** |
|---|---|---|
| **Distinctive Absent** | 949(*437.73*) | 913(*440.72*) |
| **Distinctive Present** | 870(*356.91*) | 863.5(*352.85*) |
| **Typical Absent** | 869(*363.80*) | 856.5(*371.14*) |
| **Typical Present** | 884(*239.31*) | 897.25(*324.40*) |

*Note. N = 80*

**3.3.1. Exploratory Analyses**

Due to issues with counterbalancing, more participants saw one order of blocked test stimuli more than another. We decided to explore the possible effect of order block further.

To examine the possible effect of block order on our findings, 2 (Distinctiveness: distinctive vs typical faces) x 2 (Image type: hair absent vs hair present) x 2 (Order: Hair absent block first vs Hair present block first) mixed factorial ANOVA was conducted on hit rate.

For Experiment 1a, there was a significant main effect of image type, $F(1, 78) = 21.17$, $p <$ 001, $\eta^2 = .031$, but no main effect of distinctiveness, $F(1, 78) = 0.33$, $p = .578$, $\eta^2 = .001$, nor a main effect of order, $F(1, 78) = 2.85$, $p = .096$, $\eta^2 = .017$. A significant interaction was found between image type and distinctiveness, $F(1, 78) = 11.14$, $p = .001$, $\eta^2 = .014$. Additionally, a significant interaction was found between image type and order, $F(1, 78) = 6.96$, $p = .010$, $\eta^2 = .010$. No significant interaction was found between distinctiveness and order, $F(1, 78) = .33$, $p = .568$, $\eta^2 = .001$, nor a three way interaction between the variables, $F(1, 78) = .006$, $p = .936$, $\eta^2 = .000$. Results are displayed in Figure 6.

Post-hoc Tukey's honest significant difference tests revealed that when participants saw bald images first, they had significantly better performance when viewing hair images than bald images, $t(78) = 6.01$, $p < .001$, $d = .719$, whereas when participants viewed hair images first, participants did not significantly differ between hair images and bald images, $t(78) = 1.23$, $p = .610$, $d = .048$. Interestingly, participants had higher recognition rates when they saw the bald image block second compared to when they saw it first, $t(78) = 2.69$, $p = .041$, $d = .670$. On the other hand, participants did not significantly differ in recognition rates when they saw the hair image block second compared to first, $t(78) = .329$, $p = .988$, $d = .037$.

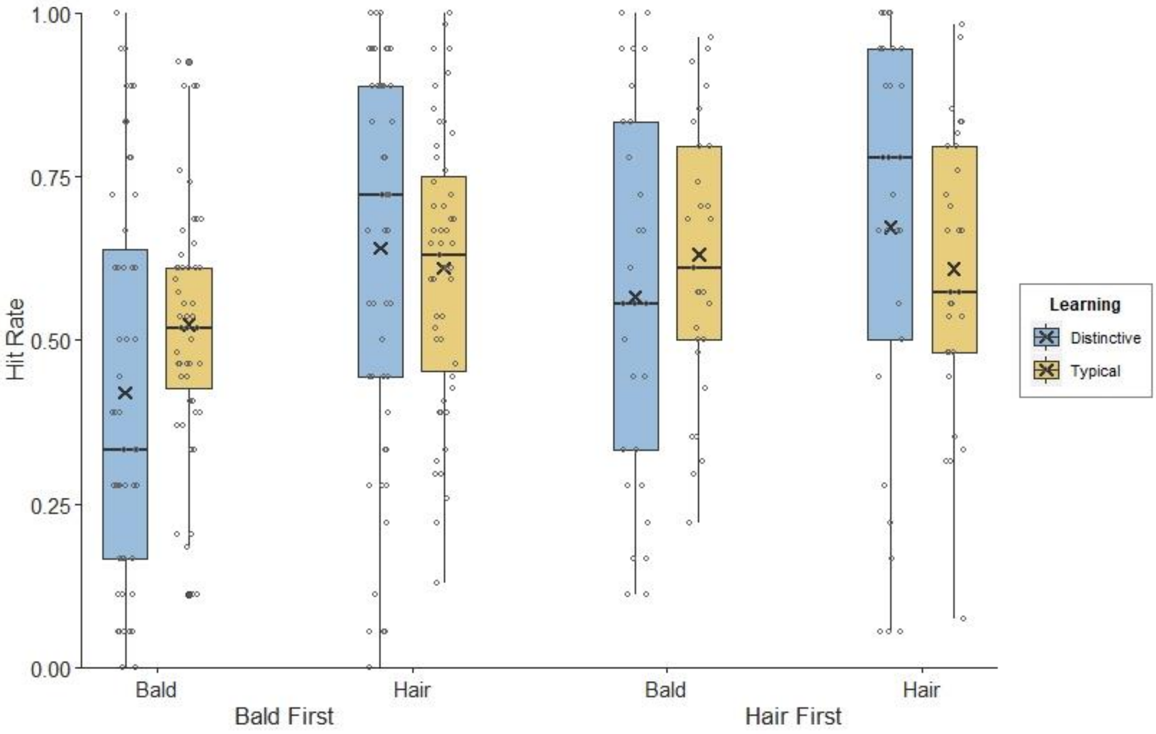


*Figure 6.* Order Effects of Experiment 1a. Hit rates are measured on the y-axis, with order (which recognition block was viewed first) and image type on the x-axis. Mean values are represented by the cross, the line within the boxplot represents the median, the boxplot itself represents the interquartile range, and the dots show individual data points.

### 3.3. Discussion

In Experiment 1a, we did not find a distinctiveness advantage overall as we expected but did find that distinctive faces recognition was more disrupted by the feature of interest being unavailable than for typical faces. In Experiment 1a, we found participants had worse recognition performance for the distinctive than typical faces when hair was not available for recognition. As predicted, the difference in impairment was larger for the distinctive faces than for the typical faces, indicating that removing the hair (i.e., the diagnostic feature) was more disruptive to recognition than for typical faces. This finding provides evidence that the learning context is influencing what we encode in a face.

As mentioned previously, hair as the distinctive feature may have led to participants only encoding very coarse information about the face (i.e., primarily hair colour) and that led to a very coarse representation of the face. This is in line with the predicted results as we expected coarse encoding of the face with a focus on the hair. In terms of representations of the weighted representation model for distinctive faces, we expect weighting to be placed on the hair to the detriment of other features (Devue, in prep). Interestingly, the encoding of hair was to the detriment of the other features in the distinctive faces. By contrast, for the typical faces, recognition was less disrupted when the bald images were presented. In Experiment 1a, participants may have encoded other aspects of the typical faces as well since the effect sizes were similar between typical and distinctive faces when hair was available compared to when it was unavailable. This encoding of the salient feature to the detriment of other features is a crucial assumption to the weighted representation model, and the present findings suggest that context plays a role in building a weighted representation (Devue, in prep), at least in initial stages of learning. Interestingly, for typical faces, weighting was still somewhat placed on hair (as there was a disruption to recognition when hair was not available for recognition).

For Experiment 1a, recognition of typical faces was less impaired between the bald images and hair images than for distinctive faces. Previous research has shown that any change of external features from learning to test can cause recognition impairment primarily due to holistic interference (Toseeb et al., 2012). Holistic interference refers to the face being recognized as a whole unit. When one aspect of the face is disrupted/changed (such as hairstyle) this then disrupts our recognition as we now view the face with this change and cannot stop viewing the face holistically. Therefore, we have issues recognizing the face with this different feature added (Toseeb et al., 2012). Even with the possible effects of holistic interference that would affect all faces, typical faces were still better recognized when hair was not available than distinctive faces, suggesting there was further encoding of the face.

The fact that there was still impairment between hair present and hair absent trials also suggests that hair was still encoded (to a degree) for the typical faces meaning that recognition was still disrupted due to holistic interference (when hair was removed for recognition). For typical faces, there was less of a reliance on hair for recognition (because there was less impairment between hair present and absent trials). Because typical faces relied less on hair, typical faces had more of a reliance on the less salient inner features of the face as we predicted. The findings also provide further evidence of perceptual learning when viewing similar comparators that is, viewing similar faces can lead to enhanced or greater processing of the differences within faces (Mundy et al., 2007). As with previous literature, the group context led to finer discrimination of typical faces when they are quite similar in appearance to one another (Mundy et al., 2007).

For Experiment 1a, while the findings do partially support our hypothesis, there are some limitations. An issue with how Testable randomizes block orders led more participants to be in one set of blocks the most compared to the other. As a consequence, the order of test images was not entirely counterbalanced and led to more participants viewing the bald face images first, before images that included hair. Our exploratory analysis showed that the block order did affect participants' recognition performance when it came to image type, with larger disruption to performance from images with hair to bald image when bald images were shown first. The consequence of this is that our findings may be interpreted as only partial evidence of weighted representations and that a further replication is necessary to confirm the pattern of results is an accurate reflection of our representation for faces. However, if the bald image block is first more often, you might expect recognition to become more reliant on inner features if it initially involves no hair information. This is because participants may become aware that multiple faces (including unlearned faces) contain the same lack of hair information and may start to focus on internal features to attempt to differentiate them.

Experiment 1b was completed to improve upon the previous experiments counterbalancing issues.

## 4. Experiment 1b

The aim of Experiment 1b was to rectify the issues with counterbalancing found in the previous experiment. In Experiment 1b, the methods remained the same as the past experiment but with 16 versions of the task so that participants were evenly distributed between the versions.

### 4.1. Methods

#### 4.1.1. Participants and Exclusions

We recruited 87 participants (*Mean age* = 19.52 ± 2.30, range 18-31, female 73, male 13, 1 gender unspecified) and excluded seven participants (based on too fast RTs and attention check failures), leaving 80 participants (67 female, 12 male, and 1 gender unspecified; *Mean age* = 19.41 ± 2.27, range 18-31 years) from Victoria University of Wellington. Informed consent was obtained prior to participating. Participants were given course credit for their participation. This study was approved by the School of Psychology Human Ethics Committee.

#### 4.1.2. Stimuli and Procedure

As with the previous experiment, each of the four faces was viewed as distinctive in a version of the task to account for any existing physical distinctiveness of the faces. To account for the fact that hair colour may be distinctive within a group (i.e., blonde hair may stand out more than brown), each individual face was distinctive in both a blonde and a brown hair version. To correct the counterbalancing issue with testable identified in Experiment 1a, each of the original eight versions of the experiment was duplicated to create 16 versions with eight of these versions having the order of image presentation reversed for the recognition task. In total, 16 versions of the task were created (eight blonde and eight brown-haired versions; eight with hair absent trials first and eight with hair present trials first). Participants were randomly assigned to one of the 16 versions (five participants per version). All other aspects of the stimuli and procedure remained the same from experiment 1a.

