Optimising EvoFIT by Reducing the Number of Faces Shown during Composite Construction

by

Elizabeth H. Jackson

A thesis submitted in partial fulfilment for the requirements for the degree of Doctor of Philosophy at the University of Central Lancashire

March 31, 2023

Firstly, I would like to thank my supervisors. I express my deepest gratitude to my director of studies, Dr Beth Helen Richardson, for her constant patience, kindness and feedback throughout this project. I also could not have undertaken this journey without Professor Charlie Frowd, whose passion and enthusiasm for his topic has been an inspiration to me throughout my time at university and Dr Cristina Fodarella, whose support and advice has been invaluable.

I would also like to thank my parents for their love and compassion, for encouraging me in all my pursuits and for inspiring me to work hard. I would not have had the courage to undertake this journey without your support.

Lastly, words cannot express my gratitude to my wonderful partner, Callum, for your un-wavering and unconditional love, encouragement and understanding throughout my years of study.

The creation of EvoFIT facial composite images enables perpetrators of crime to be identified and subsequently detained. It is therefore important to ensure that the composite construction procedure is optimised to create the most recognisable images possible. During the creation of a facial composite, eyewitnesses view and compare many images of facial shapes and textures to select those which best resemble the perpetrator. However, viewing many images may overwhelm witness working memory, resulting in cognitive overload and resultant memory capacity and decision-making deficits. Yet, there is currently no literature exploring the impact of cognitive load during composite construction. This thesis aims to bridge this gap in the literature by investigating the impact of cognitive load during EvoFIT construction and to investigate the importance of face *Shape* and *Texture* for composite construction to further optimise the construction procedure.

In five experiments, composite images created using different population sizes to manipulate the cognitive load during the construction procedure were assessed for likeness through composite naming and likeness ratings. The results demonstrated that reducing cognitive load during the construction procedure by decreasing the population size and, therefore, displaying fewer face images to participants, was beneficial for composite likeness. Moreover, reducing the population size for selection of the face shape was particularly important, indicating that face *Shape* plays a more important role than face *Texture* in the creation of a recognisable EvoFIT facial composite. Overall, this thesis demonstrates the benefits of reducing cognitive load during EvoFIT composite construction, particularly for selection of the face shape, and develops a theoretically informed construction procedure to increase the number of criminals identified through EvoFIT facial composites.

TABLE OF CONTENTS

Preface

A facial composite is an image of a face created from memory by an eyewitness to crime, with the assistance of a police practitioner. Facial composites are most commonly created in cases with limited evidence (such as CCTV or DNA), where the eyewitness, who is typically *unfamiliar* with the perpetrator, was able to view the face. The resulting image is published in the media and distributed to local police stations with the aim that somebody familiar with the perpetrator will recognise them based on the composite image (Frowd et al., 2012).

To develop the construction procedure, make improvements to the composite system, or improve our understanding of theories of facial recognition, this process is tested and replicated in a laboratory. In this case, participants unfamiliar with a set of famous individuals, for example, sporting personalities or soap stars, act as the eyewitnesses by viewing an image or a video of the individual (referred to as the 'target') and, after a time delay, attempt to recreate their face from memory using the composite system. A separate group of participants who are familiar with the individuals in the target group, then view the composite images and attempt to name the individuals. Composites that are named frequently are deemed to be the most accurate, and those which are not named correctly, are deemed inaccurate. Alternative measures for composite likeness are also used in research; however, composite naming is considered to be the most ecologically valid method.

Composite systems and procedures can vary widely from country to country. In the USA, the well-established collaboration between the police and the media in searching for perpetrators of crime results in a large number of identifications (Miles, 2005). However, not all of these identifications are correct, the consequences of

which can be devastating for innocent people arrested for crimes they did not commit. In 2018, the Innocence Project reported that 69% of 367 cases involved eyewitness misidentification, 27% of which involved the use of a composite sketch (Innocence Project, 2018). In one such case, Kirk Bloodsworth, an American military veteran, was arrested for the murder of a young girl after the police released a composite image depicting a person who looked like him, and he was later identified in a police line-up by witnesses. He was sentenced to death for the murder at his first trial, which was later changed to two life sentences (Junkin, 2005). After nine years in prison, Bloodsworth's innocence was proclaimed through DNA analysis with the support of the Innocence Project, an organisation that works to rectify and prevent wrongful convictions (Innocenceproject.org). Cases such as this have resulted in legal activists in the USA campaigning against the use of facial composites in police investigations due to their potential link with miscarriages of justice.

A reduction in the number of incorrect identifications from facial composites in the UK compared to the US may be caused by differences in the skill level of the police practitioner guiding the construction procedure (Frowd et al., 2007). The literature clearly demonstrates a relationship between a practitioner's level of expertise or experience and the accuracy with which a facial composite resembles the target (Wogalter et al., 1991). In the UK, police forces undergo rigorous training to learn composite construction procedures based on standardised training programs (Association of Chief Police Officers, 2005; Davies et al., 1986). Yet in the USA, there appears to be no standard training programs for officers to learn to use facial composite systems effectively (McQuiston-Surrett et al., 2006). In addition, the number of incorrect identifications may be influenced by the composite systems

utilised in the USA, which are considered outdated compared to the modern, evidence driven systems in the UK (Haridat, 2016).

Facial composite systems used by police forces in the UK draw closely on the psychological literature around face memory (Dianiska et al., 2021). This focus on psychology is particularly true for evolutionary composite systems (e.g., EvoFIT and EFIT 6), which focus on eyewitnesses selecting whole face images as opposed to selecting individual features to replicate day-to-day face recognition. These systems also allow for the creation of more identifiable composite images (Frowd et al., 2007, 2019) which, in turn, lead to more correct identifications (Wogalter et al., 1991).

Nevertheless, even when constructed in ideal circumstances, such as the laboratory with a short time delay between viewing the target and creating the composite image, identification is not 100%. It is therefore imperative to conduct research, such as that described in the present thesis, to continue developing facial composite systems and procedures to optimise composite construction. This will ultimately increase the identification of images produced, and reduce the risk of misidentification.

1

LITERATURE REVIEW

This chapter will outline the research aims of this thesis and summarise the relevant literature to provide a theoretical background for the research. This literature review will focus on three key aspects relevant to this thesis research: facial recognition, cognitive load and the use of facial composite systems. As cognitive load has not yet been used to understand facial composite construction, theories of memory and information processing will be discussed in the context of facial recognition and will be applied to facial composite construction.

To create a facial composite using EvoFIT, eyewitnesses view screens displaying face images and select faces which resemble the target face, typically the face of a perpetrator of crime (Frowd, Hancock et al., 2011). However, it is theorised that viewing many screens of face images during this process may overwhelm working memory. The consequences of overwhelming working memory include impaired memory and decision making, which could have detrimental effects on eyewitnesses'

ability to remember the target face and, therefore, select the most accurate faces during composite construction, resulting in an inaccurate composite image. This thesis aims to reduce the number of screens used during the EvoFIT construction process to test whether easing the strain on working memory will enable eyewitnesses to create more accurate composite images.

Key research questions addressed in this thesis are as follows:

i) What is the optimum number of screens to use during EvoFIT composite construction?

ii) Does cognitive load effect participants' abilities to utilise each stage of the composite construction procedure?

iii) Is face shape or texture more important for the construction of a recognisable composite?

Providing answers to these three questions is important both theoretically, to understand the impact of cognitive load on witness memory and practically, to increase the number of perpetrators of crime that are successfully identified.

Facial Recognition

Face recognition is an automatic, holistic process whereby a face is viewed as a whole instead of a collection of facial features (Richler et al., 2009). Early research attempting to develop and test a model of facial recognition relied on face models involving images of simple, line-drawn faces varying in the shape and distance between features (eyes, nose, and mouth: Valentine, 1999). The value of these face models is that they allow facial features to be altered in size and position easily and in a controlled manner. As face models contain simple, line-drawn features and do not

include detail or shading information, specific points on the face can be readily identified, and these points can be used to describe the face as a set of values on a fixed number of dimensions (Reed, 1972). The ability to define faces in such a way indicates that facial features are represented within a multidimensional space (Valentine, 1991). However, as natural faces are much more complex than schematic face images, simple face models do not help us understand fully how faces are processed in the real world, which is much more complex.

One aspect of face perception that has been important for understanding how we recognise faces in a more naturalistic environment is *distinctiveness*. Faces considered distinctive contain unique or prominent features, or perhaps have an unusual configuration, such as eyes being much closer together or further apart than average (e.g., Jackie Kennedy). Distinctive faces are recognised more quickly than average or typical faces (Benson & Perrett, 1994), although classifying them as faces is slower compared to typical faces (Valentine & Bruce, 1986).

Valentine (1991) proposed a framework to account for the effects of distinctiveness in face processing and recognition. This framework argues that the effects of distinctiveness in face processing can be interpreted by considering faces as being located in 'face space'. In this face space, more average faces (i.e., those with typical-looking features) are located towards the centre and more distinctive faces (i.e., those with more unique or prominent features, or configuration) are located towards the periphery. Distinctive faces (those that reside towards the periphery) are recognised more easily because they appear further from neighbouring faces in the space and so are not so easily mistaken for these other faces. In contrast, average faces (those towards the centre) are more difficult to identify because they must be

distinguished from similar-looking faces that tend to be clustered together (Valentine, 1999).

One method used to increase the distinctiveness of a face image in facial recognition research is caricaturing. Facial caricatures exaggerate a-typical facial features in shape and/or texture, such as lengthening a nose or deepening the colour of rosy cheeks to "individuate" the face, differentiating it from other faces (Perkins, 1975). Caricatured faces are typically rated as more distinctive and are identified more quickly and accurately than unedited faces (Bartlett et al., 1984; Rhodes et al., 1987). On the other hand, facial anti-caricatures minimise a-typical features in shape and/or texture to make them less noticeable, rendering the overall face more typical. Consequently, anti-caricatured faces are identified more slowly and less accurately than unedited faces (Lee & Perrett, 2000).

More specifically, to create a caricature for facial shape, set points on each face are given x-y co-ordinates. The position of each point is compared to a 'face norm', which is created by averaging the position of points on many faces. The distance between each point on the target face and the corresponding point on the average face is then calculated. Caricatures are created by multiplying the difference between each point on the target face and that point on the norm face by a fixed percentage. Therefore, points that were calculated to be in a similar position to the face norm are moved less than points which are further from the face norm. Accordingly, if a point on the target face is positioned 1mm away from that same point on the norm face, and the face was caricatured to 10%, the point on the target face would be moved by 1.1mm to create the caricature (Benson & Perrett, 1991). While research clearly demonstrates the impact of distinctiveness on face recognition,

it is not the only important factor in face processing; the roles of face shape and texture are also vital for recognition of faces.

Facial Shape and Texture

Face shape is defined as the shape of the head, the geometry of individual facial features, and the configuration between them, referred to as second-order configuration. Face texture is defined as luminance, hue, and saturation, which are colour-based properties determined by the reflectance of the skin surface and tissue (Itz et al., 2017). The role of face shape and texture in recognising familiar and unfamiliar faces has been investigated using various methods, typically demonstrating that face shape is the most important for *unfamiliar* faces and face texture is the most important for *familiar* faces.

The importance of face texture for familiar face recognition was demonstrated in Bruce et al. (1991). In their experiment, 'head models' were created by scanning the three-dimensional surface of the face with a laser and using Computer Aided Design (CAD) software to create a 3-D head model. The surface of the head model was manipulated to vary the level of textural information available and the angle at which the head model was manipulated to alter the face shape information. Participants unfamiliar with the individuals on whom the head models were based were invited to match the head models to the target photographs. The results demonstrated that head models with higher surface density, and therefore more textural information, were matched with the target photograph more frequently than those displayed with less surface density. More specifically, head models displayed with 100% surface density achieved 42% correct identification rates, whereas those displayed with 25% surface density achieved 29.6% correct identification rates. This

pattern of results demonstrates the importance of face texture for recognition of unfamiliar faces as head models with more texture displayed were identified more frequently than those displayed with less texture.

The findings of Bruce et al. (1991) are supported by literature on face recognition. In Lee and Perret (1997), a series of experiments demonstrated the importance of colour information for recognition of familiar faces. In one experiment, colour information was removed from face images to produce a greyscale image and participants were invited to identify the familiar individual based on either the coloured or greyscale image. The results demonstrated that correct identification of the familiar identities was significantly higher for the coloured image than the greyscale image, indicating that colour information is somewhat important for familiar face recognition. In Experiments 2 and 3, the colour contrast of the face image was increased, enhancing the colour information (and therefore the textural information) of the face and a 'mask' was applied onto the face images to manipulate the shape of the face to that of an average "face prototype", to reduce the shape information in the face. Participants were invited to identify the familiar identities based on the original face image, the colour caricatured face image or the 'masked' image. The results demonstrated that correct identification rates were higher for the caricatured face images than the original images. Yet, the results also demonstrated little difference in correct identification between the original images and 'masked' images, indicating that face shape is not important for familiar face recognition. In these three experiments, the importance of face texture for recognition of familiar faces is clearly demonstrated, but the final two experiments also demonstrate that face shape is of little importance for familiar face recognition, supporting the results of Bruce et al. (1991).

In Russell and Sinha (2007), photographs of familiar faces were modified to reflect the average face shape or texture based on a "face prototype", similar to that used in Lee and Perrett (1997). Face images were displayed with (i) the original face shape but average face texture, (ii) the original face texture but average face shape, (iii) or the original face shape and texture (unedited face image). Participants were invited to view face images from one of the three conditions and asked to identify the familiar individual based on the images viewed. Images with the original texture were identified more frequently than those with the original shape, a result which was exaggerated for female Caucasian faces, where a very large difference was seen between the two conditions. This pattern of results supports previous findings that textural information is more important than shape information for identification of familiar face images, as images with an average shape were identified more frequently than those with an average texture.

Furthermore, Rogers et al. (2022) created hybrid faces by combining the face shape of one celebrity with the face texture of a different celebrity within one face image. Participants were invited to view the face image and identify the individual depicted. Some participants were familiar with both celebrities, some participants were unfamiliar with both celebrities, while the remaining participants were familiar with just one of the celebrities. Unsurprisingly, the results demonstrated that familiar celebrities were identified more frequently than unfamiliar celebrities, although familiar celebrities were identified more frequently based on face texture than shape. Additionally, the effect size for familiar celebrities identified based on the face texture was far larger compared to that of familiar celebrities identified based on the face shape. This result indicates that face texture is more important for the recognition of familiar faces than face shape.

Despite the evidence demonstrating the importance of face texture for familiar face recognition, it is suggested that face shape also plays a role. In Knight and Johnston (1997), original face images, and negative face images (distorting the texture information), were displayed stationary (as a photograph) or moving (a video of the target speaking in an interview-like situation, during which the movement reveals information about the shape of the face). When the facial texture was distorted, participants identified the familiar targets more accurately based on the video than the static image, demonstrating that, when limited textural information is available, face shape is used for familiar face recognition.