### 4.1.4. Design and Analyses plans

As Experiment 1b was not pre-planned (and was a replication of Experiment 1a), no pre-planned analyses were pre-registered. The same analyses as in Experiment 1a were performed, and additional exploratory analyses were conducted.

## 4.2. Results

### 4.2.1. Pre-planned Analyses

*Hit rate*. Results are presented in Figure 7. There was a significant main effect of image type, $F(1, 79) = 11.89$, $p < 001$, $\eta^2 = .029$. As in Experiment 1a, hit rates were larger when participants saw images containing hair information than when no hair information was available during the recognition task. In addition, there was no main effect of distinctiveness, $F(1, 79) = 0.80$, $p = .374$, $\eta^2 = .003$. As in Experiment 1a, participants were not better at recognition distinctive faces over typical faces. No significant interaction was found between image type and distinctiveness, $F(1, 79) = 0.19$, $p = .662$, $\eta^2 = .000$.

Despite the absence of a significant interaction, follow-up paired samples t-tests were conducted as we previously planned them in our pre-registration. Follow-up paired samples t-tests showed participants did not significantly differ in performance on hair present trials, between distinctive ($M = .70 \pm .25$) and typical faces ($M = .67 \pm .15$), $t(79) = 1.09$, $p = .279$, $d = .122$. Likewise, participants' performance did not significantly differ between typical ($M = .60 \pm .19$) and distinctive faces ($M = .61 \pm .27$) for hair absent trials, $t(79) = .45$, $p = .651$, $d = .051$.

Paired samples t-tests for distinctive faces showed participants had better performance on hair present than hair absent trials, $t(79) = 2.36$, $p = .021$, $d = .263$. Additionally, for typical faces, participants also performed better on hair present than hair absent trials, $t(79) = 3.49$, $p < .001$, $d = .390$. The above findings indicate that participants were able to recognize both distinctive and typical faces better when hair was available for recognition than when it was absent. This time, the effect sizes were larger for typical faces suggesting participants had similar levels of disruption to recognition for both typical and distinctive faces when hair was unavailable during test.

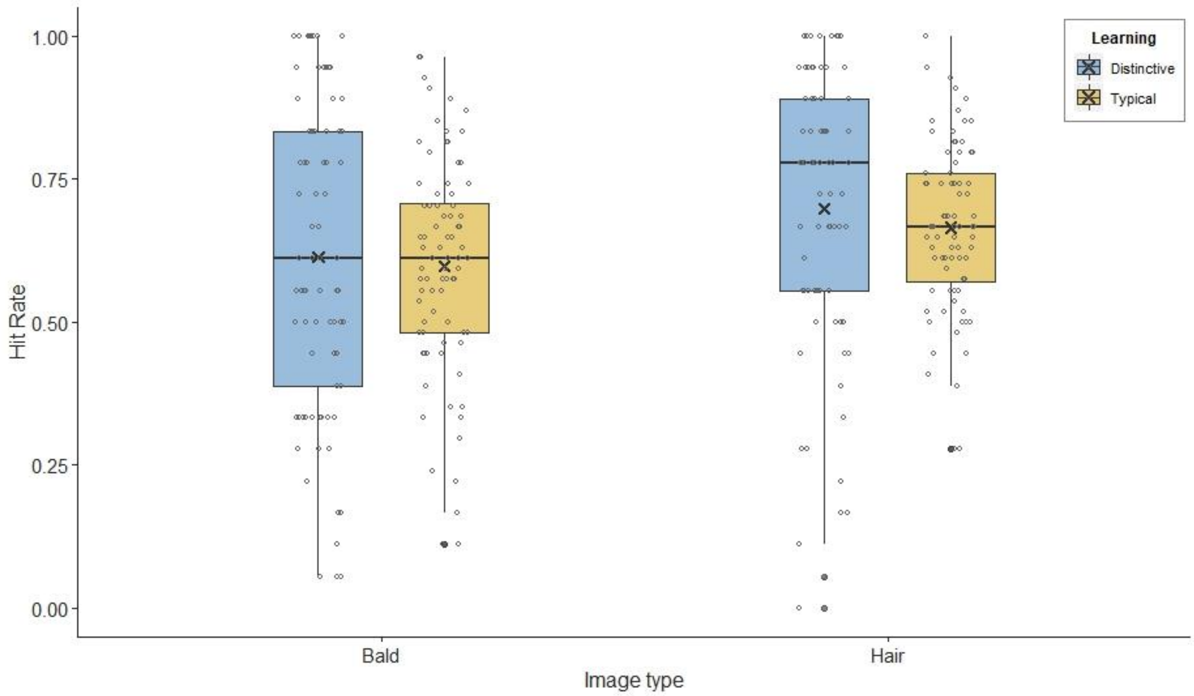


*Figure 7.* Hit Rate for Experiment 1b as a function of bald distinctive/typical face conditions and hair distinctive/typical face conditions. Mean values are represented by the cross, the line within the boxplot represents the median, the boxplot itself represents the interquartile range, and the dots show individual data points.

*False Alarms*. Results are displayed in Figure 8. There was no main effect of distinctiveness, $F(1, 79) = 0.67$, $p = .415$, $\eta^2 = .002$, whereby participants falsely identified foils as target faces for both distinctive and typical faces. However, a main effect of image type was found, $F(1, 79) = 8.72$, $p = .004$, $\eta^2 = .023$. No significant interaction was found between distinctiveness and image type, $F(1, 79) = 0.94$, $p = .336$, $\eta^2 = .001$.

Follow-up paired samples t-tests for distinctive faces showed larger false alarms for hair present ($M = .34 \pm .20$) than hair absent trials ($M = .28 \pm .18$), $t(79) = 3.00$, $p = .004$, $d = .335$, indicating participants were more likely to falsely say that a foil was a face they had seen in the learning phase when it had hair than when it did not.

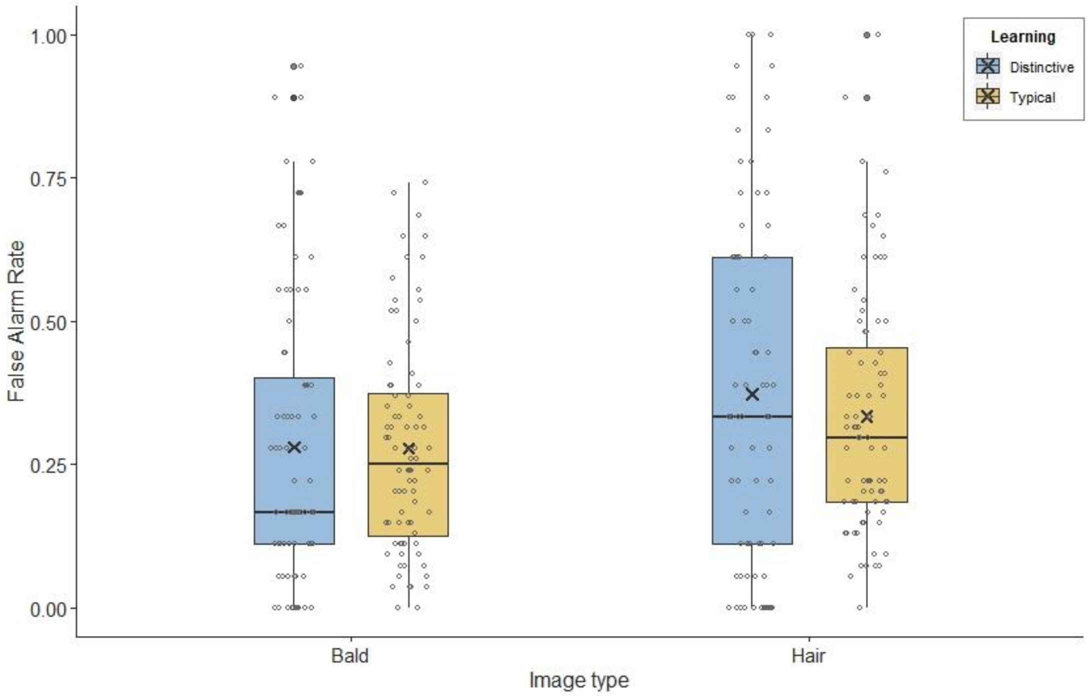


*Figure 8.* False alarm rates for Experiment 1b as a function of bald distinctive/typical face conditions and hair distinctive/typical face conditions. Mean values are represented by the cross, the line within the boxplot represents the median, the boxplot itself represents the interquartile range, and the dots show individual data points.

*Reaction time*. Median RTs indicated there were no speed-accuracy trade-offs found (range = 856-949 ms) with hit and false alarm having a similar pattern of RTs. Refer to Table 4.

Table 4. *Reaction times corresponding to hit rate and false alarms*

| **Median(*SD*)** | **Hit Rates** | **False Alarms** |
|---|---|---|
| **Distinctive Absent** | 1121.98(*443.39*) | 1078.36(*401.48*) |
| **Distinctive Present** | 1047.18(*435.31*) | 1142.44(*680.81*) |
| **Typical Absent** | 1045.44(*338.76*) | 1057.49(*405.05*) |
| **Typical Present** | 995.99(*404.60*) | 1097.31(*575.19*) |

*Note. N = 80*

**4.2.2. Exploratory Analyses**

As the counterbalancing was fixed in Experiment 1b, we decided to investigate if there was still an effect of order on our findings. As with Experiment 1a, a 2 (Distinctiveness: distinctive vs typical faces) x 2 (Image type: hair absent vs hair present) x 2 (Order: Hair absent block first vs Hair present block first) mixed factorial ANOVA was conducted on hit rate.

For Experiment 1b, there was a significant main effect of image type, $F(1, 78) = 13.53$, $p < .001$, $\eta^2 = .029$, but no main effect of distinctiveness, $F(1, 78) = .79$, $p = .377$, $\eta^2 = .003$. In contrast to Experiment 1a, there was a main effect of order, $F(1, 78) = 4.76$, $p = .032$, $\eta^2 = .020$, whereby participants had better recognition performance overall when they viewed the hair images first, followed by bald images than when they viewed bald images first, followed by hair images. No significant interaction was found between image type and distinctiveness, $F(1, 78) = .194$, $p = .661$, $\eta^2 = .000$. No significant interaction was found between distinctiveness and order, $F(1, 78) = .194$, $p = .661$, $\eta^2 = .001$. Additionally, a significant interaction was found between image type and order, $F(1, 78) = 11.85$, $p < .001$, $\eta^2 = .025$. No significant interaction was found between all three variables, $F(1, 78) = 1.71$, $p = .194$, $\eta^2 = .003$. Results are displayed in Figure 9.

As in Experiment 1a, post-hoc Tukey's tests to follow up on the significant image and order type interaction showed that when bald images were viewed first, participants had higher hit rates for the hair images than bald images, $t(78) = 5.04$, $p < .001$, $d = .852$. Similarly to Experiment 1a, when participants viewed hair images first, participants did not significantly differ between hair images and bald images, $t(78) = 0.17$, $p = .998$, $d = .101$. As with Experiment 1a, participants had larger hit rates when they saw the bald image block after viewing hair images first, $t(78) = 3.77$, $p = .001$, $d = .786$. Additionally, participants did not significantly differ in recognition rates between viewing the hair block first or second, $t(78) = 0.21$, $p = .997$, $d = .007$.

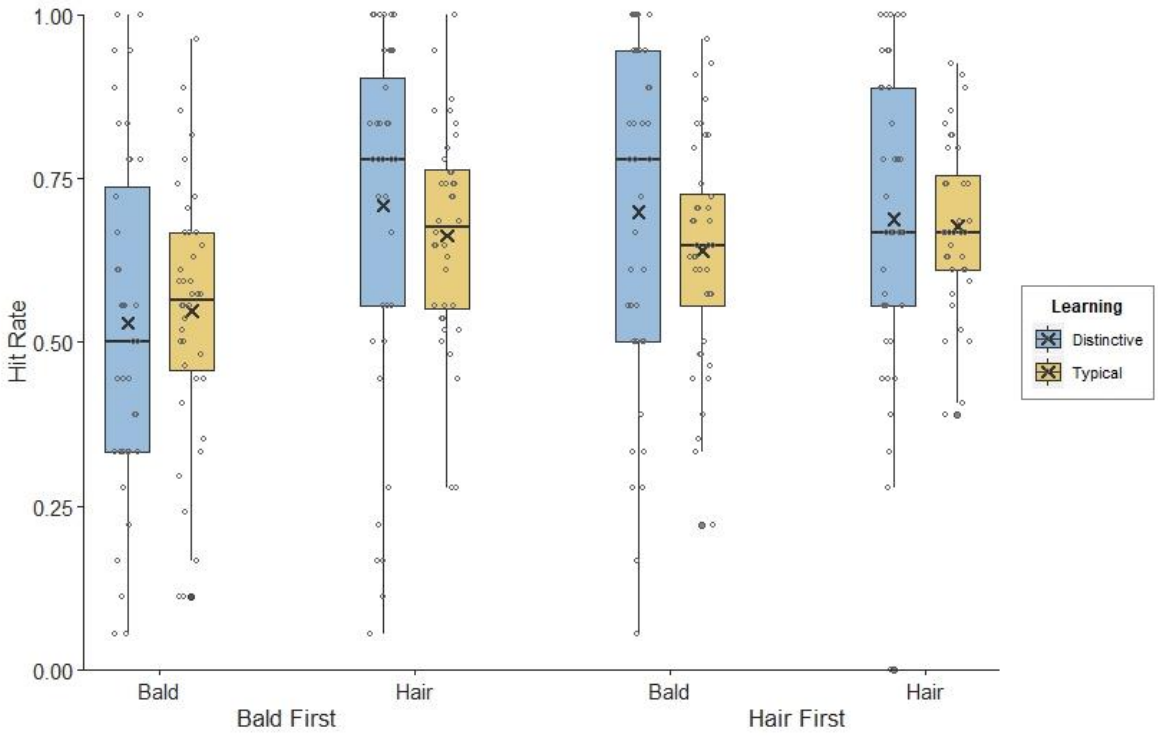


*Figure 9.* Order Effects of Experiment 1b. Hit rates are measured on the y-axis, with order (which recognition block was viewed first) and image type on the x-axis. Mean values are represented by the cross, the line within the boxplot represents the median, the boxplot itself represents the interquartile range, and the dots show individual data points.

### 4.3. Discussion

Contrary to the first experiment, in Experiment 1b, no interaction was found between image type and distinctiveness. As well, false alarm rates were similarly large for typical and distinctive faces when foils contained hair information (whereas Experiment 1a found no significant differences in FA patterns overall) The findings of Experiment 1b did not match our predictions as hair being unavailable was similarly disruptive for both distinctive and typical faces. The assumption that a salient feature will be encoded to the detriment of other features in the weighted representation model is not supported here (Devue, in prep). As both distinctive and typical faces have comparable effect sizes, it seems participants' representations for both faces relied to some degree on hair as a diagnostic feature.

This is also at odds with what the weighted representation model would predict, as hair would not be viewed as a diagnostic feature for typical faces, and therefore weighting should not be placed on it (Devue, in prep). One possibility is that hair was perceived as a diagnostic feature for typical faces. Despite the fact that hair information was kept the same for all typical faces, our findings suggest that people weighted representations on hair regardless of whether hair was diagnostic or not. As the weighted representation develops over time, it may be that hair is seen as diagnostic for both typical and distinctive faces because it is initially helpful for recognition when first encoding and learning a face. Some evidence suggests hair can be important for recognizing newly learned faces and unfamiliar faces (Bartel et al., 2018; Toseeb et al., 2012). Although others debate, that hair is not essential for our representation of faces (Murphy et al., 2014).

We also see that holistic interference may be the main culprit of the disruption to the performance from the hair to bald images. As explained previously, once hair information which was at first incorporated into the processing of these faces is gone during the test phase, recognition becomes disrupted because the faces are less recognizable without the same hair information that was incorporated into processing the face during the learning phase (Toseeb et al., 2012). Here we see that typical and distinctive faces are comparably disrupted by holistic interference, again indicating recognition for both types of faces relied on hair information.

As with Experiment 1a, we again found an interaction between order and image type, whereby participants had better recognition performance when viewing hair images if they had viewed bald images first. In contrast to Experiment 1a, there was a main effect of order, whereby recognition performance was better for both hair and bald images for hair as the first block than when participants saw both sets of images for the bald block first.

This difference between Experiment 1a and 1b is puzzling. It may be that with comparable sample sizes for counterbalancing in Experiment 1b that we see less power for a difference to occur. Alternatively, it may be that hair may draw attention away from other aspects of the face and therefore is still relied upon to a greater extent even when it is not distinctive or diagnostic in a particular learning context. It is possible that hair, because it is peripheral information and requires the processing of only coarse information for recognition to occur, is favoured initially when trying to learn a face. Indeed, previous literature has shown for unfamiliar faces peripheral information is relied upon for recognition to occur (Ellis et al., 1979; Young et al., 1985). It may be that because our Experiment was over one day, that memory of these faces is not developed enough for participants to build a representation without the reliance on this information.

Although we did not find the same interaction in Experiment 1b we do see that again there was no main effect of distinctiveness suggesting no distinctiveness advantage was found compared to previous literatures' findings (Valentine et al., 1986a). As well the lack of a distinctiveness advantage within a group context is unlike what would be expected by some of the literature.

Past research has argued group context should enhance further processing of a face that 'pops out' by bringing attention away from the other faces towards the more salient face (von Restorff, 1933, as cited in Hosie & Milne, 1996). Our findings do not support this idea of additional processing to the salient face within a group. Instead, the encoding of the distinctive face (primarily through encoding the hair) did not lead to the greater encoding of other aspects of the face. It may be that because hair can be encoded with just coarse information that the representation of the distinctive faces was much less fine-grained and much coarser than for typical faces, with the distinctive faces having a larger reliance on hair for recognition to occur.

As mentioned in the weighted representation model, representations of faces may build from a coarse-to-fine strategy. In the group context, the distinctive face may be viewed more parsimoniously, with participants realizing hair is an easy and highly efficient way to remember the distinctive face, and therefore less processing is put towards details of other aspects of the face. This may also explain why there was no distinctiveness advantage (in hit rates for hair present trials) for the distinctive faces. If recognition were dependent on encoding hair as the main feature of the face, your overall representation would be quite coarse.

In addition, the decision criterion was similarly liberal between distinctive and typical faces as evidenced by our analysis of criterion (see appendix A). It may be that because only coarse information was necessary for recognition of both typical and distinctive faces, participants were more likely to judge a face as one they had seen for both types of faces. This may also account for the lack of a distinctiveness advantage within Experiment 1.

Unlike previous literature which shows either a higher hit rate for distinctive faces (Valentine, 1986a) or a decrease in false alarms (Bartlett et al., 1984), we found no significant advantage for distinctive faces in either increased hit rates or decreased false alarms when hair was present compared to typical faces across Experiment 1. Participants' reliance on hair may have been at the overall expense of encoding of other features that would aid recognition. Further, because the hair was relied on too much for the distinctive face, typical faces may have received additional attention and processing time, leading to comparable recognition when hair information was available.

The findings of Experiment 1 also do not directly contradict what would be expected with the averaging hypothesis (Burton et al. 2005; Murphy et al., 2015). You might expect that each of the faces would be equally well remembered with or without hair as it is uninformative in the representation of the face.

However, you might also argue that as hair did not change between learning and test, it was therefore not variable and included in the representation of the face. Indeed, the hair remained stable over learning and would presumably be seen as invariant. Although hair and external features are generally assumed as stable features (unless otherwise stated), in the averaging hypothesis, they do not play an important role in the long-term representations of a face (Murphy et al., 2015). Within the weighted representation model, you would predict hair information would be encoded if it was seen as diagnostic to the face. Especially during initial encoding hair may be seen as diagnostic and may be included in the representation (at least as coarse information). Because this thesis is only looking at the initial encoding of a face, we cannot make claims about the later refinement of representations that might occur. It is possible that if you were to view these same set of faces in the same group context again that you may find the same reliance on hair for recognition. In terms of the real world, this may apply to when you only view individuals in specific contexts (e.g., at work, at the gym, etc.) where a coarse representation is enough for recognition and helpful within that given context. However, this would not help for recognition in other contexts in which you may see those individuals.

Additionally, some may argue that by manipulating distinctiveness through hair colour, we are only testing the influence of external or peripheral information on our face representations although recent research suggests that hair might be an essential aspect of recognition (Abudurham et al., 2018). Indeed, some researchers have argued hair is not as central to face representations as inner invariant features are (Burton et al., 2016; Murphy et al., 2015). Instead, it has been argued that our representations are more reliant on internal features of the face with peripheral information (which is often variable) being discarded

(Murphy et al., 2015). Other researchers have also argued that recognition performance is improved when participants place more attention and therefore encoding on internal features of the face rather than peripheral information (Burton et al., 2016; Dowsett et al., 2016). To counter the argument that our findings may only apply to peripheral aspects of the face, a follow-up study was conducted manipulating a central internal feature, namely the eyes.