Similarly, Butcher et al. (2011) demonstrated that *unfamiliar* faces were recognised more easily when they were moving than when still. In this experiment, same-race or other-race unfamiliar faces were encoded when moving or when stationary. Participants were then invited to select the 20 faces that had previously been seen from a collection of 40 faces. The results demonstrated that faces which has been viewed in motion were identified more accurately than faces which had been viewed when stationary. This effect, whereby moving faces were recognised more accurately than stationary faces, was consistent for same-race and other-race faces. As in Knight and Johnston (1997), moving faces revealed more information about the 3D structure of the face, displaying more information about the face shape than stationary faces. Thus, information about the face shape plays a role in unfamiliar face recognition, as it does for familiar face recognition in Knight and Johnston (1997).

Benson and Perrett (1991) further demonstrated the importance of face shape for recognition of unfamiliar faces, caricatured the face shape of seven caricatured familiar face images (celebrities). Deviations from the original face were accentuated by a fraction, with results indicating that face images caricatured to +32% were

named faster than those caricatured to +16% and faces that were not caricatured. Further, correct identifications were highest for face images caricatured to +16%. These results indicate that caricaturing face shape may increase the speed with which participants perceive a face but that caricaturing a face too much may decrease likeness of the face image to such an extent that identification is impeded. Therefore, it can be inferred that face shape is important for the recognition of familiar faces because increasing the distinctiveness of the face shape increases the likelihood of recognition; however, moving too far from the original face (i.e., caricaturing to +32%) decreases the likelihood of recognition.

This thesis manipulates the amount of face shape and texture information viewed during EvoFIT composite construction. The literature highlights the importance of face shape over face texture during unfamiliar face recognition (which occurs during facial composite construction). Therefore, it is predicted that face shape will be more important than face texture during the composite construction procedure.

Face Recognition during Facial Composite Construction

Face recognition is an automatic process and, although recognition of familiar faces seems effortless, recognising unfamiliar faces is a difficult task for humans to perform well (Bruce & Young, 1986; Burton et al., 2015). While difficulty in recognising unfamiliar faces does not impact on our day-to-day lives, it does become important in a situation where an eyewitness must create a facial composite of a perpetrator of a crime or, identify a perpetrator in a line-up. Eyewitnesses are often invited to pick out a perpetrator from a line-up once an identification has been made based on the composite image created by the same eyewitness. Consequently, it is important that the process of composite construction does not disrupt the eyewitness's memory of

the target so much that they are unable to later identify the perpetrator. However, there is conflicting evidence for the impact of composite construction on witness memory (Cornish, 1987; Davies et al., 1978; Kempen & Tredoux, 2012).

Davies et al. (1978) demonstrated the benefit of inviting participants to create a facial composite prior to identifying a target in a line-up. In this experiment, participants viewed a target photograph, and created a facial composite after either an approximate 48-hour delay, a three-week delay or not at all (control group). Participants then viewed 30 photographs sequentially, one of which was the target. In both the 48-hour condition and the three-week condition, participants who created a facial composite were subsequently more likely to correctly identify targets than participants who did not create a facial composite.

However, this finding was not replicated in future research. In Cornish (1987), participants viewed a composite image which had been created by a researcher using Identikit, an early facial composite system, after which two thirds of the participants attempted to recreate the target composite using the same system (Groups 1 and 2), and one third of the participants did not, acting as a control (Group 3). For each participant in Group 1, researchers created five facial composites using Identikit which resembled the composite image. These five composites were displayed alongside the target composite, and the participant was invited to select the original target composite image from the six images viewed. For each participant in Groups 2 and 3, five composite images created by researchers (randomly selected from the pool of images used for Group 1), as well as the original target composite were displayed. Participants in these groups were also invited to view the six composite images and select the original composite images from those displayed.

The results demonstrated that participants who created an Identikit composite and viewed foils based on their composite (Group 1) made 86% incorrect choices when selecting the target composite from the six images. Participants who created an Identikit composite and viewed foils based on a different composite (Group 2) made 78% incorrect choices when attempting to select the target composite from the six images. Yet, participants who did not create a composite and viewed foils based on a different composite (Group 3) made 56% incorrect choices when selecting the target composite from the six images. This study demonstrates that creating a facial composite of a target results in more incorrect judgements when later attempting to select said target from a line-up. This result is supported by Wells et al. (2005), who demonstrated that only 10% of participants who created a composite of a target using FACES were later able to recognise the target from a line-up, compared to 44% of participants who viewed the facial composite and the target (but did not create one themselves) and 84% of participants who viewed the target only.

Kempen and Tredoux (2012) demonstrated the limitations of viewing many face images on the ability to accurately recognise a 'learned face' through police line-up research. In this experiment, Kempen and Tredoux compared witnesses' ability to recognise a target face in a police line-up between three conditions. In the first condition, participants viewed the target and recreated a composite of the target face, using the facial composite system FACES. In the second condition, participants viewed the target and the facial composite constructed by participants in the first condition. In the third condition, participants viewed the target face only. Witnesses who had created a facial composite of the target were less able to later recognise the target in a line-up than witnesses who had only viewed the facial composite or only viewed the target.

One explanation for the negative impact of creating a facial composite on witness memory may be due to the large number of face images, or partial face images viewed during the construction procedure. Lindsay et al. (1994) suggest that eyewitnesses are less able to recognise a target after creating their facial composite due to viewing too many face images. In this experiment, participants watched a video of a staged crime and promptly described the target. Fifteen minutes later, participants were asked to sort through mugshots to find the target face. In both groups, the target mugshot appeared as the 150th image, but participants were asked to view all of the mugshots before presenting their chosen image. Participants who viewed 510 mug shot images were less able to select the correct target than participants who viewed 200 mugshot images. This finding indicated that viewing many images of faces interferes with participants' abilities to select the correct target face, perhaps because individuals become “overloaded” while viewing a great many faces, and this affects their memory or performance.

Facial composites constructed in the above experiments (see, Cornish, 1987; Davies et al., 1978; Kempen & Tredoux, 2012) were done so using featural facial composite systems, as opposed to a holistic system which reflects natural day-to-day recognition of faces (Frowd, Pitchford et al., 2012). The use of featural systems in these experiments means that participants must select individual facial features (such as eyes, nose and mouth) in isolation, which are then combined to create a face image. Therefore, the reason for poor identification rates of targets in a line-up after the creation of a facial composite may be due to the unnatural process of creating a facial composite using a featural system, which may impair the participant’s memory of the target face (Kempen and Tredoux, 2012; Wells et al., 2005, 2007). Furthermore, facial composites constructed using a featural system are typically less accurate than those

constructed using holistic system (see Zahradnikova et al., 2018; Frowd et al., 2015 for a review of featural and holistic composite systems). If a participant creates a facial composite image that is a poor representation of the target, they may not recognise the target from a line-up as they are seeking to identify the individual from the composite image, and not the target.

To mitigate these problems, Davis et al. (2014) explored the impact of composite construction using a modern, holistic system (EFIT-V) on line-up identification. Participants created a facial composite of a target using either E-FIT (featural system) or EFIT-V (holistic system) before being invited to select the target identity from a video line-up. In a target-present line-up, 64% of participants who created a composite using E-FIT correctly identified the target, with 18% selecting a foil (a different individual in the line-up). Conversely, 70% of participants who created a composite using EFIT-V correctly identified the target, with only 10% selecting a foil. Although the difference in correct naming between these two groups was small, the finding indicates that composite construction using a holistic system does not disrupt the memory of the target as much as composite construction using a featural system. This pattern of results was replicated in a target-absent line-up, 59% of participants who created a composite using E-FIT selected a foil, and 41% correctly stated that the target was absent, whereas 56% of participants who created a composite using EFIT-V selected a foil, with 44% correctly stating that the target was absent. Interestingly, participants who viewed the target, but did not create or view a facial composite, performed poorly compared to participants in the two experimental groups, with 45% of participants correctly identifying the target (Davis et al., 2014).

In a second experiment, a 30-minute delay was implemented between participants viewing the target and creating a facial composite using EFIT-V (cf. no

delay in the first experiment). Forty-eight percent of participants who created a facial composite correctly identified the target from the line-up, with 17% selecting a foil and 34% incorrectly stating that the target was absent. In comparison, 35% of participants who did not create a facial composite correctly identified the target from the line-up, with 31% selecting a foil and 34% incorrectly stating that the target was absent (Davis et al., 2014). Again, although the difference in percentage of correct identification between participants who created a facial composite and those who did not was small, creating a composite of the target was somewhat beneficial for remembering the target. Although this experiment extended the period of time between participants viewing the target face and creating the facial composite to 30-minutes, a short period of time compared to the 24-hour delay that is considered reasonable for composite construction with the police.

In a somewhat similar experiment, Pike et al. (2019) also demonstrated that target identification in a police line-up is most accurate after construction of a facial composite using a holistic system (EFIT-V) compared to a featural system (E-FIT). In this experiment, all participants viewed the target face and, 2-days later, returned for the second part of the experiment. In the second part of the experiment, one third of the participants did not create a facial composite, and went straight into the identification task, one third of the participants created a facial composite using E-FIT and one third of the participants created a facial composite using EFIT-V. 73% of participants who created a composite using EFIT-V were able to correctly identify the target, with 10% of participants selecting a foil, 65% of participants who created a composite using E-FIT correctly identified the target, with 12% selecting a foil, and 61% of participants who did not create a facial composite correctly identified the target, with 12% selecting a foil. These results demonstrate that, even with an

ecologically valid delay of 2-days between viewing the target photograph and creating the facial composite, using a holistic facial composite system such as EFIT-V to create the composite image is beneficial, not harmful for composite construction.

The similarity in line-up findings was highlighted in Sporer et al. (2020) and Tredoux et al. (2021). Sporer et al. (2020) analysed the correct and incorrect identifications (target present and absent) in 15 experiments whereby participants attempt to select a target from a line-up after creating a facial composite. The results revealed little effect of composite construction on correct and incorrect line-up identifications. However, the findings also indicated that viewing somebody else's misleading composite image (i.e., a composite image that does not resemble the target) may reduce correct line-up identifications. In a meta-analysis of 23 studies, Tredoux et al. (2021) demonstrated there was no significant effect of composite construction on the accuracy of target identification. In support of these findings, Tsourrai and Davis (2020) revealed no significant impact of composite construction using EFIT-6 (an updated version of E-FIT V) on the accuracy of line-up identification.

These experiments (Davis et al., 2014; Pike et al., 2019; Tredoux et al., 2021; Tsourrai & Davis., 2020) demonstrate that viewing many faces during composite construction had no negative impact on the recognition of a target in a police line-up. However, it is important to recognise the differences between the system used in the experiment and the system used in this thesis. One large difference between EFIT-V and EvoFIT is the number of faces that are presented in each face-array, as well as the number of face arrays viewed during the construction process. During composite construction using EFIT-V, participants view nine faces per screen, and as many screens as they would like to create the composite image. On the other hand, during

composite construction using EvoFIT, participants view 18 faces per screen and 20 screens throughout the construction process (360 unique faces). For participants creating a facial composite using EFIT-V to view 360 face images, they would need to view 40 face arrays during the construction process. As the number of screens viewed during E-FIT-V is decided by the eyewitness, it is possible that, if participants start becoming fatigued and struggle making decisions or remembering details of the target, they may halt the face selection process, accepting the composite image as the best likeness, and limiting the negative impact that viewing too many faces has on their ability to later select the target from a line-up. In contrast, during EvoFIT construction, participants view a set number of face arrays, and so cannot halt the face selection procedure upon feeling fatigued.

To date, no published research has explored the ability of eyewitnesses to recognise a target from a line-up after constructing an EvoFIT composite. However, Frowd and Grieve (2019) demonstrated that reducing the number of face images viewed during the EvoFIT construction procedure resulted in the production of more identifiable composite images. It may be theorised that participants who viewed many face images during the construction procedure were less able to utilise the tools designed to enhance composite likeness, resulting in less identifiable composite images. To explore this result in depth, the current PhD research aims to replicate the methodology in Frowd and Grieve (2019) with a crucial difference, a further reduction in the number of face images viewed.

Information Processing

To understand the impact of viewing many face images during the process of facial composite construction, it is important to first develop an understanding of information processing in memory. In 1968, Atkinson and Shiffrin designed the Modal Model of Memory, more commonly referred to as the Multi-Store Model of Memory, which contains three separate memory stores: the Sensory Register, the Short-Term Store and the Long-Term Store. The Sensory Register contains five separate registers, each receiving inputs from one of the five senses, the Short-Term Store receives input from the Sensory Resister, and the Long-Term Store and can typically hold between five and nine pieces of information (Cowan, 2001; Miller, 1956) for up to 20 seconds (Peterson & Peterson, 1959). Information in the Short-Term Store is lost through displacement or decay and remembered through rehearsal. The Long-Term Store, in contrast, holds information indefinitely and has unlimited capacity. The Long-Term Store encodes information semantically, and therefore, information which is given meaning is stored here. According to this model, and relevant to the current project, faces which are momentarily viewed but bear no importance to the current context are briefly stored in the Sensory Register. In contrast, faces which bear some importance for a brief period of time (e.g., a cashier in a supermarket) are stored in the Short-Term Store; and faces which are considered relevant to the current context, such as those of a friend, family or colleague are stored in the Long-Term Store.

Despite support for the Multi-Store Model of Memory, including its influence on theories such as the Serial Position Effect (Glanzer & Cunitz, 1966; Murdock, 1962) and explanations for well-known cases of memory loss such as HM (Schoville & Milner, 1957) and KF (Shallice & Warrington, 1970), it has been argued that this

explanation of the Short-Term Store may be too simplistic. These critics argue that the Short-Term Store should be divided into multiple components so that information can be processed as well as stored (Baddeley & Hitch, 1974), a notion that was addressed by Atkinson and Shiffrin (1968) but lacked evidence at the time the model was created.

To address the oversimplicity of the Short-Term store, Baddeley and Hitch (1974, 2010) proposed a model of Working Memory which has different systems for different types of information (e.g., auditory memory and visuo-spatial memory). Working memory is a brain system that provides temporary storage and manipulation of information necessary for complex cognitive tasks such as language comprehension, learning and reasoning (Baddeley, 1992; Baddeley & Hitch, 1974). The three sub-components that make up working memory are the Central Executive, which is the attentional controlling system; the Phonological Loop, which stores and rehearses speech-based information, and the Visuospatial Sketch Pad, which manipulates image information. The Central Executive also coordinates information form the Phonological Loop and the Visuospatial Sketch Pad, which are often referred to as the 'slave' storage mechanisms (Baddeley & Hitch, 1974, 2001, 2010).

Various theories attempt to explain how information is perceived and processed in working memory. Two popular views that emphasise the importance of working memory and the sensory registers are Early Selection and Late Selection. Early Selection states that perceptual processing capacity is limited and that only information attended to is perceived. However, Late Selection states that perception is an automatic process with unlimited capacity; therefore, perception of information is mandatory and cannot be prevented at will (Pohl et al., 2010).