## 5. Experiment 2

The aim of Experiment 2 was to examine if the learning context could affect how faces were encoded and how their representations were weighted when the distinctive/diagnostic feature was an internal feature. Participants viewed a set of faces with one face being made distinctive by having a different eye colour than the other faces. During the recognition task, participants were tested on their recognition when the diagnostic feature was available (i.e., block with eyes open) and when the diagnostic feature was unavailable (i.e., the block with eyes closed).

### 5.1. Methods

#### 5.1.1. Participants and Exclusions

Exclusion criteria were identical to those in Experiment 1. We recruited 97 participants and excluded 17, leaving 80 participants (63 female, 15 male, and 2 gender unspecified; *Mean age* = 19.93 ± 3.58, range 18-40).

#### 5.1.2. Stimuli and Procedure

The same learned and unlearned faces as Experiment 1a and 1b were used here. This time, the images did not include hair and were manipulated in terms of eye colour (see Figure 10). The images were split into three conditions: a blue-eyed, a brown-eyed, and an eye closed condition.  The eye closed condition was meant to be equivalent to the bald image condition, in which the diagnostic feature (this time eye colour) was not available for recognition.

To account for the fact that eye colour may be distinctive within a group (i.e., blue eyes may stand out more than brown), each individual face was distinctive in both a blue and a brown-eyed version. In total, 16 versions of the task were created (eight blue and eight brown-eyed versions; eight with eye-closed trials first and eight with eye-open trials first). Participants were randomly assigned to one of the 16 versions (five participants per version). The same procedure as in the previous two experiments was used here.

*Figure 10.* Example stimuli for the learning phase.

### 5.1.4. Design and Analyses plans

After initial data collection but before data completion and before any analyses were conducted, the plans were pre-registered on Open Science Framework at [https://osf.io/qcd6u].

The same analyses as in Experiment 1a were conducted in Experiment 2 to examine the impact of context on the saliency of internal features.

## 5.2. Results

### 5.2.1. Pre-planned Analyses

*Hit rates*. For hit rate, a significant main effect of image type was found, $F(1, 79) = 18.47$, $p < .001$, $\eta^2 = .036$, whereby participants had better recognition of eye open images than eye closed images. As well, a significant main effect of distinctiveness was found, $F(1,79) = 4.30$, $p = .041$, $\eta^2 = .015$, showing participants had higher hit rates for distinctive than typical faces. However, no significant interaction was found between distinctiveness and image type, $F(1, 79) = 2.17$, $p = .145$, $\eta^2 = .005$. Results are displayed in Figure 11.

Due to our specific predictions, follow-up paired samples t-tests were still conducted. Follow-up paired samples t-tests showed participants did not significantly differ in recognition rates between distinctive ($M = .63 \pm .26$) and typical faces ($M = .60 \pm .59$) for eye closed trials, $t(79) = .74$, $p = .463$, $d = .083$. However, for eye open trials, participants had significantly better recognition rates for distinctive ($M = .74 \pm .25$) over typical faces ($M = .66 \pm .19$), $t(79) = 2.48$, $p = .015$, $d = .277$, suggesting a recognition advantage for distinctive faces. Moreover, performance was higher for eye open trials than eye closed trials for distinctive faces, $t(79) = 3.26$, $p = .002$, $d = .365$. Similarly, for typical faces, performance was better for eye open than eye closed trials, $t(79) = 2.83$, $p = .006$, $d = .317$. Unlike our predictions that distinctive face performance would be more disrupted by the diagnostic feature being unavailable, both the typical and distinctive face performance was significantly better for eyes open than eye closed trials and the effect sizes were in the same ballpark.

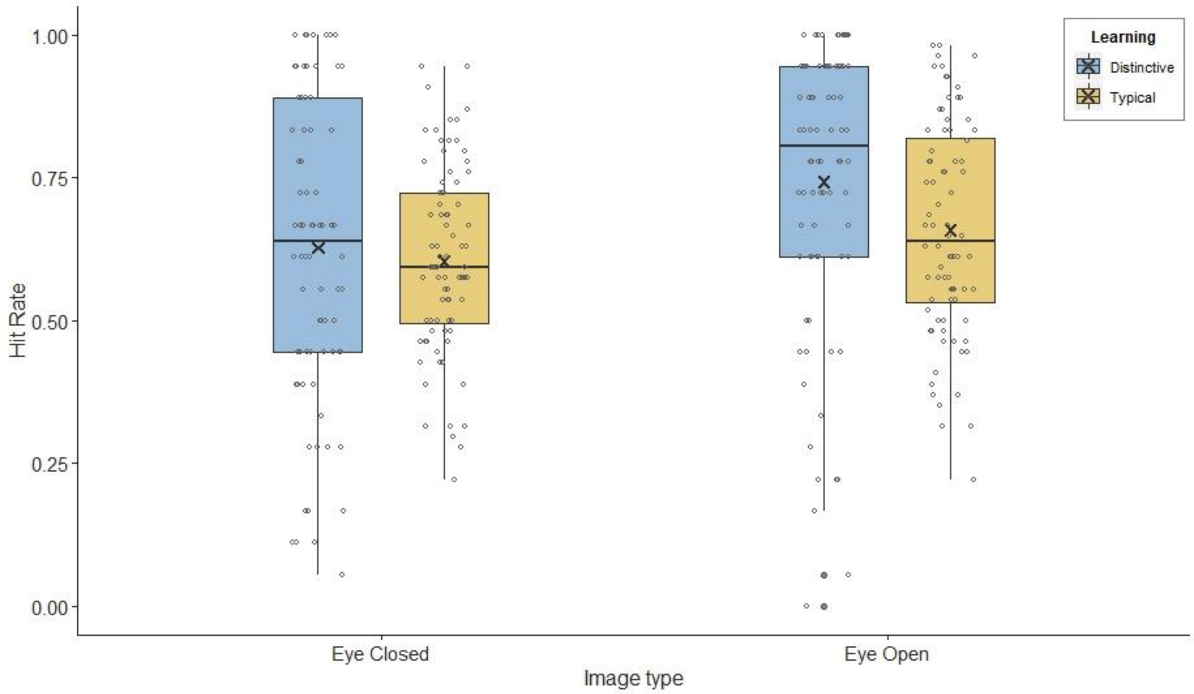


*Figure 11*. Hit Rate for Experiment 2 as a function of eye-closed and eye-open distinctive and typical conditions. Mean values are represented by the cross, the line within the boxplot represents the median, the boxplot itself represents the interquartile range, and the dots show individual data points.

*False Alarms*. For false alarms, no main effect of distinctiveness was found, $F(1, 79) = 0.32$, $p = .572$, $\eta^2 = .001$. Additionally, no main effect of image type was found, $F(1, 79) = 0.06$, $p = .234$, $\eta^2 = .003$. No significant interaction was found between distinctiveness and image type, $F(1, 79) = 9.65$, $p = .860$, $\eta^2 = .000$. Results are displayed in Figure 12.

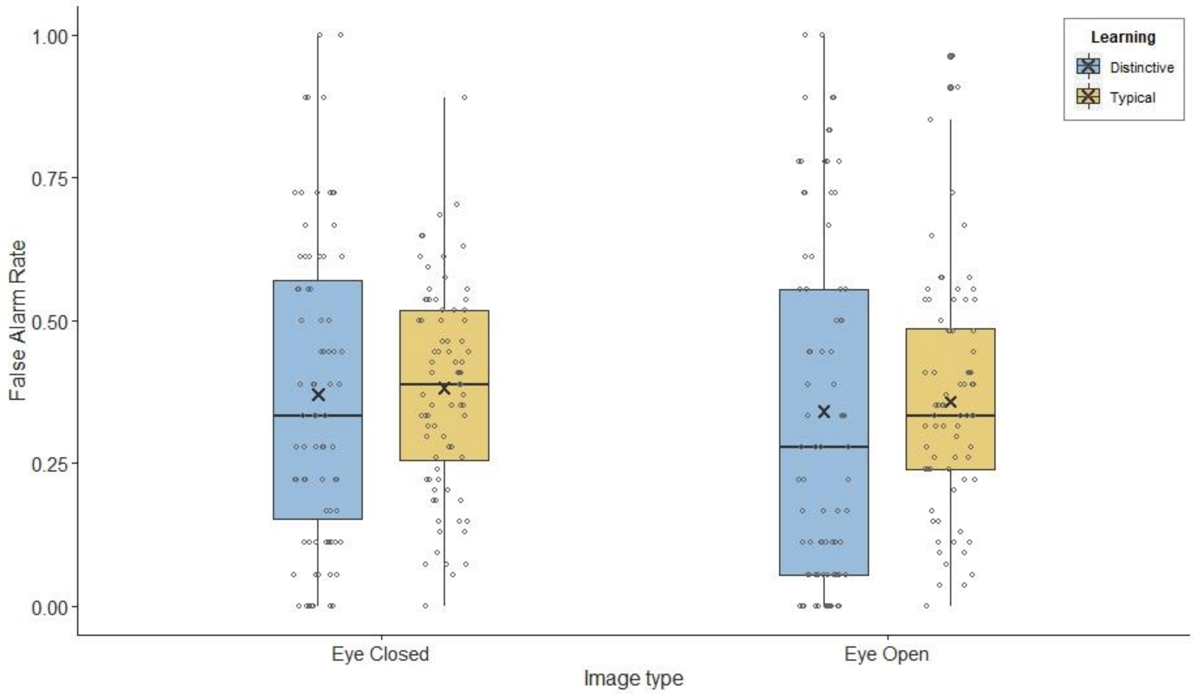


*Figure 12.* False Alarm Rate for Experiment 2 as a function of eye-closed and eye-open distinctive and typical conditions. Mean values are represented by the cross, the line within the boxplot represents the median, the boxplot itself represents the interquartile range, and the dots show individual data points.

*Reaction time*. Median RTs indicated there were no speed-accuracy trade-offs (range = 856-981 ms). RTs are displayed in Table 5.

.

Table 5. *Reaction times corresponding to Hit rates and False alarms*

| **Median(*SD*)** | **Hit Rates** | **False Alarms** |
|---|---|---|
| **Distinctive Absent** | 879(*275.65*) | 901.5(*314.31*) |
| **Distinctive Present** | 856.5(*285.07*) | 981.5(*429.10*) |
| **Typical Absent** | 865.5(*205.68*) | 922.75(*260.25*) |
| **Typical Present** | 909(*227.68*) | 967(*272.60*) |

Note. *N = 80*

### 5.2.2. Exploratory Analyses

For consistency, the same exploratory analyses on block order was conducted here. A 2 (Distinctiveness: distinctive vs typical faces) x 2 (Image type: Eye-closed vs Eye-open) x 2 (Order: Eye-closed block first vs Eye-open block first) mixed factorial ANOVA was conducted on hit rate.