Measures such as the dichotic listening paradigm for hearing and the selective looking paradigm for viewing demonstrated that unattended information often goes unnoticed, supporting the Early Selection view (Bookbinder & Osman, 1979). Selective-attention tasks, such as the Flanker Task, during which a participant must state the direction a target arrow is facing while ignoring the direction of adjacent arrows, which may be facing in the same or opposite direction as the target arrow (Eriksen & Eriksen, 1974), demonstrated slower responses with irrelevant distractors present. This result supports the Late Selection view, as performance in the relevant task suffers when irrelevant information is present, suggesting that irrelevant information cannot be ignored. A solution to the Early and Late Selection debate may be a hybrid model of attention proposed by Lavie (1995, 2001), coined the Perceptual Load Theory, which encompasses mechanisms of both the Early and Late Selection.

The Perceptual Load Theory states that the level of perceptual load, which is the amount of information involved in processing task-relevant stimuli, dictates the efficiency of selective attention. The higher the perceptual load, the more likely we are to process irrelevant distractors (Lavie, 1995, 2010). When under high levels of perceptual load, eyewitnesses are likely to still remember key details about a scenario, for example, remembering information about a central character. However, they are less likely to remember peripheral information, such as information about a seemingly unimportant person walking past. Furthermore, eyewitnesses under high perceptual load are more receptive to suggestion, showing increased susceptibility to leading questions and were less likely to remember auditory details during the crime scenario (Murphey & Greene, 2016).

An alternative to the Perceptual Load Theory (Lavie, 1995; 2010) is the Theory of Visual Working Memory, which focuses solely on visual information. This

theory defines Visual Working Memory as the active maintenance of visual information to serve the needs of an ongoing task; that is, the amount of visual information that can be maintained in memory at once (Luck & Vogel, 2013). Visual Working Memory stores information about the position, shape, colour and texture of items, but appears to be limited to between three and five simple objects (Xu, 2002); however, this number may vary depending on the type of item, pattern of items and the task (Brady et al., 2011).

The Theory of Visual Working Memory suggests that familiar and unfamiliar faces are stored differently, which may impact their real-time identity processing (Gamborata & Sessa, 2019). Evidence for this phenomenon comes from change detection tasks, whereby faces are displayed before a retention interval, after which the faces are displayed again but with a change to one or more of the faces (Woodman et al., 2012). Changes to familiar faces were identified more quickly and accurately than changes to unfamiliar faces. However, as faces are complex objects which change depending on viewing angle and facial expression, fewer faces are typically stored in the Visual Working Memory than simple objects such as shapes. More specifically, only one to two faces may be stored in the Visual Working Memory at one time (Jackson & Raymond, 2010). In relation to a face recognition task, such as that during the first stage of EvoFIT composite construction, discussed in detail later, the notion that only one or two faces are stored in working memory at one time indicates that selecting faces from a screen displaying many options (here, 18 per screen) may be a difficult task, particularly when the task is repeated several times consecutively. However, alternative information processing theories may offer a more optimistic view.

The Cognitive Load Theory by Sweller (1988, 2010) states that working memory load is limited and that an overwhelming load on working memory reduces memory capacity and decision-making ability, reducing overall performance on a task. The Cognitive Load Theory states that 'load' comes from three sources: intrinsic load, extraneous load and germane load. Intrinsic load results directly from the task's complexity and depends on the interactivity of elements in the task and the learner's prior knowledge (Klepsch et al., 2017). The elemental interactivity refers to the number of elements that the individual must process simultaneously in working memory while completing the task.

Low elemental interactivity indicates that few elements must be processed simultaneously during the task, whereas high elemental interactivity indicates that many elements must be processed simultaneously (Sweller, 2010). Extraneous load is dependent on the instructional design of the material. High extraneous cognitive load indicates that the learner is investing mental resources into a processes irrelevant to the task itself, such as searching for or repressing information (Klepsch et al., 2017). Germane load is associated with the number of working memory resources devoted to facilitating learning and transferring information from working memory into long-term memory as well as connecting newly learned information to pre-existing information (Paas et al., 2003; Sweller, 2010).

To date, the impact of cognitive load has been explored thoroughly with relation to education. However, it is also important to understand the impact of cognitive load on other types of complex tasks where the results can have a large impact on society, for example, the creation of a facial composite image. Therefore, the following sections will present the relevance of cognitive load in the context of facial composite construction.

Facial Composite Systems

A facial composite is a portrait of a sought individual created by a forensic practitioner based on the memory of the face held by a witness or victim of a crime (McQuiston-Surrett et al., 2006). The resulting image is published in newspapers and via other forms of media and may be distributed to local police stations (Frowd & Hepton, 2009). The purpose of a facial composite is to identify a perpetrator based on the recognition of the individual by somebody familiar, such as a well-known neighbour, a colleague, family members or a police officer recognising the individual from a previous crime (Frowd et al., 2012).

Facial composites are most often used when a perpetrator is unfamiliar to the eyewitness, and when the eyewitness is able to see the face of the perpetrator (i.e., the perpetrator is not wearing a full-face covering) and can be used alongside other evidence. However, the creation of a facial composite image is particularly useful when there is a lack of CCTV footage, or where the face has been concealed in the footage, or where there is no DNA evidence available. Put simply, a composite is needed when there is no identifying information available, or when recovery of this information would take so long as for other crime to be committed. As a facial composite is often the only image available of the perpetrator, it is important that facial composites images are as identifiable as possible. An overview of the different types of composite systems are outlined below.

Sketch

The earliest police facial composites were hand drawn by a forensic or sketch artist. Following an in-depth cognitive interview (CI) to obtain a detailed verbal description of the face of the perpetrator from a witness, an artist would create an initial sketch (Frowd et al., 2015). The eyewitness would then view the sketch and suggest changes; revisions are made to the image and this process is repeated until the eyewitness is satisfied that the best likeness has been achieved.

The interview used is typically the CI, initially designed by Geiselman et al. (1986) to aid eyewitnesses in recalling the details of crimes. The overall aim of the CI is to obtain the most accurate, detailed and thorough description of the offender, while minimising recall of false information. Free recall with follow up questions from the interviewer to gain more detailed information is used often to obtain a description of the face (Fodarella et al., 2015). Before the introduction of the CI in standard practice, eyewitness reports were often incomplete, unreliable, and malleable during questioning, a procedure attributed to many miscarriages of justice (Sobel & Pridgen, 1982).

A strength of the CI approach is that the interview can vary in depth, depending on the scenario (Beatty & Willis, 2010). For example, CI to gain a description of a perpetrator for the creation of a forensic sketch is more detailed than that for construction of a composite using other modern composite construction techniques (e.g., E-FIT or EvoFIT, described below). The interviewer creating the composite sketch must obtain a description of the offender that is detailed enough from which to create a drawing. In comparison, practitioners using a mechanical or computerised composite system have a more secondary role in the outcome of the composite image, as facial features or whole face images are not hand-drawn by the

practitioner as they are for sketch images. Therefore, it is important for a sketch artist to gain an accurate, detailed description of the offender compared to practitioners using mechanical or computerised composite systems.

Conversely, the detail obtained in the CI may be somewhat less important if the artist utilises a reference catalogue of sketched faces or individual facial features for the eyewitness to select that resemble the offender (Kuivaniemi-Smith et al., 2014). The literature demonstrates that viewing many similar faces during modern composite construction (Kempen & Tredoux, 2012) or a police line-up (Lindsay et al., 1994) can negatively affect the witness's memory of the target face by overloading the working memory. Therefore, it may be reasonable to presume that viewing previously sketched faces or facial features in a reference catalogue can also interfere with the accuracy of the witness's memory of the target face.

An alternative technique for creating a sketch that eliminates the need to view multiple sketched faces or features is outlined by Nejati and colleagues (2011). In this method, an eyewitness is invited to draw a sketch of the target face and to provide necessary information about the face, such as the sex, age, race as well as selecting the colour of their skin, hair and eyes from colour palettes. The eyewitness is then asked to draw a set of faces, typically three or four, with photographs present. The face sketches are mapped using computer software and inconsistencies between the photographs and the drawings of photographs are calculated, revealing the eyewitness's individual biases. Using this knowledge, the sketch of the face can be edited to remove individual biases.

Despite potential problems that can arise from viewing example faces or facial regions, freehand drawing can be advantageous compared to other composite systems. A sketch or forensic artist has more freedom over the appearance of the composite

image and can alter the face in any way to ensure that the eyewitness is satisfied that the best likeness has been achieved (Homa, 1983). This artistic freedom to make changes to the composite image may explain why sketched composite images are typically morrecognisable than those created using alternative facial composite systems (Frowd, Carson, Ness, Richardson et al., 2005). In a comparison between composites created by a sketch artist and those created using a variety of systems, sketches were the most accurate.

Mechanical Systems

In the 1960s and 1970s, mechanical-driven systems were introduced. These systems rely on selecting facial features that resemble the target and positioning them to create an image of a face (Davies et al., 2000), and are thus referred to as featural systems. To create the facial composite using a mechanical system, an eyewitness works with a technician who assists in the construction the composite image. Two popular mechanical systems were Photofit and Identikit. Identikit required the witness to select line drawn facial features on transparent slides that resemble those of the target, and these were stacked on top of each other to create an image of a face (Davies et al., 1978; Garcia-Solley, 2019). Alternatively, Photofit required the witness to choose printed black and white photographs of features and place these into a mechanical frame to create an image of a face. A clear slide could be placed on top and used to add unique details, such as a mole or scar (Garneau, 1973).

One strength of mechanical systems is that they are simple to use and do not require extensive training or talent (cf. forensic artists). When using Photofit, the features slot into a mechanical frame, and so there is no skill required to position the features according to the witness's description (Lindsay et al., 2013). Furthermore, the

mechanical frame means that the task of composite construction may also be easier for eyewitnesses, which may aid in the construction of an accurate composite image. When using Photofit or Identikit, eyewitnesses select facial features that best match their memory of the target, and so features are not drawn to match the witness's description for each composite, as they are for forensic sketches. However, some skill on the part of the Identikit and Photofit 'technicians' may be required to accurately draw unique details onto the clear slide (Gibling & Bennett, 1994).

Still, there are several limitations to using a mechanical system. Many studies have demonstrated that both Photofit and Identikit produced composites of poor accuracy, based on the likeness judgements and inaccurate naming of composite images (for example, see, Ellis, 1975, Ellis et al., 1978; Laughery & Fowler, 1980). Furthermore, the accuracy of composites constructed using Photofit in comparison to other composite systems is demonstrated on Page 41 (Frowd, Carson, Ness, Richardson et al., 2005).

In Laughery and Fowler (1980), composites were created using Identikit or Sketch immediately after an 8-minute exposure period or with the target present. In both conditions, composites drawn by a sketch artist were more accurate than those created using Identikit with a Technician. Furthermore, there was little difference in accuracy between Identikit composites constructed with the target-present and the target-absent. This result suggests that, even an Identikit created under ideal conditions (i.e., with immediate construction) does not accurately resemble the target. Such a finding indicates that perpetrators of crime are unlikely to be identified based on a facial composite image constructed using Identikit.

In a similar study by Ellis and colleagues (1978), participants were asked to recreate two facial composites that had been constructed using Photofit- one with the

composite image present and one from memory. The results revealed little difference in composite accuracy between composites created with the target-present and absent. This finding suggests that participants cannot create an accurate face image using the Photofit system, even when the target image is available and created using the exact features available for the participants to select. Further, Photofit facial composites created with the target present were less accurate than face images sketched by untrained participants in the same conditions.

One explanation for the limitations of Identikit and Photofit is the limited number of features from which to choose when creating a composite (cf. sketches limited only by the artist's ability). Each system contained multiple options for the composite: eyes, nose, mouth and external features such as hair and ears. However, the number of available features was not enough to accurately represent all faces in general (Laughery & Wogalter, 1989). In addition, the features that were available could not be altered in size, shape or spatially. Therefore, eyewitnesses were unable to create as accurate a composite image as possible. A second explanation for the poor results by mechanical systems is the focus on the method of selection of individual facial features (McIntyre et al., 2016). As previously explained, people process faces holistically and struggle to accurately choose isolated features (Taylor, 2012). This type of featural system does not reflect day-to-day facial recognition, limiting the construction of accurate composites.

A further explanation is the lack of organisation in the system. Here, eyewitnesses will inevitably view many facial features as part of face construction. Although untested, it may be appropriate to apply information processing theories here and these may suggest that viewing too many facial features as a result of the lack of organisation may overwhelm working memory, reducing memory and

decision-making ability. Rectification of these issues was attempted with the development of computerised composite systems.

Computerised Systems

The digitalisation of information led to the development of computerised facial composite systems. Typically, such systems consist of isolated facial features, which are selected individually, or in the context of a whole face, by an eyewitness. As such, these systems are also referred to as featural systems.

The first promising computerised facial composite system was Mac-a-Mug Pro. This system was developed using line-drawn features, similar to those used in Identikit. However, Mac-a-Mug Pro had a vast library of features and accessories (McQuinston-Surrett et al., 2006). The process of composite construction using Mac-a-Mug Pro is similar to that of the aforementioned mechanical systems. The first step is to input a description of the target into the computer system, after which features matching the description are displayed on screen and eyewitnesses select facial features that most accurately represent the target from those displayed. The selected features are assembled and displayed as a face on a screen, ready for further manipulation (Koehn & Kisher, 1997). Examples of manipulation include ageing the face using appropriate lines and adding unique features such as moles and tattoos. More sophisticated changes can also be made, such as enlarging individual features using the MacPaint graphics program. According to the manufacturers, the flexibility to move features independently can create almost 100 times the number of faces compared to mechanical systems (Kovera at el., 1997).

Initial studies examining the effectiveness of composites created using Mac-a-Mug Pro showed promising results. Cutler and colleagues (1988) praised the

computerised system, claiming the potential for realistic, recognisable composites to be created. However, composites in this experiment were created with direct reference to a facial photograph (cf. from memory). Further research demonstrates that, although Mac-a-Mug Pro can create an accurate likeness when referring to a photograph, improving on results from mechanical systems, when created from memory, Mac-a-Mug Pro composites were not recognisable (Koehn & Fisher, 1997; Kovera et al., 1997; Wogalter & Marwitz, 1991).

A similar alternative system is FACES (FACES Software, n.d.). Composite construction using FACES is similar to that using Mac-A-Mug Pro; however, FACES contains many more of features to select from, for example 934 pairs of eyes, 1154 noses and 915 lips (Tredoux et al., 2021). Unfortunately, the increased number of features does not appear to assist in accurate composite construction. Composites created of familiar (famous) individuals were named approximately 15% of the time (Masip et al., 2012). However, composites created of unfamiliar individuals were identified far less accurately. When created in a scenario more reflective of police composite construction (i.e., approximately 24 hours between encoding and construction), no composite images were named correctly (Frowd, Carson, Ness, McQuinston-Surrett et al., 2005; Frowd, McQuinston-Surrett et al., 2006), or correct naming has been found to be very low (Frowd, Carson, Ness, Richardson et al., 2005).

A more commonly used facial composite system was E-FIT, developed by Aspley Ltd in 1993 and widely used by British and European police. E-FIT uses a computerised version of the Photofit Library which, as previously discussed, contains photographic images of facial features. As with Mac-A-Mug Pro, a benefit of a computerised system is that hair and face modifications can be made (e.g., changing the shape, size and configuration of features), as well as unique, specific adjustments

using graphics packages (Sullivan, 2007). Such changes could not be readily made to the photographic images when creating a composite using Photofit. This ability shows a clear advantage of using a computerised system instead of a mechanical one. In a direct comparison (see, Davies et al., 2000), E-FIT and Photofit composites were created with familiar targets present or absent. When constructed with the familiar target present, the average correct naming score of E-fit composites was 58% of the 12 images but was only 30% for Photofit. Yet, when the target was absent, the average correct naming score for was only 21% for E-fit composites and 22% for Photofit composites out of the 12 images (Davies et al., 2000). Composites constructed from unfamiliar targets using either system were not named correctly at all. This finding indicates that, although E-FIT has an improved ability to create an accurate composite, the mechanisms had not been optimised to perform successfully in a realistic scenario. A similar outcome was found in Frowd, Carson, Ness, McQuinston-Surrett and colleagues (2005).

A second popular computerised facial composite system is PRO-fit. The general design and idea of PRO-fit is very similar to that of E-FIT, and both systems were popular in the United Kingdom. The similarities between the two systems include using a single face, with features able to switch in and out (Frowd, McQuinston-Surret et al., 2007). This method is typical of a modern-day feature system as the focus is on selecting and altering facial features independently rather than selecting and altering the face as a whole. Another similarity between PRO-fit and E-FIT is the disappointing outcome of composites constructed from memory. A PRO-fit constructed after a series of offences committed in the early 2000s failed to result in a conviction. This failure led to the creation of an alternative facial composite image using an evolutionary system called EvoFIT (Frowd et al., 2004), which was

deemed superior by the victim who constructed both composites (Frowd & Hancock, 2007: a full description of EvoFIT is included below). Despite the poor result in this case, E-FIT and PRO-fit have produced identifiable results in lab-based settings.

When comparing the accuracy of E-FIT or PRO-fit composites with those created using alternative systems, E-FIT and PRO-fit composite images were the most accurate, with E-FIT just outperforming PRO-fit, sketch, Photofit and EvoFIT (Frowd, Carson, Ness, Richardson et al., 2005). However, when the retention period between viewing the target and creating the composite was extended to 48 hours, the accuracy of composites constructed using both E-FIT and PRO-fit reduced considerably (Frowd, Carson, Ness, McQuinston-Surrett et al., 2005).

An explanation for the poor identification rates from computerised systems in realistic settings may be their focus on selecting individual features rather than whole face images (Kovera et al., 1997). As humans process faces holistically in day-to-day life, focusing on the individual internal features (eyes, nose and mouth) during composite construction is not optimal. However, to mitigate this limitation, some aspects of E-FIT and PRO-fit can be considered holistic in nature, such as selecting and adjusting facial features in the context of the whole face, rather than selecting and altering isolated features before assembling the face (Frowd & Hancock, 2008). A further limitation of featural systems is that the eyewitness must view many versions of each feature during composite construction while comparing between the different features included and comparing to the target face from memory. More recent developments in evolutionary systems focus on viewing and selecting whole faces during the construction process.

Evolutionary Systems

Evolutionary facial composite systems attempt to emulate day-to-day facial recognition by displaying whole faces (or whole-face regions) for eyewitnesses to select (Frowd, 2001). These systems are also referred to as Darwinian facial composite systems, as evolutionary algorithms are used to ‘breed’ together faces selected by eyewitnesses to create further generations of faces for selection.

One evolutionary system, EFIT-V (now updated to EFIT-6), aims to create realistic and accurate facial composites based on the witness' ability to perform various facial processing tasks (George et al., 2008). Unlike featural systems, eyewitnesses using EFIT-V start by selecting the target's hairstyle and then view nine faces, selecting the face that most resembles the target. Using an evolutionary algorithm, nine new faces are generated based on the 'best face' selected. The process is repeated until the witness is satisfied that the best likeness has been achieved. Adjustments can be made to this face image, such as resizing or repositioning features and altering the age of the face. Moreover, features deemed to be accurate by the eyewitness can be ‘locked’, so they do not change during these adjustments. (Valentine et al., 2010). Further changes can also be made to the face image using graphics packages such as Adobe Photoshop or Corel Paint Shop (Solomon et al., 2012).

Composite construction using EFIT-V is quicker than that using other systems (Davis et al., 2010), so eyewitnesses have time to create multiple composite images of the same target. When composites are morphed into one image, the likeness may be better than any of the individual composite images. In each composite, there are errors that occur; however, the errors between composites are unlikely to be correlated. Therefore, morphing multiple composites randomly distributes the errors, producing a

composite image that is a better likeness than any individual composites (Valentine et al., 2010).

A final evolutionary facial composite system is EvoFIT (Frowd, 2001). As with other evolutionary systems, the holistic nature of the construction process means that witnesses do not need to have a good recall of an offender's face to construct an EvoFIT and do not need to remember each facial feature individually. EvoFIT differs from other holistic systems by presenting two different types of faces for shape and texture. To create an EvoFIT composite, eyewitnesses first view screens containing "smooth faces". Smooth faces are face images which focus on the shape, position and configuration of the features. The face shape is selected by selecting "smooth faces" which best resemble the target on each screen of 18 face images; once six faces have been selected (typically from four screens) eyewitnesses then select the "best face" from the six faces selected. Afterwards, shading is displayed on the face to alter the appearance of the face texture, and the process of face selection is repeated.

The same process of selecting six faces is repeated, and the "best face" from the six is selected, after which whole-face alterations are made using *Holistic Tools* such as age, weight and masculinity, as well as unique changes to individual features using the *Shape Tool*. Hair and external features are then selected. As with many facial composite systems, changes can be made to the face using graphics packages such as Adobe Photoshop and Corel Paint Shop (EvoFIT, 2021).

EvoFIT is the system used to create facial composites during the five experiments in this thesis. EvoFIT has been extensively shaped by research, and much of the process is understood and has been refined. Furthermore, a system that has good correct naming after a forensically relevant delay of at least one day is required and, currently no other holistic systems is able to demonstrate this (i.e., in a published

research paper) using the gold standard procedure. Furthermore, the EvoFIT composite construction procedure typically displays a set number of faces (72) over a set number of screens (four). This provides a baseline from which the number of screens can be manipulated. Other holistic facial composite systems display fewer faces over more screens, but the number of screens is decided by the eyewitness. For example, if a composite image is judged to be the 'best likeness' early on during face construction, the eyewitness may view few screens, but if the composite image is not judged to be the 'best likeness', an eyewitness may view more screens. As with EFIT-V, EvoFIT uses an evolutionary algorithm to display accurate face images depending on the faces selected throughout the construction process.

Evolutionary Theory

Evolutionary algorithms (EAs) are general-purpose search techniques (Holland, 1975, 1992). The traditional theory of EAs assumes that EAs work by discovering, emphasising, and recombining "building blocks" of solutions in a highly parallel fashion. The Building Block Theory states that most of what we know about the world depends on descriptions and mechanisms constructed by elementary building blocks. Therefore, building blocks are a common feature important for all levels of human understanding, including science and innovation, and recognisable everyday objects. For example, trees are made up of leaves, branches, and a trunk, and the English written language is made up of the 26 letters in the alphabet.

Two main characteristics define building blocks. First, they must be easy to identify (once they have been discovered) and second, they must be easy to recombine to form a wide variety of structures, as can be done with actual children's building blocks (Holland, 2000). The general idea is that good solutions tend to be made of

good building blocks (Mitchell, 1998). In 1975, Holland introduced the notion of schemas (schemata) to formalise the idea of building blocks. In this system, schemas behave like a pattern-matching device. Strings are created using a template of ones, zeros, and asterisks, also known as the ternary alphabet (Mitchell, 1998). A schema matches a string at every location. Thus, a 1 matches a 1 in the string, a 0 matches a 0, and a * matches either a 1 or a 0, behaving as a 'don't care' symbol. Although all evolutionary algorithms contain the aspects discussed above, there are different types of EA that are designed for specific tasks. As alluded to by the term Evolutionary Composite Systems, EFIT-V and EvoFIT use an evolutionary algorithm as part of the composite creation process. More specifically, these systems use a type of EA, referred to as a genetic algorithm (GA), which is explained in detail in Goldberg (1989).

Evolutionary Aspects of EvoFIT

EvoFIT uses Principal Components Analysis (PCA) to represent data about a face. PCA is a statistical technique used to emphasise variation and bring out strong patterns in a data set. This technique makes it easier to explore and visualise data. In 1987, Sirovich and Kirby demonstrated that faces could be presented well using PCA. Since then, PCA has been applied to many situations, such as in forensic settings to search for targets within a mugshot album (Baker, 1999).

The first step in the EvoFIT procedure is to produce a novel face. To do this, EvoFIT selects 17 floating-point random numbers (drawn from a normal distribution). The numbers are generated and scaled by the eigenvalue of the relevant eigenface. An eigenvalue is a value used to scale a vector, which is a phenomenon that has two properties: magnitude and direction. A vector whose direction remains unchanged

when a linear transformation is applied is called an eigenvector. An eigenface is a face image represented as linear combinations of eigenvectors (Turk & Pentland, 1991). This approach to face recognition seeks to capture the variation in a collection of face images and use this information to encode and compare images of individual faces holistically. The next step during composite construction is to select faces that resemble the target face.

During EvoFIT composite construction eyewitnesses select six face images that most resemble the target. These six faces are given a higher fitness value than all other faces. As EvoFIT implements a biased roulette wheel (Goldberg, 1989), so faces with a higher fitness value are more likely to contribute to the next generation than faces with a low fitness value. Once strings have been assigned a fitness value, uniform crossover and mutation take place (see, Holland, 2000), a process that occurs without input from the eyewitness. Crossover is used, hopefully, to increase the accuracy of face images that appear in the next generation by displaying faces which contain combined features from face previously selected faces (Gallard & Esquivel, 2001; Jones & Forrest, 1995). Mutation is used to create variation in the next generation by randomly changing the fitness of a characteristic, which may result in a face being displayed that would not be otherwise generated. Implementing too much or too little mutation is undesirable. After testing various levels of mutation to determine the optimum amount, it was established that using a rate of mutation of 0.05 produced the most accurate composite (Frowd et al., 2004). This mutation rate means that 1 in 20 coefficients are replaced with an appropriately scaled random value.

Mutation is not the only factor used to help produce an identifiable composite. Once six faces have been selected for shape and texture, Baker's (1985) algorithm is

employed. This algorithm selects the two fittest parents from the 12 options as a pair of “parent” faces. A common alternative to Baker’s algorithm is the weighted roulette wheel. This roulette wheel, as described in Holland (1975), is weighted so that the fittest selections are most likely to be chosen. EvoFIT uses Baker’s algorithm instead of the weighted roulette wheel because this algorithm is less likely to select low-rated faces (cf. weighted roulette wheel) inappropriately. Another way to ensure that the fittest face images are chosen is to increase their ‘weight’. The best faces selected for shape and texture are twice weighted, and all other selected faces are treated equally. This means that the “selection pressure” for this presumably preferable face is higher, resulting in more breeding opportunities and more offspring are produced with the influence from this item. This results in face images created using the fittest characteristics, meaning that these faces resemble the target to a greater extent. However, twice weighting is not the only method used to create a successful composite. Finally, at the end of each generation, an elitist strategy is used (Mitchell,1998). The elitist mechanism means that the best face is always carried forward to the following generation. This mechanism prevents any "superior" faces from becoming "lost" through crossover or mutation operators. The elitist strategy and the other mechanisms are used together to try to ensure that composites created depict the target as accurately as possible from memory.

One important factor that can impact the success of a GA is the population size (Goldberg, 1989; Lobo & Goldberg, 2004; Lobo & Lima, 2005; Mora-Melià et al., 2017). Typically, the more complex the problem is, the higher the population size needs to be (Lobo & Goldberg, 2004). In a facial composite system, the population size is the number of faces viewed before face images are 'bred' using a genetic algorithm. For EFIT-V the population size is the number of faces displayed on one

screen (i.e., nine face images). However, the population size for EvoFIT is the number of faces split between several arrays and is therefore much more extensive (i.e., 72 faces, or more specifically, EvoFIT displays 60 unique face images, and 12 images are carried over from the previous screen). The population size utilised during EvoFIT can be adjusted easily by changing the number of faces displayed on each screen, or by changing the number of screens shown to a witness.

Reducing the number of screens during EvoFIT composite construction was used in pilot work to alter the population size in Frowd and Grieve (2019), revealing that reducing the number of screens down to two was beneficial for overall composite accuracy. By decreasing the number of screens displaying faces for selection, the number of interactive elements (faces) is reduced, which is likely to decreases the intrinsic cognitive load of the task (Sweller, 2010). By decreasing cognitive load, the overall task of facial composite construction should be less demanding, which may result in a better performance from an eyewitness and a more accurate composite. Put simply, if an eyewitness has fewer faces to view, the task may be less likely to overwhelm the working memory; thus, their ability to select the best faces and make accurate alterations to the face may improve, resulting in a more recognisable facial composite. Alternatively, reducing the population size during EvoFIT construction may result in composite images that are less accurate due to the composite not evolving as accurately since the population size is smaller. Furthermore, reducing the population size also decreases the number of face images available to select from. If eyewitnesses have a smaller variety of face options, the likelihood of there being enough faces that accurately resemble the target is smaller. Testing these contrasting theories is a focus of the current thesis.

Another central factor for facial recognition is the importance of face shape and texture. As demonstrated in the literature, the role of shape and texture for face recognition is dependent on the circumstances, whereby face shape is relatively more important for the recognition of unfamiliar faces (Bruce et al., 1991; Knight & Johnston, 1997) and face texture is relatively more vital for recognition of familiar faces (Bruce et al., 1991; Lee & Perrett, 1997; Rogers et al., 2022; Russell & Sinha, 2007). Alternatively, as a facial composite is an imperfect image of a face, information about both the face shape and texture may be needed for accurate identification. There are two techniques for selecting face shape and texture in modern composite systems.

EFIT-V combines the face shape and texture into a single representation referred to as a global appearance model (Cootes et al., 2002). A limitation of this model is that it assumes face shape and texture are of equal importance for face recognition during police composite construction; however, there is no evidence to support this assumption. As mentioned above, EvoFIT separates selection of face shape and texture during the construction procedure. First, screens of "smooth" faces containing various face shapes without texture are displayed to the eyewitness, who is invited to select the best options from those displayed. Once the process of selecting the best face shape is complete, various textures are superimposed onto the face, and the selection process repeats for face textures. Once face shapes and textures are selected, combinations of them are displayed on two separate screens to allow selection of the overall best face for that generation (Frowd et al., 2004). Later, face shapes and textures are combined into various representations of the face from which the eyewitness can select the most accurate option. Additional important factors that impact the accuracy of face recognition and memory include the interview technique

prior to facial composite construction (see, Chapter 2, pages 62-63) and focusing on the eye region during construction (see, Chapter 2, pages 63-64).