For Experiment 2, there was a significant main effect of image type, $F(1, 78) = 18.24$, $p < .001$, $\eta^2 = .036$ and a main effect of distinctiveness, $F(1, 78) = 4.25$, $p = .043$, $\eta^2 = .015$. There was no main effect of order, $F(1, 78) = .011$, $p = .918$, $\eta^2 = .000$. No significant interaction was found between image type and distinctiveness, $F(1, 78) = 2.17$, $p = .145$, $\eta^2 = .005$. Additionally, no significant interaction was found between image type and order, $F(1, 78) = .01$, $p = .922$, $\eta^2 = .000$. No significant interaction was found between distinctiveness and order, $F(1, 78) = .01$, $p = .905$, $\eta^2 = .000$, or between all three variables, $F(1, 78) = .95$, $p = .332$, $\eta^2 = .002$. Results are displayed in Figure 13.

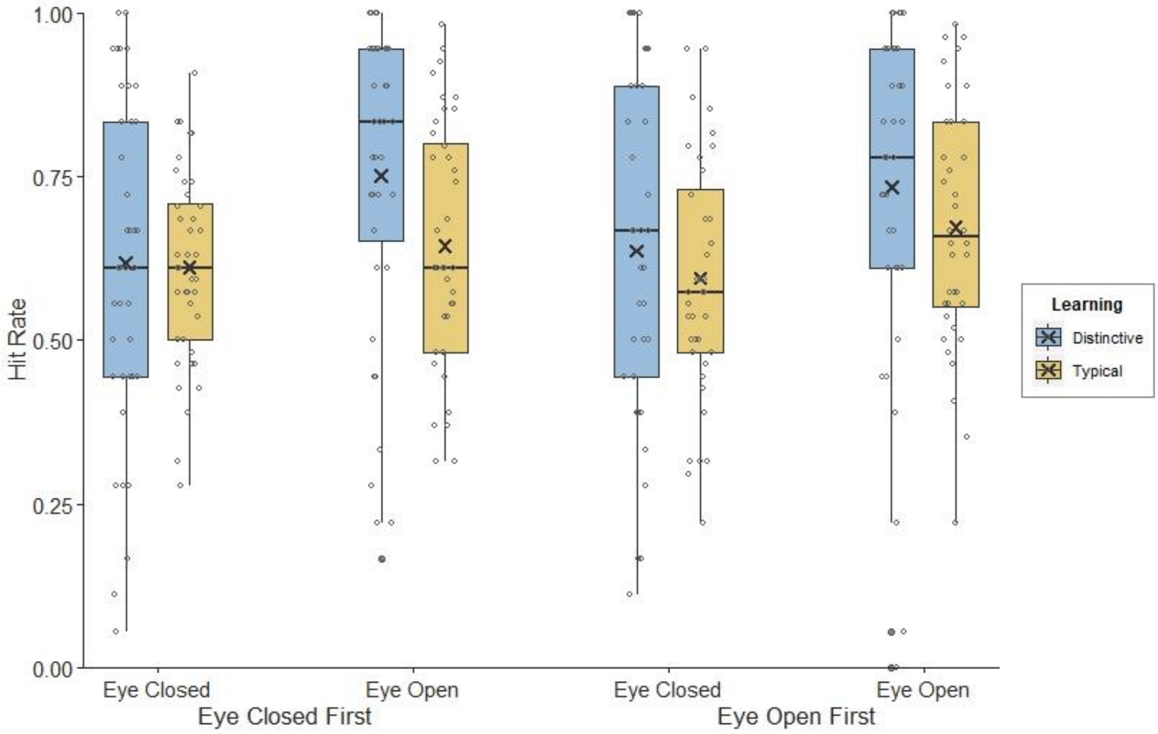


*Figure 13.* Order Effects of Experiment 2. Hit rates are measured on the y-axis, with order (which recognition block was viewed first) and image type on the x-axis. Mean values are represented by the cross, the line within the boxplot represents the median, the boxplot itself represents the interquartile range, and the dots show individual data points.

## 5.3. Discussion

For Experiment 2, we see a change in the pattern of performance compared to Experiment 1a and 1b. This time, participants had better performance for distinctive than typical faces when the diagnostic feature was available for recognition. Similar to Experiment 1b, distinctive and typical faces were similarly hurt by the absence of eye information during the recognition task. We observe different patterns of performance in all three experiments with Experiment 1b and Experiment 2 showing comparable levels of disruption when hair is not available for both distinctive and typical faces (contrary to Experiment 1a's findings) and better performance for distinctive faces when the feature of interest was available (contrary to

Experiment 1 findings). For Experiment 1a and 1b, this is likely due to the counterbalancing issues of Experiment 1a. This order effect seems especially pertinent to how encoding might change depending on what participants view first. Due to the first experiment's order effects, we decided to complete additional exploratory analyses of block order to determine if the contextual effects found in all three experiments are dependent on the order that images are presented.

In our second experiment, eyes were used to examine the influence of context on the way facial features are encoded. Our findings largely mirrored the pattern of performance found in the first experiment, but there were some noticeable differences. To start, there were higher hit rates and greater numbers of correct recognition of the distinctive faces compared to typical faces, showing the distinctiveness advantage found in the broader literature. Participants had better recognition of distinctive than typical faces when the eyes were open and their colour visible during the recognition task. This was similar to past research showing an increased hit rate for distinctive faces (Valentine, 1986a). Although Experiment 1 did not find our predicted distinctiveness advantage, Experiment 2 does align with our initial prediction of better recognition performance for distinctive over typical faces.

Additionally, participants did not differ significantly in performance for distinctive and typical faces when the eyes were closed, removing the distinctive feature. Unlike in Experiment 1, where the distinctive feature impaired the encoding of other features, in Experiment 2, the distinctive feature enhanced or facilitated the encoding of additional/surrounding facial information. Based on the weighted representation model, one would expect individuals to encode salient features to the detriment of other features (Devue, in prep). Instead, the current findings show an enhancement of recognition for distinctive faces when the salient feature is both available and unavailable indicating the representation is weighted on both the eyes and other internal facial features. This enhancement mirrors what would be expected based on the Restorff effect in that the salient face had greater attention and therefore, further processing of features occurred (von Restorff, 1933, as cited in Hosie & Milne, 1996). This greater refinement and recognition advantage are especially interesting given the fact that the difference between eye colours in the group learning was only subtly different and much less noticeable than the difference in hair colour in Experiment 1.

As with Experiment 1, the performance was better when the feature of interest was available (eye-open condition) than when it was unavailable (eye-closed) for both distinctive and typical faces. Counter to our predictions; the effect size was similar between distinctive and typical faces' hit rates. For both distinctive and typical faces, recognition performance was improved when the eyes were available compared to when they were not.

## 6. General Discussion

In these two experiments, we have investigated the effect of context on the saliency of facial features. Over the two studies, participants were given learning and then recognition tasks. In the learning phase, we presented participants with images of faces with one of these faces being distinctive to the others in hair or eye colour. During the test phase, participants saw images of these faces with the feature of interest (in the hair/eye-open condition) and without the feature of interest (in the bald/eyes-closed condition). We predicted that recognition rates (i.e., hit rates) would be higher for the distinctive over the typical faces when the feature of interest was visible, but that recognition would be more disrupted for distinctive faces than typical faces when the feature of interest was not available for recognition (through comparison of effect sizes). In Experiment 1, hair colour was manipulated to make one face distinctive compared to a group of three other faces. In Experiment 2, eye colour was manipulated to make one face distinctive compared to three other faces. In Experiment 1, we found hair was encoded to the detriment of other features for the distinctive faces but that distinctive face recognition was not better than typical face recognition when hair was available, partially supporting our hypothesis. In Experiment 2, we found eyes were encoded and facilitated further encoding of other aspects of the face for distinctive faces contradicting our prediction that the distinctive feature would be encoded to the detriment of other features. The two studies provide evidence that context does affect what we find distinctive within a face and how that affects the encoding and later representations of faces. Below we discuss the implications of our findings in terms of past theories, the weighted representation model, and possible order effects within our experiment.

Both experiments provide evidence that the learning context did influence what features were encoded during learning. This is in line with the weighted representation model, which suggests that distinctiveness affects the weighting of your face representations (Devue, in prep). For Experiment 1, encoding involved more coarse information (i.e., hair) and led to a very coarse representation of the face that was not reliant on other less salient aspects of the face. By contrast, encoding in Experiment 2 involved both the eyes and other aspects of the face. In terms of the weighted representation model, the fact that other facial features were also encoded and recognized in the distinctive faces contradicts the model's assumptions that distinctive features will be encoded to the detriment of other features.

This is important for our current understanding of face representations. While current theories hypothesise within-person variability plays a crucial role in learning a face (Murphy et al., 2015; Burton et al., 2008) it is apparent that distinctiveness (which is related to between-person variability in the weighted representation model), as manipulated by context, may play a crucial role in the acquisition of diagnostic information from a face. Indeed, learning context induces a between-person comparison, and in our experiment specifically, it is used as an artificial way of making one face distinctive within a group. This between-person comparison seems to be especially important as it affects how we memorize and recognize a face. As mentioned previously, it may be that over time, when the learning context changes, variability plays more of a role in how we adapt our representations. However, if a face (such as the distinctive face in this experiment) continues to seem salient based on the same features, its representation may remain overly reliant on that one feature (as is the case for Experiment 1). This would be in line with past research showing decreases in recognition for iconic celebrities when their iconic features are not available (e.g., Cindy Crawford's mole or the Pope's hat in Carbon's 2008 study).

Why there might be this difference in representations between Experiment 1 and Experiment 2 is up for debate. It may be that, as eyes are at the centre of the face, the distinctive colour enhances the processing of other inner features around them. Because eyes are often rated as the most informative feature of faces generally (Fraser et al., 1990), it may be that eyes are focussed on as a diagnostic feature in general and that the experimental manipulation further helped to direct processing towards the eyes and surrounding areas.

Interestingly, hair usually draws attention when initially encoding a face (Bartel et al., 2018), and therefore, when hair becomes especially diagnostic (as in Experiment 1), the processing is moved away from the central aspects of the face and towards the hair instead to the detriment of the other features. This is in line with what the weighted representation model would assume; as we try to be parsimonious with our encoding of a face, we will initially encode what is seen as most diagnostic about a face (in this case hair seems especially important) with the assumption that encoding that distinctive feature will be to the detriment of other features (Devue, in prep). However, the results of Experiment 2 suggest that individual features may differ in how diagnostic they are and how the representation may be weighted. It may be that dependent on the feature and where it lies within the face may lead to different encoding and coarse-to-fine strategies.

**6.1. Order Effects**

The findings from the Experiment 1a and 1b indicate that the effects of image type were dependent on what order participants saw the stimuli. Crucially, the order could not account for the differences in performance for distinctive face recognition in any of the three experiments.

For Experiment 1a, when participants saw bald images first, they had better recognition performance for hair images than bald images. In contrast, when hair images were presented first, there was no difference between hair and bald image recognition performance. Here it seems order may play a role in participants' recognition abilities when it comes to image type; however, due to our unbalanced counterbalancing within the experiment, we cannot be sure.

For Experiment 1b, participants had higher hit rates when viewing hair images than bald images when bald images were viewed first compared to when hair images were viewed first (similarly to Experiment 1a).

In Experiment 2, the order of presentation did not affect recognition performance. Unlike in Experiment 1a and 1b, participants did not differ in recognition abilities of image types when the experiment manipulated distinctiveness in eye colour.

Why might there be this order effect between the two experiments? One explanation was first introduced by Hintzman (2011). Hintzman (2011) argued that current memory researchers tend to hold a process-pure view of study and test phases, in that, most paradigms used to study memory include a study and test phase in which one you presumably encode (study phase) and the other you use retrieval processes (test phase). Instead, these two processes may be more interdependent than researchers have currently assumed. Hinztman

(2011) also noted that because information is changing around us constantly, rather than new experiences overwriting our memory that we instead use the current situation as a reminder of new information on a given topic (i.e., an update on our current representation). It may be that the effect of order in our two experiments is because we are continually encoding and reupdating our representations every time we view a face.

For Experiment 1, when participants view the bald image trials before the hair images, participants may realize the faces that they previously had encoded no longer have the key features that they relied on previously when studying, they now update their representations by encoding more of the inner features of the faces (as this becomes more diagnostic when hair is no longer available). This means that when participants get to the hair present trials, they have encoded more of the inner features and now can recognize the faces using both inner feature and hair information. In comparison, when participants saw the hair images first, participants were able to see that hair was not as diagnostic as in the learning phase due to foils with the same hair colour being included within the block and start to encode inner features more readily. In return, when bald images are viewed recognition is less reliant on hair for both faces leading to better recognition.

In both Experiment 1a and 1b, recognition performance increased when participants first saw images without the hair and then later with the hair available. This, in part, maybe due to participants updating their representations of the faces from the first recognition test block to the next.

For Experiment 2, no order effects were found with similar advantages with the eye-open condition than the eye-closed condition, no matter if eye-open or eye-closed images were shown first. It is possible for Experiment 2 that encoding the rest of the face was helpful enough for recognition in all conditions and orders of stimuli.

Unlike for Experiment 1, because eye information may have already led to focussing the face representation on the central aspects of the face, the representation during the two encoding phases was not disrupted by the order in which participants viewed the images.

Based on the weighted representation model, we assume that stability (or lack thereof) will lead to encoding features that are more diagnostic of recognition (Devue, in prep). As Hintzman (2011) argued, encoding and recognition should be updated between the learning and recognition phases of an experiment. Here, the weighted representation model can account for this change in representation based on the order. When participants in the recognition phase are initially shown images without a feature that was previously seen as diagnostic (i.e., seeing the bald images first), stability is lost, and they must weigh their representation on other aspects of the faces, leading to refinement of other aspects. The effect of order shows how our representation may change weighting when a feature becomes unavailable for recognition (changing the stability of a given face), while this occurred by fully obscuring or removing the feature entirely in this experiment; in the real world, this may occur with small changes to a person's appearance (for example a haircut or hair colour change). It seems we update and refine our representations of faces based on both our previous understanding of how the face was diagnostic and on later encounters that determine what features remain stable and diagnostic of identity.

**6.2. Limitations**

One limitation of this study is the artificial nature of the test stimuli. We used artificially generated faces in order to control perceived distinctiveness and for practical reasons. Although, artificially generated faces have many benefits in terms of controlling our manipulation of distinctiveness, it does not fully reflect the way we learn and recognise real faces. For example, some research has shown different neural activity between participants viewing real or artificial face stimuli (Wheatley et al., 2011). A future study using ecological stimuli will provide further evidence of how context can affect perceived distinctiveness, and whether this can influence our representations. Using ecological stimuli would require some changes to the methods. For example, faces would have to be rated on structural distinctiveness and equated before trying to manipulate hair or eye colour.

**6.3. Future Directions**

The current experiments do show that the initial learning of a face is influenced by context and other faces that are being learned. Future research could also investigate the longer-term contribution of contextual information on our representations. There is evidence that the advantage for distinctive faces remains over 4-weeks but not over 6-weeks (Metzger, 2006). As the weighted representation model is built on the idea that our representations are dynamic, you might expect different weighting over time as participants become more accustomed to the stimuli they are learning. To test this idea, participants could repeat learning and be re-tested on their recognition of the same faces from Experiment 1 and 2, with the same format but repeated over extended periods (e.g., weeks).

It may be that the initial encoding is reliant on the distinctive feature but that the weighting changes over time as stability may vary as you reencounter a face multiple times. How the weighting changes may depend on the relative stability of faces, something that was not investigated in these studies. If the colour of the distinctive feature in a face changed, you would expect the initial reliance on that feature to decrease for the hair condition as that feature becomes less stable leading to other features being seen as more diagnostic. For the eye condition, you might expect more of a reliance on other features of the face as well, since the distinctive eyes already facilitate recognition of other features, this might lead to even more refinement of other features because the eyes are no longer stable and diagnostic of an identity. Indeed, for both types of distinctive faces, if the distinctive feature does not maintain stability, you might expect other areas of the face to become encoded and for greater refinement of those areas to occur.

In addition, due to the difference in findings between Experiment 1 and 2, it is evident that future studies need to investigate other features of the face regarding context. Eyes are generally seen as quite diagnostic features (Fraser et al., 1990), and hair is often seen as aiding recognition when faces are not yet familiarized (Ellis et al., 1979; Young et al., 1985). A follow-up experiment could investigate another feature such as the mouth which is still central to the face but often seen as less diagnostic to investigate contextual distinctiveness. Another external/peripheral feature like the ears may be used as well (while not central to the face they require more refined processing than hair with its coarse information).

**6.4. General conclusion**

Overall, this study examined the influence of context on our initial encoding and representation of a face. In Experiment 1 recognition was best when hair was available during the test phase and was disrupted when hair was not available, but overall we did not see better recognition for the distinctive faces over the typical ones. By contrast, in Experiment 2, recognition was better for the distinctive faces than the typical (giving a distinctiveness advantage typical in the literature). As well, recognition being better for eye open versus eye closed conditions for both distinctive and typical faces similar to Experiment 1, allowing us to conclude that encoding was affected by context. Our findings support the idea that diagnostic features of a face can affect how we encode and later recognize a face, and suggest that context can determine how distinctive we perceive a face to be (especially when first learning a face).

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *The Quarterly Journal of Experimental Psychology*, *68*(10), 2041-2050. https://doi.org/10.1080/17470218.2014.1003949

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision, 16*(3):40, 1-18. https://doi.org/10.1167/16.3.40.

Abudarham, N., Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. *Cognition, 182*, 73-83. https://doi.org/10.1016/j.cognition.2018.09.002

Bartel, S. J., Toews, K., Gronhovd, L., & Prime, S. L. (2018). "Do I Know You?" Altering hairstyle affects facial recognition. *Visual Cognition, 26*(3), 149-155. https://doi.org/10.1080/13506285.2017.1394412

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory and Cognition, 12*, 219-228. https://doi.org/10.3758/BF03197669

Baker, K. A., Laurence, S., & Mondloch, C. J., (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity, *Cognition*, *161*, 19-30. https://doi.org/10.1016/j.cognition.2016.12.012

Bassett, D. S., Bullmore, E. T., Meyer-Lindenberg, A., Apud, J. A., Weinberger, D. R., & Coppola, R. (2009). Cognitive fitness of cost-efficient brain functional networks. *Proceedings of the National Academy of Sciences*, *106*(28), 11747–11752. http://doi.org/10.1073/pnas.0903641106

Bruce, V., Henderson, Z., Newman, C., & Burton, A.M. (2001). Matching Identities of Familiar and Unfamiliar Faces Caught on CCTV Images. *Journal of Experimental Psychology: Applied, 7*(3), 207-218. https://doi.org/10.1037/1076-898X.7.3.207

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

Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from Variation: Representations of Faces Derived from Multiple Instances, *Cognitive Science, 40*(1), 202-223. https://doi.org/10.1111/cogs.12231

Burton, A. M., Wilson, S., Cowan M., & Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence From Security Surveillance. *Psychological Science, 10*(3), 243-248. https://doi.org/10.1111/1467-9280.00144

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*(3), 256-284. https://doi.org/10.1016/j.cogpsych.2005.06.003

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

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception, 31*, 985-994. https://doi.org/10.1068/p3335

Davidenko, N., & Ramscar, M. (2005). Distinctiveness Effects in Face Memory Vanish with Well-Controlled Distractors. *Proceedings of the Cognitive Science Society, 27*(27). Retrieved from http://www.escholarship.org/uc/item/0df031kt

Devue, C., Wride, A., & Grimshaw, G. M. (2018). New insights on real-world human face recognition. Journal of Experimental Psychology: General, 148(6), 994–1007. https://doi.org/10.1037/xge0000493

Devue, C. (2020). A Dynamic Weighting Account of Familiarisation with Faces in the Real-world. Manuscript in preparation.