Eyewitness Interview

Face recollection and memory impact the success of witness memory in the legal system, and therefore, aiding the memory of the eyewitness via interviewing techniques is important for the construction of an accurate facial composite image. During a cognitive interview (CI) the eyewitness freely recalls a description of the face to a practitioner (Gieselman et al., 1986). This interview technique is used by the police to obtain details about a crime, as explained Chapter 4.

A more modern interview technique, designed specifically for use before composite creation is the holistic-cognitive interview (H-CI: Frowd et al., 2008). The H-CI starts by asking participants to recall a description of the target face freely. The interview also asks the eyewitness to silently reflect the characteristics of the target by asking them to rate the target on seven characteristics which are read out sequentially (Frowd et al., 2011). An H-CI is beneficial before composite construction using a feature based or a holistic system (Frowd et al., 2015; Skelton et al., 2011). However, when using a holistic system such as EvoFIT, the interview process repeats, inviting the eyewitness to rate only the eye region on the same characteristics, as the eye region is important for face recognition (Portch et al., 2017).

The first experiment in this thesis utilised a self-administered cognitive interview (see, Fisher & Geiselman, 2010). Although this interview technique is rarely used in research (for an example, see Martin et al., 2018), it allowed the whole construction process to be completed in a self-administered way. The second

experiment utilised the more common cognitive interview as the construction procedure in this experiment was not self-administered. From Experiment 3 onwards, a holistic-cognitive interview was used to better resemble current police practises.

Current Thesis

This thesis theorises that when eyewitnesses compare many faces to each other and to their memory of the target face during EvoFIT composite construction, the number of interacting elements is high, raising the intrinsic cognitive load of the task (Paas et al., 2003). The high number of interacting elements overloads working memory capacity, impeding an eyewitness's decision-making ability and memory capacity (Sweller, 1988), two factors which are important when deciding which face images resemble one's memory of the target face. If reducing the number of face arrays during the 'face selection' stage of EvoFIT composite construction reduces 'elemental interactivity', that is, the number of elements in the task, this manipulation should also reduce the task's intrinsic load. If the intrinsic load of the task is reduced, eyewitnesses may be better able to remember the target face (the face they are attempting to recreate) and may make decisions about the face more accurately, improving the overall likeness of the final composite image.

The overarching aim of this thesis is to manipulate intrinsic cognitive load during EvoFIT composite construction by manipulating the number of faces (interactive elements) viewed during the construction process. In Frowd and Grieve (2019), composites constructed using *Two Screens* during the face selection procedure were more accurate than those constructed using the typical *Four Screens*. Henceforth, the current thesis aims to replicate and extend this experimentation,

reducing the number of screens incrementally from *Four* to *One Screen*. The results of the experiments will provide a unique understanding of the impact that cognitive load has on the ability of witnesses to create accurate composite images and may inform the development of a more effective procedure for producing facial composites.

Furthermore, there is currently no understanding of the relative importance of face *Shape* and *Texture* during the construction of an EvoFIT facial composite. Therefore, in the later experiments, the number of screens viewed during face selection will be reduced individually for the selection of face *Shape* and *Texture*. This procedure should provide valuable information about the importance of face *Shape* and *Texture* during EvoFIT composite construction and may establish an optimised procedure more effective for composite construction.

2

METHODOLOGY

The previous chapter outlined the literature on facial recognition, cognitive load and facial composite construction. Theories of memory and information processing were discussed in the context of facial composite construction, highlighting the impact that cognitive overload may have on eyewitnesses during composite construction. This chapter will outline the underlying research philosophy in this thesis as well as the methodology implemented to manipulate cognitive load during EvoFIT composite construction and will explain how composite accuracy is measured.

The research advances the use of the EvoFIT system to address key questions related to human memory and facial processing by decreasing the population size during composite construction (i.e., reducing the number of faces viewed and compared by participants at each stage of the composite construction process). This chapter will detail the research design selected for the five experiments and justify the key design choices, namely the research philosophy, research type, strategy, the time

horizon, sampling, data collection and analysis. This chapter will also discuss methodological limitations.

Research Philosophy

The underlying research philosophy in this thesis is Positivism due to its reliance of empirical evidence to study human behaviour. Positivism is a philosophical belief that one should not go beyond the boundaries of what can be observed using the scientific method. As scientific knowledge is testable, research can only be proven by empirical means; thus, arguments, belief and intuition are inadequate (Hacking, 1981).

Although positivism is not considered to be an adequate philosophy to study the full range of human experience, it has been hugely influential as it is still used to study many a phenomenon which can be measured empirically. Three key features of Positivism are that it is useful for testing theories, it aims to predict behaviour, and it looks for hard rules or laws (Bryant, 1985). The objective and logical nature of empirical research encourages an unbiased and systematic measure of human ability. This thesis aims to empirically measure the ability of participants to create an accurate facial composite using *Four* different population sizes: *One, Two*, *Three* or *Four Screens* of face images to select the face *Shape* and *Texture*. In other words, this thesis aims to test the theory of cognitive load in relation to composite construction, predicting that participants will create the most accurate facial composites images using fewer screens of faces.

Research Type

The research type selected for this thesis was indictive, as the theory is generated from the collected data and is therefore exploratory in nature. Inductive research was chosen as this area of research is fairly new, with only one previously existing experiment (Frowd & Grieve 2019). Therefore, an exploratory, rather than a confirmatory approach is most appropriate.

Furthermore, the research adopts a quantitative methodology. Quantitative experiments are most suitable to understanding the effect of population size during EvoFIT composite construction as they encourage the production of generalisable information.

Research Strategy

Research in this thesis is conducted using experiments, as they enable a comparison between a control (composites constructed using the standard approach in EvoFIT with *Four Screens*) and experimental factors (composites constructed using a reduced number of screens). This method of research is controlled and, in ideal circumstances, would be undertaken in a laboratory to ensure control over the artificial environment.

The designed experiments enabled careful manipulation of controlled variables, allowing the research to be replicable (Schiewe, 1988). Undertaking quantitative experiments like those in the current research are the standard methodology in EvoFIT research. There are parallels in the lab and the real world, so it is important to know how effective composites are in the lab to improve composite construction for police investigations. Using untested composite construction methods

in actual cases, as a field experiment, would be unethical; therefore, a lab-based experiment is deemed most appropriate.

Time Horizon

This thesis relied on cross-sectional data, whereby the data in each experiment was collected at one point in time, rather than at multiple time points. Cross-sectional data was used because the research aims did not rely on the collection of data at different points in time. Furthermore, participants in the current experiments were recruited to be either familiar or unfamiliar with the target identities (celebrities) viewed in the experiments, depending on the requirements in each part of the experiment. If data were collected at multiple time points, a participant's familiarity with the identity may change between the time points; for example, if a participant started watching a soap opera featuring the targets from the experiment.

Sampling Strategy

The research in this thesis employed probability sampling, which involves a random selection of participants from a population. This sampling method aimed to develop findings that are generalisable to the general population. However, it should be noted that it is difficult to generate a sample that reflects the whole population, as participants were all somewhat interested in research and signed up for a participant recruitment website. Therefore, the random sample should be generalisable to this population, but may not truly reflect the general population.

The number of participants for each experiment in this thesis was determined by looking back to historical data in EvoFIT research. Ten participants per condition

for both face construction and naming have been used in similar research since these number of participants lead to a design with sufficient power to be able to detect at least a medium effect size, should one exist (Fodarella et al., 2015; Frowd et al., 2008; Giannou et al., 2021).

Brown et al. (unpublished) estimated the number of participants required to construct and correctly name composites using G*Power (Faul et al., 2009). The aim was to detect at least a medium effect size for by-items and by-participants analyses (*Odds ratio* = 2.5) based on mean correct naming of 25%, which should be achievable in this experiment based on previous research (see, Erikson et al., 2022; Fodarella et al., 2021; Giannou et al., 2021). The result was 105.5 responses per group. Each experiment is designed to include 10 participants for construction and at least 10 participants for naming in each condition, resulting in a minimum of 100 responses per group.

Each experiment involved three separate groups of participants. The first group of participants constructing the facial composites must not have created a facial composite using EvoFIT in the past six months to avoid practice-effects and to ensure that the manipulation remains unknown to the participant during construction. Moreover, participants creating the facial composite must be unfamiliar with the target, replicating the most usual scenario where an eyewitness creates a composite of an unknown perpetrator.

The second group of participants, tasked with naming the composites must, instead, have been familiar with the targets to recognise them based on the composite images. An *a-priori* rule was employed for participants in the second group, whereby participants must name eight of the 10 targets correctly from their photographs to be deemed familiar. In early EvoFIT experiments (see, Frowd, Carson, Ness, Richardson

et al., 2005), the *a-priori* rule was for participants to name 50% of the targets correctly, so that they had the opportunity to name at least 50% of the composites correctly. However, only knowing 50% of the targets may lead to less effective data, since the estimate from each person will be more variable (as a score out of 5 will be less than a score out of 10). Therefore, the *a-priori* rule was raised to 80% correct target naming. Although it would be preferable for participants to name 100% of the targets correctly, this would make it more difficult to recruit participants. Therefore, 80% correct naming is the requirement in this research, as is common in published facial composite research (Fodarella et al., 2021)

The third group of participants who were asked to rate the composite likeness compared to the target photograph must also be unfamiliar with the targets to prevent under or over-rating based on the participant's recognition of the composite. Participants may under-rate a composite they did not personally recognise a target but over-rate a target they did recognise. An *a-priori* rule was employed for participants in this group: Participants rating the likeness of facial composite images should, in general, not be familiar with the target identities. Therefore, participants in these groups were invited to inform the researcher if they recognised any of the individuals based on their photograph during the experiment. Data from participants who were able to name two or more of the identities from the photographs were not included.

Data Collection

Each experiment included Part 1 (Composite Construction) and Part 2 (Composite Evaluation). Composite Evaluation comprised Part 2a: Composite Naming, Part 2b: Final Composite Image Rating, and Part 2c: Intermediate Composite Rating.

Part 1 (Composite Construction)

For each of the five experiments, participants were recruited using opportunity sampling via the undergraduate participation system SONA, and websites for participant recruitment: Call for Participants and Prolific Academic. These online systems are designed with the purpose of recruiting participants for various experiments; SONA is the system used for participant recruitment at UCLan, and Call for Participants and Prolific Academic have no affiliation with the university. On all three websites, the researcher posted an advertisement for their experiment as well as the participant criteria and any reward for completing the experiment. Participants signed up to arrange a time and date for the experiment to take place via videoconferencing.

Participant rewards for taking part in the experiment were course credit (for participants from SONA), £5 online shopping vouchers (for participants from Call for Participants) or £5 cash (for participants from Prolific Academic). A £5 payment was appropriate as the experiment took approximately one hour. Underpayment may cause low participation rates, and an overpayment may be unethical as individuals uncomfortable with the experiment may simply participate for the high reward (Bentley & Thacker, 2004). As all of the experiments took place remotely, participants were required to access a PC or laptop and video conferencing software, such as Microsoft Teams or Skype.

Target Encoding

Part 1 took place over two days. On the first day, participants viewed a photograph of an unfamiliar face (referred to as the 'target') for 30 seconds. Thirty-seconds is

currently a standard length of time used in EvoFIT research (e.g., Frowd et al., 2007; Giannou et al., 2021), a reasonable length of time to allow participants to view the whole face in detail. Although some research allows participants to view the target for 60 seconds (Fodarella, 2020), this small difference in time is not considered to have a significant impact in relation to an eyewitness (Polluzo et al., 2019).

Each experiment involved 10 target faces. Targets used in all five experiments were celebrities, so that they would be familiar to the group of participants who were recruited to name them, but unfamiliar to participants recruited to create composite images and rate the images for likeness to the target. In Experiments 1 and 2, target faces were England International Footballers. In Experiments 3-5, Soap Actors from Coronation Street, EastEnders and Emmerdale were chosen (see Appendices 1-6 for target photographs).

In all five experiments, 40 participants were recruited to create facial composites. As there were 10 targets in each experiment, *Four* composites were created of each target. In Experiments 1-3, 10 composites were created in each level of the condition Screens (*One Screen, Two Screens, Three Screens, Four Screens*). In Experiment 4, 10 participants were created in both levels of the condition *Shape* (*Two Screens, Four Screens*) and *Texture* (*Two Screens, Four Screens*). Similarly in Experiment 5, 10 participants were created in both levels of the condition *Shape* (*One Screen, Two Screens*) and *Texture* (*One Screen, Two Screens*).

As participants viewed the target photograph, they were asked whether they recognised the individual in the image. If they did report the face to be known, they were given a different target to inspect, and this process was repeated. It was unlikely that any participant would have recognised all of the targets, as the advertisement stipulated that only participants who were unfamiliar with the target group (England

International Footballers or actors/actresses from the relevant television soap depending on the experiment) should take part. In this thesis, no participant recruited to create a composite image was able to recognise all of the targets. If they had, they would have been informed that they do not meet the requirements for the experiment and that they were unable to take part.

Furthermore, the researcher did not know the target identities until after data collection, so she would not have been able to subconsciously bias faces selected by participants. To avoid the researcher seeing the target identities, when participants viewed the target face during encoding, each participant was sent a password-protected Word document and the corresponding password to open the document and view the photograph.

Construction Procedure

Twenty-*Four* hours after participants viewed a target face, they were sent a link to EvoFIT Online, where on-screen instructions directed the composite construction (Experiment 1) or, were interviewed by the researcher using a Cognitive Interview (CI: Experiment 2) or an H-CI in (Experiments 3-5).

In a real-world scenario, it is unlikely that an eyewitness will have the opportunity to create a facial composite image sooner than 24 hours after a crime has taken place, although this does happen occasionally. Therefore, including a 24-hour delay between encoding and composite construction in this thesis replicates a scenario where an eyewitness creates a facial composite with the police (Martin et al., 2018). Although there are cases where eyewitnesses are unable to create a facial composite of a perpetrator after only 24 hours, and must wait longer, memory decline for recall of faces is greatest up until 24 hours and declines less rapidly after this length of time

(Kramer, 2021). Therefore, it was not deemed necessary to include a longer delay between encoding and composite construction in this thesis, even though a longer delay may occur during a real investigation and would be more ecologically valid.