Dowsett, A. J., Sandford, A., & Burton, A. M. (2016). Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. *The Quarterly Journal of Experimental Psychology, 69*(1), 1-10. https://doi.org/10.1080/17470218.2015.1017513

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*(4), 576-585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1980). The Deterioration of Verbal Descriptions of Faces Over Different Delay Intervals. *Journal of Police Science and Administration, 8*(1), 101-106. Retrieved from https://www.researchgate.net/publication/252068683_The_deterioration_of_verbal_descriptions_of_faces_over_different_delay_intervals

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

Fisher, G. H., & Cox, R. L. (1975). Recognizing human faces. *Applied Ergonomics, 6*(2), 104-109. https://doi.org/10.1016/0003-6870(75)90303-8

Fraser, I. H., Craig, G. L., & Parker, D. M. (1990). Reaction time measures of feature saliency in schematic faces. *Perception, 19*, 661-673. https://doi.org/10.1068/p190661

Friedman, M. P., Reid, S. R., & Carterette, E. C. (1971). Feature saliency and recognition memory for schematic faces. *Perception & Psychophysics, 10*(1), 47-50. https://doi.org/10.3758/BF03205767

Goldstein, A. G., & Mackenberg, E. J. (1966). Recognition of human faces from isolated facial features: a developmental study. *Psychon. Sci, 6*, 149-150. https://doi.org/10.3758/BF03328001

Gibson, E. J. (1969). *Principles of perceptual learning and development.* Appelton-Century-Crofts.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*(9), 300-337. https://doi.org/10.1016/s1364-6613(00)01519-9

Hintzman, D. L. (2011). Research Strategy in the Study of Memory: Fads, Fallacies, and the Search for the "Coordinates of Truth". *Perspectives on Psychological Science, 6*(3), 253-271. https://doi.org/10.1177/1745691611406924

Hosie, J. A., & Milne, A. B. (1996). The Effect of Experimental Design on Memory for Typical and Distinctive Faces. *Memory, 4*(2), 175-197. https://doi.org/10.1080/096582196389031

Jenkins, R., & Burton, A. M. (2008). 100% Accuracy in Automatic Face Recognition. *Science, 319,* 435. https://doi.org/10.1126/science.1149656

Jenkins, R., White, D., Montfort, X. V., & 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

Jones, S. P., Dwyer, D. M., & Lewis, M. B. (2015) Learning Faces: Similar Comparator Faces Do Not Improve Performance. *PLoS ONE, 10*(1): e0116707. https://doi.org/10.137/journal.pone.0116707.

Klatzky, R. L., & Forrest, F. H. (1984). Recognizing familiar and unfamiliar faces. *Memory & Cognition, 12*(1), 60-70. https://doi.org/10.3758/BF03196998

Langdell, T. (1978). Recognition of faces: An approach to study of autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 19,* 255-268. https://doi.org/10.1111/j.1469-7610.1978.tb00468.x

Laughery, K. R., Alexander, J. F., & Lane, A. B. (1971). Recognition of human faces: Effects of target exposure time, target position, pose position, and type of photograph. *Journal of Applied Psychology, 55*(5), 477-483. https://doi.org/10.1037/h0031646

Light, L., Kayra-Stuart, F. & Hollander, S. (1979). Recogniton memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 212-228. https://doi.org/10.1037/0278-7393.5.3.212

Lobmaier, J. S., & Mast, F. W. (2007). Perception of Novel Faces: The Parts Have it. *Perception, 36*(11), 1660-1673. https://doi.org/10.1068/p5642

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. (2018a). Improving Identity Matching of Newly Encountered Faces: Effects of Multi-image Training. *Journal of Applied Research in Memory and Cognition, 7*, 280-290. https://doi.org/10.1016/j.jarmac.2017.10.005

Matthews, C. M., & Mondloch, C. J. (2018b). Finding an unfamiliar face in a line-up: Viewing multiple images of the target is beneficial on target-present trials but costly on target-absent trials. *British Journal of Psychology, 109*, 758-776. https://doi.org/10.1111/bjop.12301

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *TRENDS in Cognitive Sciences*, *6*(6), 255-260. https://doi.org/10.1016/S1364-6613(02)01903-4

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34* (4), 865-876. https://doi.org/10.3758/BF03193433

Menon, N., Kemp, R. I., & White, D. (2018). More than a sum of parts : robust face recognition by integrating variation. *Royal Society Open Science, 5*, 172381. https://doi.org/10.1098/rsos.172381

Menon, N., White, D., & Kemp, R. I. (2015). Variation in Photos of the Same Face Drives Improvements in Identity Verification. *Perception, 44*(11), 1332–1341. http://doi.org/10.1177/0301006615599902

Metzger, M., & Bridges, K. R. (2004). Spectacles, Distinctiveness, and Face Recognition: A Web-Based Experiment. *CyberPsychology & Behavior, 7*(1), 113-117. https://doi.org/10.1089/109493104322820183

Metzger, M. M. (2006). Face Distinctiveness and Delayed Testing: Differential Effects on Performance and Confidence. *The Journal of General Psychology*, *133*(2), 209-216. https://doi.org/10.3200/GENP.133.2.209-216

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

Mundy, M. E., Honey, R. C., Dwyer, D. M. (2007). Simultaneous Presentation of Similar Stimuli Produces Perceptual Learning in Human Picture Processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(2), 124-138. https://doi.org/10.1037/0097-7403.33.2.124

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity, *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 577-581. http://dx.doi.org/10.1037/xhp0000049

Oliva, A. (2005). Gist of the scene. In *Neurobiology of Attention* (pp. 251–256). http://doi.org/10.1016/B978-012375731-9/50045-8

Ramon, M. (2015). Perception of global facial geometry is modulated through experience. PeerJ 3:e850. https://doi.org/10.7717/peerj.850

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

Robins, E., Susilo, T., Ritchie, K., Devue, C. (2018). Within-person variability promotes learning of internal facial features and facilitates perceptual discrimination and memory. OSF Preprints. https://doi.org/10.31219/osf.io/5scnm

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica, 128*(2), 274-289. https://doi.org/10.1016/j.actpsy.2008.02.003

Ruiz-Soler, M., & Beltran, F. S. (2012). The Relative saliences of facial features when differentiating faces based on an interference paradigm. *Journal of Nonverbal behaviour, 36,* 191-203. https://doi.org/10.1007/s10919-012-0131-z

Seamon, J. G., Stolz, J. A., Bass, D. H., & Chatinover, A. I. (1978). Recognition of facial features in immediate memory. *Bulletin of the Psychonomic Society, 12*(3), 231-234. https://doi.org/10.3758/BF03329680

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

Singular Inversions (2019). *FaceGen 3.1 Modeller* [Computer Software]. https://facegen.com/index.htm

Shepherd, J., Davies, G., & Ellis, H. (1981). Studies of cue saliency. In G. Davies, H. Ellis, and J. Shepherd (Eds.), *Perceiving and remembering faces*, (pp. 103–131). Academic Press.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, *51*(6), 599–606. http://doi.org/10.3758/BF03211656

Toseeb, U., Keeble, D. R. T., & Bryant, E. J. (2012). The Significance of Hair for Face Recognition. *PLoS ONE 7*(3), e341144, https://doi.org/10.1371/journal.pone.0034144

Valentine, T. (1991). A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition. *The Quarterly Journal of Experimental Psychology, 43A*(2), 161-204. https://doi.org/10.1080/14640749108400966

Valentine, T., & Bruce, V. (1986a). The effects of distinctiveness in recognising and classifying faces. *Perception, 15*(5), 525-535. https://doi.org/10.1068/p150525

Valentine, T., & Bruce, V. (1986b). Recognizing Familiar Faces: The Role of Distinctiveness and Familiarity. *Canadian Journal of Psychology, 40*(3), 300-305. http://dx.doi.org/10.1037/h0080101

Valentine, T., & Ferrara, A. (1991). Typicality in categorization, recognition and identification: Evidence from face recognition. *British Journal of Psychology, 82*, 87-102. https://doi.org/10.1111/j.2044-8295.1991.tb02384.x

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology, 69*(10), 1996-2019. https://doi.org/10.1080/17470218.2014.990392

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20*(3), 291-302. https://doi.org/10.3758/BF03199666

Weyrauch, B., Huang, J., Heisele, B., & Blanz, V. (2004). Component-based Face Recognition with 3D Morphable Models, First IEEE Workshop on Face Processing in Video, Washington, D. C., 2004. Retrieved from http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html

Wickham, L. H. V., Morris, P. E., & Fritz, C. O. (2000). Facial distinctiveness: Its measurement, distribution, and influence on immediate and delayed recognition. *British Journal of Psychology, 91*, 99-123. https://doi.org/10.1348/000712600161709

Wheatley, T., Weinberg, A., Looser, C., Moran, T., & Hajcak, G. (2011). Mind Perception: Real but Not Artificial Faces Sustain Neural Activity beyond the N170/VPP. *PLoS ONE, 6*(3): e17960. https://doi.org/10.1371/journal.pone.0017960

Young, A. W., & Burton, A. M. (2017). Are We Face Experts? *Trends in Cognitive Sciences*, *22*(2), 100-110. http://doi.org/10.1016/j.tics.2017.11.007

Young, A. W., & Ellis, H. D. (1989). Semantic Processing. In A. W. Young & H. D. Ellis (Eds.), *Handbook of Research on Face Processing.* Elsevier: Amsterdam.

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception*, *14*(6), 737–746. http://doi.org/10.1068/p140737

# Appendices

## Appendix A – Sensitivity and Criterion Analysis for Experiment 1 and 2

### Experiment 1a

*Sensitivity*. For d', there was a significant main effect of image type, $F(1, 79) = 14.61$, $p < .001$, $\eta^2 = .019$, with d' being larger for hair present images than hair absent images. There was no main effect of distinctiveness, $F(1, 79) = .06$, $p = .807$, $\eta^2 = .000$. A significant interaction between distinctiveness and image type was found, $F(1, 79) = 4.75$, $p = .032$, $\eta^2 = .005$. Post-hoc Tukey HSD tests revealed, that participants had increased sensitivity for distinctive hair present ($M = 1.14 \pm 1.68$) than hair absent trials ($M = .60 \pm 1.43$), $t(79) = 4.30$, $p < .001$. For typical faces, there was no difference in sensitivity between hair present ($M = .99 \pm .95$) and hair absent trials ($M = .82 \pm .87$), $t(79) = 1.38$, $p = .516$. When hair was present, participants were more sensitive for distinctive than for typical faces, $t(79) = 3.56$, $p = .003$, and when hair was unavailable participants were similarly sensitive for distinctive and typical faces, $t(79) = 1.68$, $p = .339$. D' is displayed in Figure 14.