The Interview: The first step in the construction of a facial composite is the interview. The typical interview used is the CI (Experiment 2). In Experiment 1, no interview was conducted by the researcher. Instead, composite construction occurred via EvoFIT Online, using a self-assessment style (Martin et al., 2018). In Experiments 2 onwards, composite construction occurred using the EvoFIT App and so a cognitive interview was conducted by the researcher prior to composite construction. During this interview, the researcher asked each participant to freely recall a description of the target face seen 24 hours previous. A Verbal Recall Sheet (see, Appendix 6) was used to record the face description: a procedure used by the police and replicated in research (Marsh et al., 2015; Pitchford et al., 2017).

In Experiments 3-5, the researcher conducted a more modern, extensive interview, referred to as the Holistic-Cognitive Interview (H-CI). An H-CI starts in the same way as a CI, with a free recall of the face. When eyewitnesses recall a description of a face, they typically describe each feature individually, which may encourage them to perceive the face as a collection of individual features as opposed to a holistic image. Consequently, after recalling a description of the face, it is important to invite the eyewitness to re-focus the attention on the whole face (Frowd et al., 2012). The interviewer does this by asking the eyewitness to think about the personality characteristics associated with the target face, based on their face, for one minute. Once this minute has passed, the researcher explains that the interviewer will read aloud a list of seven characteristics (intelligence, friendliness, kindness, selfishness, arrogance, aggressiveness, and distinctiveness) and, for each

characteristic, the eyewitness should rate whether the target would be is low, medium or high for their given face (Frowd et al., 2008).

A recent addition to the H-CI considers the importance of the eye region for face recognition, and the emphasis that is placed on the eye region during the selection of faces to create an EvoFIT facial composite (Fodarella et al., 2017). During face recognition, the human gaze predominantly centres on the eye region (Barton et al., 2006). Faces displayed with the eye region covered were identified less frequently than those with any other feature covered (Peterson et al., 2008; Royer et al., 2018), demonstrating the importance of this region for face recognition. The eye region is considered very important for unfamiliar face-matching or recognition tasks and is crucial for interpreting facial expressions for familiar faces (Calvo et al., 2018; Nelson & Mondloch, 2014). The eye region is also an important factor in non-verbal communication from birth throughout adulthood (Jongerius et al., 2020; Kleinke, 1986).

With so much focus on the eye region in day-to-day life (Hjelmas & Wroldsen, 1999; Royer et al., 2018), the importance of this region during facial composite construction is understandable. Therefore, after eyewitnesses have rated the target face on each of the seven characteristics during the H-CI, the researcher explains that they will repeat the same list of seven characteristics, but that the eyewitness should now rate whether the target was low, medium or high for each characteristic based only on the eye region (Frowd et al., 2019). Asking eyewitnesses to rate the characteristics of the target based only on the eye region encourages them to focus on this region, potentially making it easier for the eyewitnesses to also focus on the eye region when selecting faces during construction. Interestingly, when this

step was forgone, but face images were selected based on the eye region (as is standard procedure), composite likeness was heavily reduced (Portch et al., 2017).

Face Selection: For all experiments, composite construction took place online. Experiment 1 was a self-administered interview whereby the participant moved through the on-screen instructions by themselves. For the rest of the experiments, the app was used by the researcher who shared her screen with the eyewitness, meaning that the researcher and the eyewitness viewed the screen in real-time, but the researcher had control of the mouse. After selecting the appropriate age and race database for the target, participants start to select face images which resemble the target.

Throughout the face selection stage of EvoFIT composite construction, the researcher instructed participants to ignore the width of the faces, as this can be changed later, and to focus on the eye region when selecting face images which resemble the target, as per the standardised procedure. To start, *Four* arrays containing face shapes are typically displayed to the eyewitness, with each face array containing 18 faces, displayed as three rows of six faces. An example face array for selection of the face shape is presented in Figure 1.

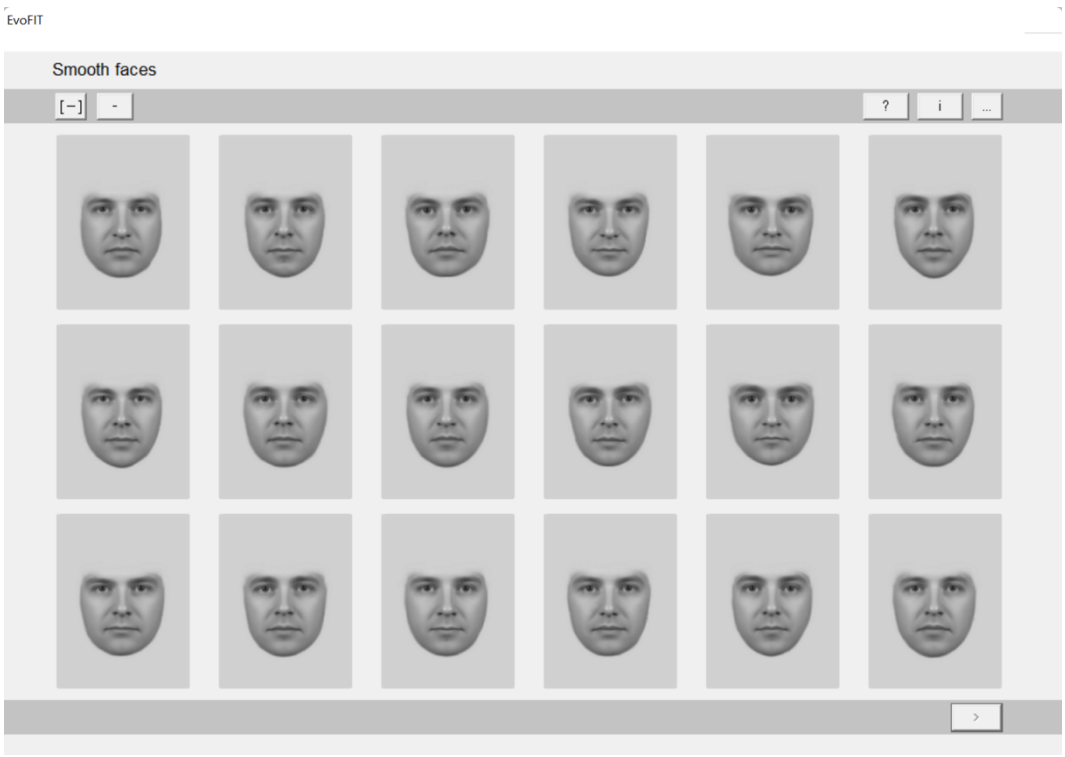


Figure 1. An example face array containing 18 smooth face images for selection by the eyewitness.

As the current research aims to manipulate the number of face arrays, or screens, viewed by a participant, *One, Two, Three* or *Four* face arrays were displayed, depending on the experimental condition. For the current method of EvoFIT construction, eyewitnesses selected two faces which resemble the target from the first three face arrays, and then had an opportunity to swap any of the selected faces in a fourth face array.

In this research, participants selected six faces altogether, despite the number of face arrays viewed: Participants in the *One Screen* condition selected six faces from *One* screen; participants in the *Two Screens* condition selected three faces on each of two screens; participants in the *Three Screens* condition selected two faces from each of two screens; and participants in the *Four Screens* condition selected two faces from the first *Three Screens and* were able to swap faces on the fourth screen. Once six faces were selected, the researcher invited participants to view the faces and

select which face, from the six, resembled the best resembled the target based on the eye region. This face was used as the base and various facial textures were applied to it. An example face array for selection of the face shape is presented in Figure 2.

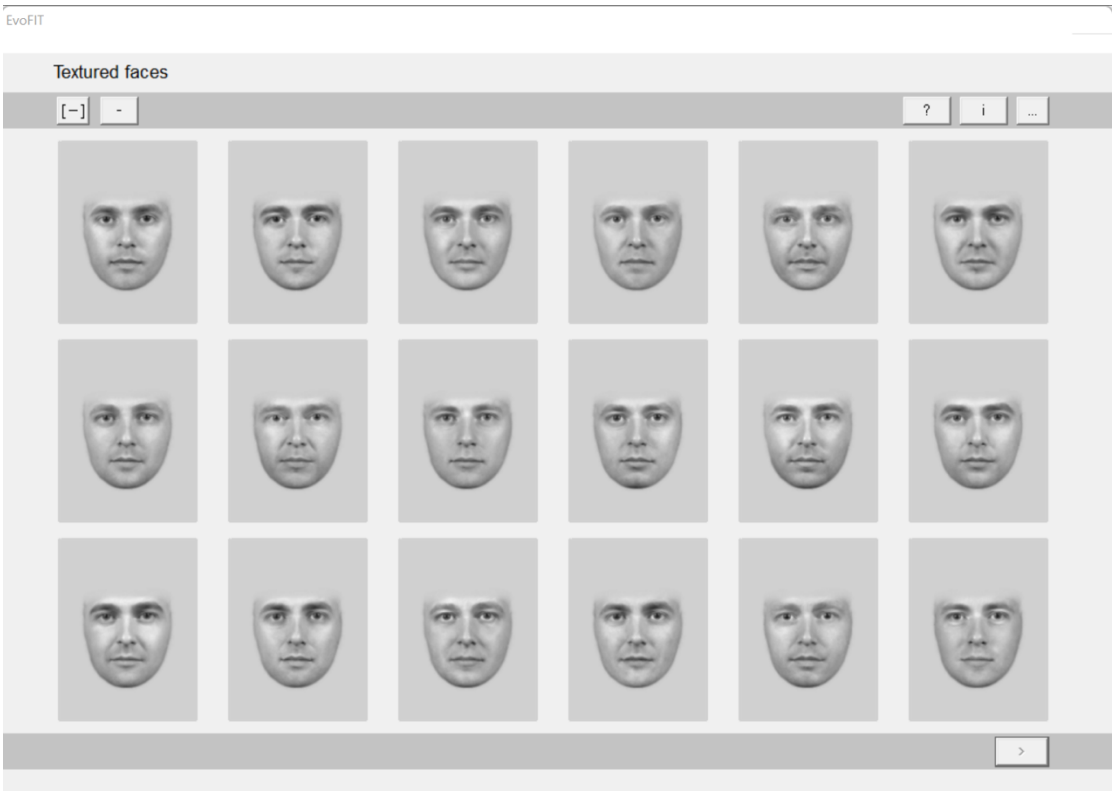


Figure 2. An example face array containing 18 textured face images for selection by the eyewitness.

So, the researcher invited participants to select six face images from *One, Two*, *Three* or *Four* face arrays, as per the experimental condition. Next, instead of selecting the best face from these six, two new face arrays were displayed to the participant, containing a variety of face shape and texture combinations that may resemble the target face, based on the information learned by the genetic algorithm (as demonstrated in Figure 3).

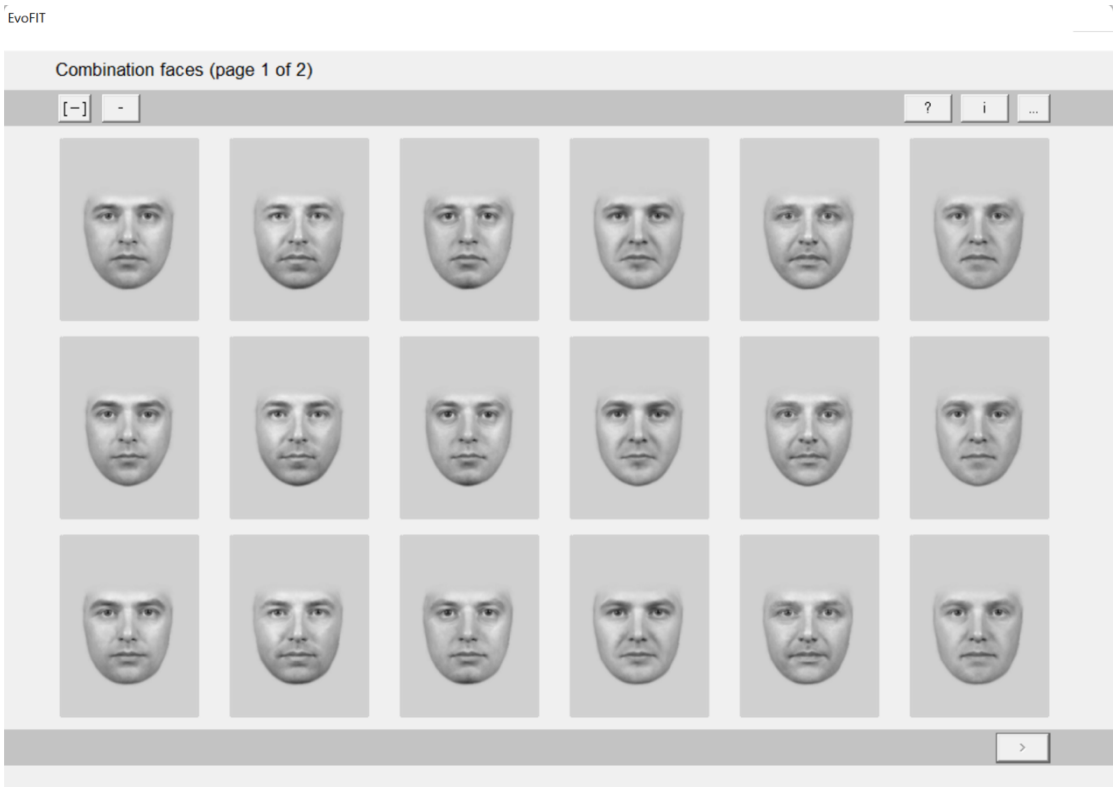


Figure 2. An example face array containing combinations of faces previously selected by the eyewitness.

All participants viewed both screens of face arrays and selected one face from each screen. Once two faces have been selected, based on the eye region, participants choose between the two faces. A further screen was then displayed, containing an overview of the previously selected faces. Participants were invited to select the best face from those displayed, ideally resulting in the selection of the most accurate face from the *First Generation*. In the current research, the researcher saved this face image (as '*First Generation*') for rating in Part 2c of the experiment. The process repeated exactly, and the genetic algorithm produced face images based on those selected as "best fit" during the *First Generation*. After selecting the best likenesses for face shape, texture and combinations of shape and texture, the most accurate face from the *Second Generation* was saved (as '2nd Generation) for rating in Part 2c of the experiment. Typically, participants are then given the option to evolve the

composite image again, which would repeat the *Second Generation*. However, as the number of screens viewed during the construction procedure was controlled in this thesis, participants were not given this option.

Image Enhancement: Once a face image has been selected, the likeness of the image to the target face can be enhanced using the *Holistic Tools* and the *Shape Tool*, and then external features (i.e., hair, neck and ears) can be selected. The aim of image enhancement is to increase the accuracy of the facial composite, and so participants are invited to focus on making changes to the whole face as opposed to focusing on the eye region, which was a requirement during face selection.

The first image enhancement tool used is *Holistic Tools*. Using *Holistic Tools*, participants are able to make 15 different changes to their evolved face. To do this, there is a sliding scale to manipulate how extreme the change should be and in which direction. For example, when changing the age, participants can make the face look older by sliding the tool to the right, and younger by sliding the tool to the left. By sliding the tool further from the central point (i.e., all the way to the end of the sliding bar), the change is more extreme, but by keeping the tool close to the central point, the change is more subtle (demonstrated in Figure 4). Participants go through each of the 15 changes with the researcher and look at the face at each point on the sliding bar (there are 11 points on the bar in total, including the central point which does not change the face). Once participants have selected the point at which the face most resembles the target for all 15 changes, they then use this same tool to change the shading of the face image.

Figure 3. Composite age change displayed at three different points on the scale during Holistic Tools. The composite is manipulated to appear younger (left), older (right), and with no change (middle).

Although a facial composite is displayed in greyscale, changes to the shade of facial features can increase or decrease the images identifiability. Participants do not have to go through all available changes to alter the shade of the face (as they do with the 15 changes to the face discussed above) but are able to select the areas of the face that they would like to lighten or darken, using the sliding bar. Using this tool, participants a lighten or darken the eyebrows, iris, cheekbones and eyebags. Participants are also able to create the illusion of stubble by darkening the area around the jaw and above the top lip. Once a participant is satisfied that the best likeness has been created using the tools, they view an image of the composite before the changes and after the changes and are invited to select the face image that most resembles the target. This step allows participants to undo all of the changes made using *Holistic Tools* if they have reduced, rather than enhanced, the likeness of the composite image. The image selected here is saved (as '*Holistic Tools*') for rating in Part 2c of the experiment.

Once alterations are made to the face using *Holistic Tools*, participants have a choice of using the *Shape Tool* to make further changes to the face or of selecting the external features (hair, neck and ears) for the composite image. The biggest difference between *Holistic Tools* and the *Shape Tool* is that, whereas *Holistic Tools* makes changes which affect the whole face, the *Shape Tool* makes changes that affect a pair (or group) of features, individual features, or even individual points on the face. Each feature on an EvoFIT composite can be changed, as demonstrated in Figure 5.

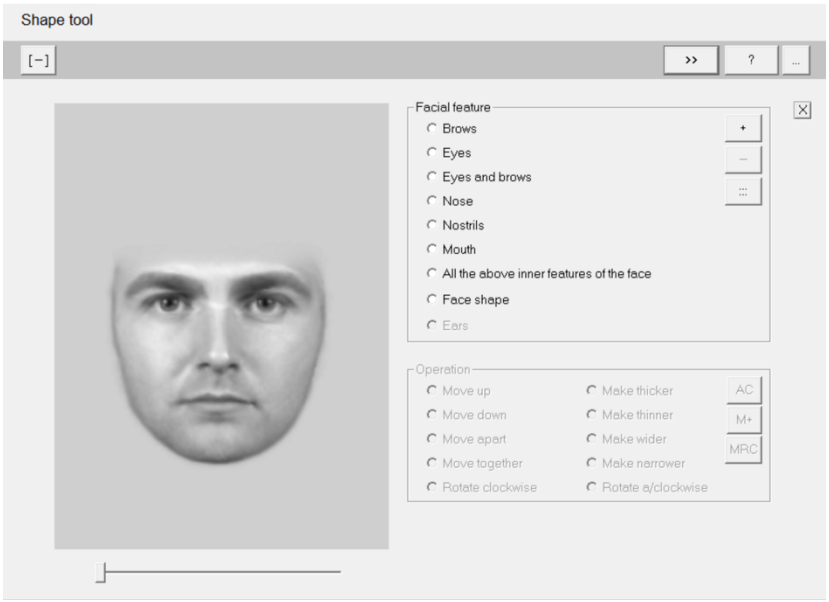


Figure 4. A demonstration of the shape tool during EvoFIT composite construction.

Each feature on an EvoFIT composite is outlined by many points, which are visible when entering the *Shape Tool*. Each point on the face can be moved up, down, left or right. When a collection of points (two or more) are selected, the area within the points can be made larger (by moving the points further apart) or closer together (by moving the points closer together). When a participant selects a feature, for example the nose, all of the points that are related to the nose are automatically selected, without the researcher having to select individual points. However, if a very specific part of the

nose needs to be moved or resized to increase the likeness of the composite, for example, the inner corner of the right nostril, the points related to this area can be selected by the researcher. Typical changes made using the *Shape Tool* include making a feature appear larger or smaller, changing the angle of a feature or a part of a feature (such as turning up the corners of the mouth), and moving a feature. An example of a facial composite before and after thinning the eyebrows is displayed in Figure 6.

Figure 5. A demonstration of a change made using the Shape Tool, with the original face (left) and the face with the eyebrows thinned (right).

These changes are often quite subtle, and make a composite image appear less generic, increasing the uniqueness and ideally the identifiability of the composite. Once a participant believes that they have achieved the best likeness using the *Shape Tool*, they view an image of the composite before the changes and after the changes side-by-side and are invited to select the face image that most resembles the target.

This step, again, allows participants to undo changes made using the *Shape Tool* if they have reduced, and not enhanced, the likeness of the composite.

Selecting the external features for a composite uses a similar process to face selection. The researcher filters the external feature options, allowing the researcher to select the length, colour and style of the target's hair based off information in the initial interview so that the most relevant hairstyles are displayed first. Then, the facial composite image that has been created by the participant is displayed with various external features in an array of 18 faces (three rows of six face images), as demonstrated in Figure 7.

Figure 6. An example face array displaying different external features for selection by the eyewitness.

Participants are made aware that small changes to the external features can be made here, such as the colour (shade) of the hair being lightened or darkened and the parting of the hair being reflected. Unlike face selection, participants are able to select as many hair options as they like and are able to view options multiple times. Once participants have viewed all of the screens, and/or have selected as many of the hair options as they deem appropriate, a face array containing only the external feature

options that have been selected is displayed. From here, participants are able to select the most accurate external features, or a small group of external features, which resemble the target and repeat this process until only one option is selected. Once external features have been chosen, participants are able to make further changes to the face using *Holistic Tools* or the *Shape Tool* or, if the best likeness has been achieved, the composite image is saved at this point. An image of the composite internal features at this point is also saved (as '*Final Image*') for rating in Part 2c of the experiment.

Although this general process of composite construction remains the same throughout Experiments 2-5, the interview conducted, and the number of screens viewed during face selection is dependent on the experiment. In Experiment 2, a CI was conducted prior to composite construction, and in Experiment 3-5 an H-CI was conducted. Furthermore, in Experiments 2 and 3, the number of screens viewed during face selection remained the same for selection of face shape and face texture (*One, Two*, *Three* or *Four Screens* depending on the experimental condition). However, in Experiments 4 and 5, half of the participants viewed the same number of screens for selection of the face *Shape* and *Texture* (*Two* or *Four Screens* in Experiment 4 and *One* or *Two Screens* in Experiment 5). However, the other half of the participants viewed a different number of screens for selection of the face *Shape* and selection of the face *Texture* (*Two Screens* to select the face *Shape* and *Four Screens* to select the face *Texture*, or vice versa in Experiment 4 and *One Screen* to select the face *Shape* and *Two Screens* to select the face *Texture*, or vice versa in Experiment 5). The motivation for these manipulations will be explained in the individual experiments themselves.

Composites constructed throughout this PhD research were done so using remote composite construction, via video-conferencing software. This method enabled participants to view the researcher's screen but required the eyewitness to explain the position in the array of faces for selection (compared to merely pointing to the face). Although construction was not face-to-face, the researcher took all measures to replicate the process of police composite construction as far as possible. For example, and similar to a typical face-to-face procedure, the researcher made an effort to ensure that rapport was developed by conversing with the interviewee prior to the interview and taking time to describe the procedure throughout composite construction in a polite and friendly manner (Danziger, 2023).

Part 2 (Composite Evaluation)

Despite differences between experiments during composite construction, the procedure for Composite Naming (2a) and Intermediate Composite Rating (2c) remained the same for all experiments. The procedure for Final Composite Image Rating (2b) remained the same between Experiments 1-3 and Experiments 4 and 5.

As in Part 1, participants were recruited using the undergraduate participation system SONA, and participant recruitment websites Call for Participants and Prolific Academic. Participant rewards were course credit (for participants recruited via SONA), £2 shopping vouchers (for participants recruited via Call for Participants) or £2 cash (for participants recruited via Prolific Academic). Although not time limited, the experiment typically lasted between ten and twenty minutes. The experiments took place remotely, via Microsoft Teams or Skype; participants required access to a PC or laptop and video-conferencing software.

Part 2a (Composite Naming): Part 2a of the experiments implemented a between-subjects design. Participants were recruited on the basis that they were familiar with the target faces. These participants were randomly allocated to condition, for example, in Experiment 1, participants were assigned to one of four conditions (*One Screen, Two Screens, Three Screens, Four Screens*). The researcher showed each participant a PowerPoint presentation containing 10 facial composite images from one of the four conditions (*One Screen, Two Screens, Three Screens* or *Four Screens*) sequentially and asked the participant to name the individual based on the composite image or provide a "don't know" response if they did not recognise the target from the composite. Once participants had attempted to name all 10 targets, the researcher sequentially displayed the ten target photographs and asked the participant to name these. Participants who were able to correctly name eight of the 10 targets correctly were deemed sufficiently familiar with the set. However, participants who were unable to name at least eight of the targets correctly were deemed to be unfamiliar with the targets, and their data were not included in the study and a new participant was recruited to replace them in the same condition. After naming the targets, the researcher displayed the same PowerPoint presentation containing the 10 composite images and asked participants to re-attempt naming with the knowledge of at least 80% of the targets. This type of identification (cued naming) is not optimal; however, it can provide rich data when spontaneous naming is lower than expected. These data were presented and discussed in Experiment 1.

Part 2b (Composite Rating): In all experiments, Part 2b implemented a mixed design and recruited participants who were unfamiliar with the targets. In Experiments 1-3,

participants were randomly assigned to one of three rating tasks: rating *Internal Features, External Features* or *Whole Composites*.

Depending on the condition, facial composites were displayed with only the internal features visible, only the external features visible or as a whole, unedited composite. The researcher displayed each composite image (or partial composite image) next to the corresponding target photograph and asked participants to rate how alike the images were on a Likert scale of 1 (very poor likeness) to 7 (very good likeness). This measure may not be particularly useful for determining the optimum number of screens during composite construction, as other stages follow it, but it is necessary to assess whether reducing the number of screens during composite construction has an adverse effect on the accuracy of different regions of the composite. For example, constructing a composite by selecting faces from only *One Screen* may result in a composite that has accurate external features, but may give rise to unrecognisable internal features. Hence, this part of the experiment is important to assess any negative effects that may be caused by manipulating the number of screens during the construction procedure. Additionally, in Experiment 1, participants from each condition were randomly allocated to two groups: 'whole face' and 'eye region'. The 'whole face' group followed the same procedure as Experiments 2 and 3. The 'eye region' group were asked to focus on the eye region when rating each composite image.

In Experiments 4 and 5, participants were randomly assigned to one of three rating tasks: rating the face *Shape*, face *Texture*, or *Whole Composites* (a combination of face shape and texture). The researcher displayed each composite image next to the corresponding target photograph. In the shape rating condition, the researcher defined face shape to the participant as "the shape of the head and the shape and configuration

of features" and asked the participant to rate how alike the face shape was between the two images. In the texture rating condition, the researcher defined face texture as "colour-based properties of the face" and, as the composite image was presented in greyscale, made it clear that this referred to the shading information in the image. The researcher then asked the participant to rate how alike the face texture was between the two images. The researcher did not define face shape or texture in the whole face rating condition and asked the participant to rate how alike the face was between the two images. All three conditions used the same Likert scale of 1-7 as the other experiments. The purpose of rating in this way was to determine whether creating a facial composite using fewer or more screens to select the face shape than the face texture, or vice versa, impacted the accuracy of the face shape or texture in comparison to the target photograph. For example, creating a composite using *Two Screens* to select the face *Shape* but *Four Screens* to select the face *Texture* may result in the face *Shape* or the face *Texture* being far more accurate than the other.

Part 2c (Intermediate Composite Rating): In all experiments, Part 2c implemented a within-subject's design and recruited participants who were unfamiliar with the targets. As per the within-subject's design, all participants viewed composites from all conditions. The researcher displayed composite images at four stages of construction: *after the first generation, after the second generation,* after use of *Holistic Tools* and *the final image*. Furthermore, 40 random composite images were selected (one for each composite created by a participant) and one random face was displayed alongside the composite images from each stage of the construction procedure (totalling 5 face images). As only internal features are available for the *First*

Generation, *Second Generation* and after *Holistic Tools*, only the internal features were included in the *Final Image* and the random composite image.

The researcher presented the images in a random order and also displayed the target photograph for comparison, as in Part 2b. The random composite image would act as a baseline measurement and would also demonstrate at which point in the composite construction process the likeness of a composite image surpasses that of a *Random Face* when compared to the target photograph. Participants rated the likeness of each image compared to the target photograph on a Likert scale of 1-7. The purpose of rating composite images in this way was to determine the impact of reducing the population size at the beginning of composite construction (during face selection) on the participant's ability to utilise the image enhancement tools (*Holistic Tools* and the *Shape Tool*) at the end of composite construction.

Data Analysis

Preparation

The first step in the data analysis was preparing the data. In Part 2a, the data was checked to ensure that all participants were able to name 80% of the target photographs correctly. The data were then coded. Responses were coded as correct and assigned a value of 1 when participants gave the correct name for the target image and composite image. Responses were coded as incorrect and assigned a value of 0 when a wrong name or "don't know" response was given for the composite image, but the target was identified correctly. Responses were assigned a value of 2 when the target was not identified. In SPSS, cases less than two are selected so that cases are included in the analysis when a participant correctly named the identity based on the

target photograph. In Parts 2b and 2c, the dataset was checked for missing or low-effort data (whereby a participant gives each composite image the same rating, for example, rating all composites as a 3). Missing or low-effect data did not occur in any of the experiments in this thesis; however, if it did, the data would not have been included, and a new participant would have been recruited in the same condition.

Analysis of Means: In SPSS, Summarise Cases was used to analyse the means, and the number of cases for composite in Part 2a; as well as the means and standard error for composite rating in Parts 2b and 2c.

Generalised Linear Mixed Models: The regression technique Generalised Linear Mixed Models (GLMM) was used in SPSS throughout the research. GLMM is an appropriate analysis technique to use in this research because it takes the participants and the composite items into account in the model, as opposed to ANOVA, which is able to take either participants *or* items into account, but not both in the same model. Furthermore, GLMM is now considered to be the industry standard, so it is a clear choice for use in this thesis.

In each experiment, a single model was run for each predictor. Where there was only one predictor (Experiments 1-3), this single model constituted the GLMM analysis. However, where there were two predictors (Experiments 1-5), a single model was run for each predictor, and an interaction was also run between the two predictors. If the single predictors were significant, but the interaction term was not, a further model was run containing the two single predictors without the interaction. Furthermore, for experiments with three predictors (Experiments 4 and 5), a single model was run for each predictor (a, b and c) and a separate model containing the interaction between each predictor was run (a*b, a*c and b*c) as well as a three-way interaction between all three predictors (a*b*c). As per standard practice, models

containing an interaction also contained the two (or three) relevant individual predictors in a full factorial model (Field, 2018). Also, a higher order interaction was selected as a preferable model over a lower order one.