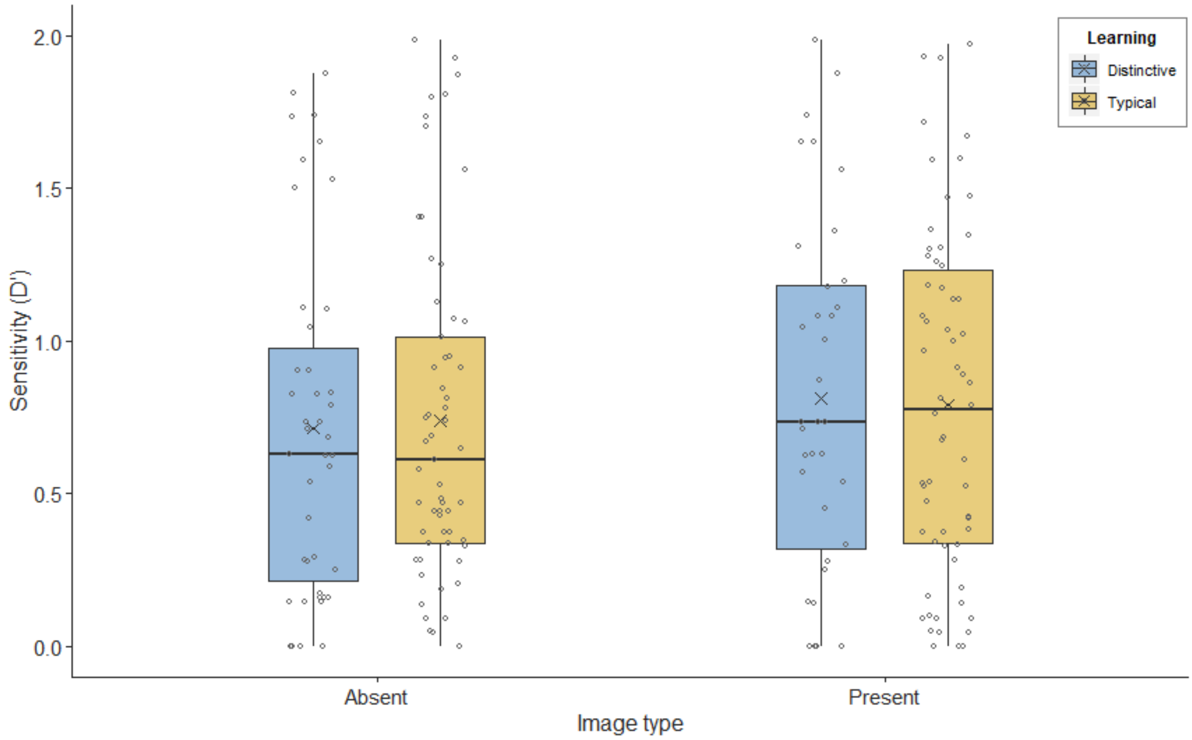


*Figure 14*. D' for Experiment 1a. Mean values are represented by the cross, the line within the boxplot represents the median, the boxplot itself represents the interquartile range, and the dots show individual data points.

*Criterion.* For criterion C, a significant main effect of image type was found, $F(1, 79) = 14.62$, $p < .001$, $\eta^2 = .028$, wherein, participants were more liberal when viewing hair present trials compared to hair absent trials. but no main effect of distinctiveness was found, $F(1, 79) = 0.06$, $p = .815$, $\eta^2 = .000$. A significant interaction between distinctiveness and image type was also found, $F(1, 79) = 7.42$, $p = .008$, $\eta^2 = .013$. Post-hoc Tukey HSD tests revealed participants differed in their bias between distinctive hair present ($M = .02 \pm .80$) and distinctive hair absent ($M = .37 \pm .69$) trials, $t(79) = 4.65$, $p < .001$ with hair absent trials being more conservative. However, participants had similar conservative biases between typical hair present ($M = .15 \pm .44$) and typical hair absent trials ($M = .21 \pm .39$), $t(79) = 0.88$, $p = .815$. When hair was present, there were no significant criterion differences between distinctive and typical faces, $t(79) = 1.39$, $p = .509$. Similarly, when hair was absent, there were no significant differences in criterion between typical and distinctive faces, $t(79) = 1.77$, $p = .292$.

**Experiment 1b**

*Sensitivity*. For d', there was a no main effect of image type, $F(1, 79) = .17$, $p = .684$, $\eta^2 = .002$ or distinctiveness, $F(1, 79) = 2.67$, $p = .106$, $\eta^2 = .033$. Additionally, no significant interaction between distinctiveness and image type was found, $F(1, 79) = .02$, $p = .888$, $\eta^2 = .000$

*Criterion*. For criterion C, a significant main effect of image type was found, $F(1, 79) = 9.85$, $p < .002$, $\eta^2 = .038$, but no main effect of distinctiveness, $F(1, 79) = 1.47$, p = .229, $\eta^2 = .004$. No significant interaction between distinctiveness and image type was found, F(1, 79) = .58, p = .450, $\eta^2 = .001$.

**Experiment 2**

*Sensitivity*. For d', there was a significant main effect of image type, $F(1, 79) = 13.89$, $p < .001$, $\eta^2 = .019$. In addition, there was a significant main effect of distinctiveness, $F(1, 79) = 14.12$, $p = .003$, $\eta^2 = .033$. A marginally significant interaction between distinctiveness and image type was also found, $F(1, 79) = 1.82$, $p = .072$, $\eta^2 = .004$. Post-hoc Tukey HSD tests revealed, that participants had increased sensitivity for distinctive eye open ($M = 1.14 \pm 1.68$) over eye closed trials ($M = .60 \pm 1.43$), $t(79) = 4.10$, $p < .001$. For typical faces, there was no difference in sensitivity between eyes open ($M = .99 \pm .95$) and eye closed conditions ($M = .82 \pm .87$), $t(79) = 1.38$, $p = .516$.

*Criterion.* For criterion C, there was no main effect of image type, $F(1, 79) = .62$, $p = .108$, $\eta^2 = .005$ or distinctiveness, $F(1, 79) = .31$, $p = .326$, $\eta^2 = .003$. Additionally, no interaction between distinctiveness and image type was found, $F(1, 79) = .15$, $p = .698$, $\eta^2 = .000$. Participants were similarly liberal in their bias between all conditions (lowest $M = -.01$, largest $M = .03$).

## Appendix B – Correlational Analysis

### Experiment 1a

*Correlations.* Pearson's Correlation coefficients are displayed in Table 6 and 7. Analysis between the CFMT (range = 48-100%, $M = .74 \pm 15$) and the four conditions of the recognition task showed a positive relationship between face processing skills and recognition performance across the conditions of the task. To elaborate, higher CFMT scores predicted better recognition for three of the four conditions (all but distinctive absent images).

Table 6. *Experiment 1a correlations between experiment condition hit rates and the CFMT scores*

| | **Distinctive Absent** | **Distinctive Present** | **Typical Absent** | **Typical Present** |
|---|---|---|---|---|
| **CFMT** | .162 | .311** | .301** | .416*** |
| **Dist. Absent** | - | - | - | - |
| **Dist. Present** | .449*** | - | - | - |
| **Typ. Absent** | .299** | .248* | - | - |
| **Typ. Present** | .200 | .269* | .705*** | - |

*Notes. Hit rate for distinctive absent, distinctive present, typical absent, typical present conditions were correlated with CFMT scores (top row) and each other (remaining rows). *p < .05, **p<.01, ***p<.001*

Table 7. *Experiment 1a correlations between experiment condition false alarm rates and the CFMT scores*

| | **Distinctive Absent** | **Distinctive Present** | **Typical Absent** | **Typical Present** |
|---|---|---|---|---|
| **CFMT** | -.309** | -.229* | -.335** | -.297*** |
| **Dist. Absent** | - | - | - | - |
| **Dist. Present** | .484*** | - | - | - |
| **Typ. Absent** | .356** | .224* | - | - |
| **Typ. Present** | .304** | .180 | .518*** | - |

*Notes. False Alarm rate for the distinctive absent, distinctive present, typical absent, typical present conditions were correlated with CFMT scores (top row) and each other (remaining rows). *p < .05, **p<.01, ***p<.001*

**Experiment 1b**

*Correlations.* Pearson's Correlation coefficients are displayed in Table 8 and 9. Analysis between the CFMT (range = 48-100%, $M$ = .74 ± 15) and the four conditions of the recognition task showed no relationships except for between CFMT scores and distinctive hair present trials

Table 8. *Experiment 1b correlations between experiment condition hit rates and the CFMT scores*

| | **Distinctive Absent** | **Distinctive Present** | **Typical Absent** | **Typical Present** |
|---|---|---|---|---|
| **CFMT** | .218 | .262* | .215 | .197 |
| **Dist. Absent** | - | - | - | - |
| **Dist. Present** | .272* | - | - | - |
| **Typ. Absent** | .084 | -.016 | - | - |
| **Typ. Present** | .030 | .203 | .483** | - |

*Notes. Hit rate for distinctive absent, distinctive present, typical absent, typical present conditions were correlated with CFMT scores (top row) and each other (remaining rows). *p < .05, **p<.01, ***p<.001*

Table 9. *Experiment 1b correlations between experiment condition false alarm rates and the CFMT scores*

| | **Distinctive Absent** | **Distinctive Present** | **Typical Absent** | **Typical Present** |
|---|---|---|---|---|
| **CFMT** | -.243* | -.267* | -.232* | -.237* |
| **Dist. Absent** | - | - | - | - |
| **Dist. Present** | .262* | - | - | - |
| **Typ. Absent** | .428*** | .428*** | - | - |
| **Typ. Present** | .248* | .248* | .511*** | - |

*Notes. False Alarm rate for the distinctive absent, distinctive present, typical absent, typical present conditions were correlated with CFMT scores (top row) and each other (remaining rows). *p < .05, **p<.01, ***p<.001*

**Experiment 2**

*Correlations.* Pearson's Correlation coefficients are displayed in Table 10 and 11. Analysis between the CFMT (range = 48-100%, $M$ = .74 ± 15) and the four conditions of the recognition task showed a positive relationship between face processing skills and recognition performance on the typical face conditions of the task.

Table 10. *Experiment 2 correlations between experiment condition hit rates and the CFMT scores*

| | Distinctive Absent | Distinctive Present | Typical Absent | Typical Present |
|---|---|---|---|---|
| **CFMT** | -.023 | -.114 | .314* | .283* |
| **Dist. Absent** | - | - | - | - |
| **Dist. Present** | .231* | - | - | - |
| **Typ. Absent** | .146 | .042 | - | - |
| **Typ. Present** | .178 | .037 | .520** | - |

*Notes. Hit rate for distinctive absent, distinctive present, typical absent, typical present conditions were correlated with CFMT scores (top row) and each other (remaining rows). *p < .05, **p<.01, ***p<.001*

Table 11. *Experiment 2 correlations between experiment condition false alarm rates and the CFMT scores*

| | Distinctive Absent | Distinctive Present | Typical Absent | Typical Present |
|---|---|---|---|---|
| **CFMT** | -.064 | -.330** | -.333** | -.327** |
| **Dist. Absent** | - | - | - | - |
| **Dist. Present** | .262* | - | - | - |
| **Typ. Absent** | .428*** | .318** | - | - |
| **Typ. Present** | .194 | .292** | .511*** | - |

*Notes. False Alarm rate for the distinctive absent, distinctive present, typical absent, typical present conditions were correlated with CFMT scores (top row) and each other (remaining rows). *p < .05, **p<.01, ***p<.001*