Generalised Linear Mixed Models

In Part 2a, GLMM modelled the between-subjects IVs *(predictors or fixed effects)* in the context of *random effects*, which were (i) participants in the naming stage of the experiment and (ii) composite items (identities). Data in this part of the experiment were binomial, with coded values as 0's and 1's. As Part 2a was a between-subjects design, only random intercepts (and not slopes) were included in the model for participants and items. In Part 2b, GLMM modelled IVs *(predictors or fixed effects)* in the context of *random effects*, which were (i) participants in the rating stage of the experiment and (ii) the composites created in Part 2a. Part 2b was a mixed design. *Task* was a between-subject factor and was run in the context of the random effects; however, *Screens* was a within-subjects factor. Therefore, *Screens* was also added as a *random slope* for (i) participants and (ii) items (see Erikson et al., 2022). In Part 2c, GLMM also modelled IVs in the context of *random slopes,* which were (i) participants and (ii) items, as in Part 2b. As this part of the experiments utilised a within-subjects design, with participants rating all composites, random slopes were added for both *Screens* and *Stages* to both participants and items. However, it is not always the case that random effects can be estimated from the data; while models were built to initially include all random effects, any effect that could not be estimated was removed, to give a resulting model that was as generalisable as possible (Barr et al., 2013).

Post-Hoc Testing

Polynomial contrasts are simulated using GLMM to determine the mathematical pattern of the data, namely a linear, cubic, quadratic or quartic trend. In Parts 2a and 2b of Experiments 1-3, polynomial contrasts were run for *Screens*, to understand the pattern of composite accuracy between the four levels of *Screens*. A linear contrast was predicted, which would indicate the accuracy of composites changes in the same direction, by a similar rate between each level of screens (i.e., composite accuracy increases or decreases as the number of screens increases or decreases).

In Part 2c of all experiments, polynomial contrasts were conducted when a predictor emerged significant in the GLMM. In all five Experiments, polynomial contrasts assessed the mathematical pattern of the data between the levels of *Stages.* A monotonic trend in the data was expected, demonstrated by significant linear and / or quadratic contrasts. A monotonic pattern would indicate the change in accuracy between each stage of construction would move in the same direction but at a different rate between the levels of construction. Hence, it was predicted that composite accuracy would increase throughout construction, but the change between each level would be smaller at the end of construction compared to the beginning of construction- an effect that would appear to fit a quadratic trend the best. In Experiments 2 and 5, polynomial contrasts also assessed the mathematical pattern of the data between the levels of *Screens.* As in parts 2a and 2b, a significant linear contrast was predicted, whereby the accuracy of composites changes in the same direction, by a similar rate, between each level of screens.

Evaluation of Methodology

All experiments in the current thesis were designed to mimic a typical investigative process as far as possible. However, lab-based experiments may not simulate real-life situations and behaviours, particularly where research involves witnesses and victims of crime. The literature demonstrates that stress can impact eyewitnesses' memories of faces (Marr et al., 2020). As eyewitnesses creating facial composite images have been through a stressful event and may also find the process of composite construction stressful, it is likely that stress has some impact on the construction of the facial composite.

Of course, it is not ethical to simulate a crime so that a facial composite can be constructed in the most life-like situation possible, nor is it ethically sound to create scenarios whereby participants feel nervous, stressed, or uncomfortable without obtaining informed consent.

Interestingly, Valentine and Mesout (2009) asked individuals who visited the Horror Labyrinth at the London Dungeon to identify the faces of actors who played a role in the attraction. They found that individuals who experienced more anxiety during the attraction were less able to identify the faces of actors. In the aforementioned experiment, participants would have still entered the Horror Labyrinth without participation in the study. However, it is not feasible, or necessary to use a similar procedure in the experimentation in the current thesis. Once a new composite construction procedure is developed through laboratory experiments, it is tested by the police using a field trial, after which changes can be made to the procedure depending on its success. Therefore, it is unnecessary to replicate the stress

faced by eyewitnesses, as this will be assessed more validly through the field trial (Frowd, Pitchford et al., 2012).

Although eyewitnesses creating facial composite images are likely to be stressed, they are also highly motivated to create an accurate composite image (Bayer, 2015). Therefore, it is crucial to explain to participants the importance of creating an accurate image, to increase their motivation. Ensuring that participants understand the seriousness of the task increases ecological validity (Hartson & Pyla, 2012), compared to a scenario whereby a participant is uninterested in creating an accurate composite image. In addition, the ecological validity of this research could be increased further by recruiting witnesses and victims of crime in the experiment and publishing the composite constructed for identification by a police officer or public member. However, if a composite is constructed in a condition which is as yet untested, and the image produced is not accurate enough to result in the perpetrator of crime being identified, a potentially dangerous individual is free as a consequence of the experiment. Furthermore, there would be a lack of control over the encoding time of the offender or the time between encoding and composite construction. There may also be some targets who are more easily identifiable, perhaps due to an unusual tattoo or scar on the face, which would be a confounding variable in the experiment. Although it would be interesting to assess how manipulating the number of screens used during composite construction impacts identification of composite images in real cases, it is not possible due to the practical matters discussed.

In the case of the current remote research, some control over the environment the participant is in while viewing the face. Due to this level of control, replication of the experiment is more likely, ensuring that any changes to the outcome are reliably due to the manipulation before changing police protocol. If police protocol is altered

to reduce the number of screens used during composite construction, but a confounding variable (such as targets with identifying features, i.e., facial scars or tattoos) impacts the identification rates, there is a chance that composites constructed are not recognisable, and perpetrators of crime are not identified. Overall, controlled lab experiments in this research allow a cause-and-effect relationship between changes to the presented population size during composite construction and composite accuracy that a field experiment would not. However, once the lab produced reliable and consistent findings, this could then be tested in the field.

Another important limitation of lab-based experiments is the potential for experimenter bias. Specifically, the researcher's expectations may affect their interaction with participants during construction. To minimise such problems, the researcher is kept unaware of the target identities during Part 1 of the experiment, and therefore cannot influence face selection or enhancement to achieve a desired result. To further mitigate against researcher bias, target identities may be changed between experiments so that a researcher does not become familiar with them. If the same targets are repeatedly used, a researcher may be able estimate a target's appearance based on the participants' description during the CI, and so, even without viewing the target photograph, there is risk of experimenter bias. In the current research, as the researcher was absent during composite construction in Experiment 1 and therefore did not become familiar with the targets; the same targets were repeated in Experiment 2.

A further limitation of the methodology is that data collection was online via video conferencing. The literature clearly demonstrates the benefits of building rapport with an eyewitness during an investigative or forensic interview (Collins et al., 2002; Vallano & Compo, 2011). This statement is also true for eyewitness

interviews, whereby an eyewitness is interviewed by the police with the aim of gaining information about the crime that has been witnessed, or a description of the perpetrator (Abbe & Brandon, 2013). However, building rapport is considered to be more difficult via video conferencing than it is face-to-face (Nash et al., 2014). To mitigate against this possibility, the interviewer spent several minutes asking open ended friendly questions, as in Nash et al. (2014), to build rapport with the participant in the initial interview (before composite construction). Although face-to-face participation in the experiments included in this thesis would be preferable, this was not possible during data collection due to the global COVID-19 pandemic. Therefore, online meetings were conducted via videoconferencing platforms such as Skype, Zoom and Teams, and the potential negative impact to rapport building as a result of videoconferencing was mitigated through purposeful rapport building at the beginning of each meeting.

On balance, the methodology used in this thesis replicates facial composite construction experienced by eyewitnesses with the police as closely as is reasonably possible without causing stress to participants or violating ethics guidelines (British Psychological Society, 2018). The methodology also replicates that generally used in composite research (Frowd et al., 2015). Indeed, in terms of online interaction, it is the case that more witnesses and victims are interviewed in this way in real cases (Frowd, 2021).

This chapter presented the positivist research philosophy followed in this thesis, explaining the focus on empirical evidence to study human behaviour before discussing further details about the research type, strategy, time horizon and sampling strategy. This chapter then discussed in detail the data collection for composite construction, providing a thorough description of the composite construction procedure as well as outlining the data collection for composite naming and likeness ratings throughout the

experiments. Following this was an explanation of the data analysis used to examine the data which follows the gold standard, using the regression technique Generalised Linear Mixed Models (GLMM) to explore patterns in the data. Finally, was an evaluation of the methodology covering the ecological validity of the research, the strengths and limitations involved in laboratory-based experimentation and the use of online data collection throughout the PhD research.

3

EXPERIMENT 1

REDUCING THE POPULATION SIZE DURING EVOFIT SELF-ADMINISTERED CONSTRUCTION

Abstract

This experiment aimed to determine the optimum number of screens used during EvoFIT Online composite construction to understand the impact of cognitive load on witness memory during the construction process. Forty participants created a facial composite of an unfamiliar celebrity using *One, Two, Three* or *Four Screens* to select the face images during the procedure. The accuracy of composites was tested through composite naming by 24 participants who were familiar with the celebrities and through composite likeness ratings by 69 participants unfamiliar with the celebrities. The results demonstrated that composites constructed using *One Screen* were identified most frequently, and composites constructed using fewer screens were

generally deemed most accurate. This finding suggests that decreasing the number of screens during EvoFIT composite construction may benefit composite likeness due to a reduction in participant cognitive load.

EvoFIT facial composite construction is frequently used by the police in the UK and internationally to aid eyewitnesses in creating a face image of a perpetrator of crime. An eyewitness must have viewed the perpetrator's face in order to create an accurate facial composite image; therefore, EvoFIT is often used by victims of face-to-face crimes, such as robbery and sexual assault. Individuals committing such crimes are very dangerous, and identifying them is important for the safety of our communities. Although facial recognition is an automatic and seemingly easy task (Richler et al., 2009), humans are notoriously bad at recognising faces which are somewhat unfamiliar to them (Burton et al., 2015; Bruce & Young, 1986). While it is relatively easy to recognise the faces of family members or friends, even when changes have been made to the face through ageing or plastic surgery (Chapman, 2018), recognising a face which has only been viewed briefly, or may have been deemed unimportant at the time (e.g., con artists or so-called "doorstep pedlars"), is very difficult (Hancock, 2000).

For composite construction to be effective as an investigative technique, composite systems must enable witnesses and victims of crime to create a facial likeness that is accurate and thus identifiable. Consequently, facial composite systems, including EvoFIT, must be continuously developed and employ evidence-based best practice. The current research aims to improve the process of composite construction and the resulting composite image by reducing the cognitive load

involved in the construction process. The cognitive load will be reduced by decreasing the number of screens (and thus the number of faces) viewed during the construction process, with the aim of creating a composite image with greater accuracy than found in current approaches.

Cognitive Load

During EvoFIT composite construction, eyewitnesses view many face images when selecting the composite facial shape and texture. It is theorised that viewing many face images may result in a high intrinsic cognitive load, which is related to the number of elements interacting in a task (Sweller, 1988). For example, many elements interact during the face selection stage of composite construction, that is, a large number of face images that are compared to each other during the task.

If the cognitive load experienced by the participant during EvoFIT composite construction becomes too high, participants are likely to experience cognitive overload (the result of information input being greater than processing capacity: Bishara, 2021), during which memory capacity and decision-making ability are diminished (Byyny, 2016). If participants do experience cognitive overload, they may struggle to remember the target face, reducing their ability to create a good likeness. Furthermore, they may be less able to make accurate decisions regarding the face images selected during the construction procedure. For these reasons, it is important to first understand the impact that cognitive load has on composite construction and to look to reduce the intrinsic cognitive load during the EvoFIT construction process to optimise the procedure and increase the identifiability of the composite images created.

Eye Region

In terms of facilitating witness memory, the eye region is considered the most important area of the face for accurate facial recognition (Royer et al., 2018). Yet, the literature suggests that other regions are also important; for example, the nose, as well as the eye region, attracts the most visual attention (Luria & Strauss, 2013), or the eye region and the mouth attract the most attention as they are involved in communication and therefore command attention (Ellis et al., 1978). Importantly, the eye region is consistently considered important for face perception, with researchers arguing that this is objectively the best region to use when identifying a face (Peterson et al., 2008; Peterson & Eckstein, 2012; Rizzo et al., 1987).

The superiority of the eye region for face recognition has been demonstrated using a variety of methods. Covering one feature of the face and inviting participants to identify the familiar individual demonstrated that faces displayed with the eye region covered were identified less frequently than faces displayed with any other feature covered (Royer et al., 2018). Tracking a participant's eye movements while asking them to identify the familiar individual in a photograph of a face resulted in an increased focus on the eyes than on any other area of a face. This result was exaggerated when asking participants to determine the facial expression shown in the photograph, demonstrating that the eye region is not only important for face recognition but also crucial for recognition of emotions and non-verbal communication (Nelson & Mondloch, 2014).

Due to the importance of the eye region for face recognition, eyewitnesses are asked to focus on the eye region when creating a facial composite using the EvoFIT system (Fodarella et al., 2017). This procedure was introduced after an identifiable composite was constructed with the police, in which the eye region more closely

resembled that of the target face. After comparing the difference in accuracy between composites constructed by selecting face images based on the whole face, or the eye region, it was clear that focusing on the eye region produced a superior composite image (Martin et al., 2018; Portch et al., 2017; Frowd et al., 2019). In the current experiment, participants are asked to focus on the eye region during face selection, as per standard practice. However, focus on this region is manipulated during a likeness rating task to understand whether the composites are more accurate when rated based on the whole face or only the eye region.

A typical measure used to assess the accuracy of facial composite images is inviting participants to rate the accuracy of face images in comparison to the target face (Gibson et al., 2009). Such a measure is important for understanding how accurately a composite resembles the target. As the current experiment manipulates the composite construction process by reducing the number of screens used for face selection, composite rating can also be used to ensure that there are no detrimental effects to the composite images as a result of changing the construction procedure. This measurement also aims to explore the accuracy of the eye region in composites constructed with fewer screens, and therefore a lower cognitive load. In addition to the eye region, any change to the construction procedure must not negatively impact composite *Internal* and *External Features*, as both regions are important for the construction of an identifiable facial composite (Frowd & Hepton, 2009).

Internal and External Features

Although it is largely agreed that the eye region is central to face recognition (Peterson et al., 2008; Peterson & Eckstein, 2012; Rizzo et al., 1987), the importance of *Internal Features* and *External Features* for face recognition has been long

debated. One view is that *Internal* and *External Features* are stored holistically, as one unit of information (Andrews et al., 2010; Axelrod & Yovel, 2010). However, an alternative view is that *Internal* and *External Features* are stored separately, which explains why a face can typically still be recognised despite a change to the hairstyle (Chan & Ryan, 2012). Under the assumption that *Internal* and *External Features* are perceived and stored separately, it may be important to understand whether internal or external features of a face are the most critical for face recognition, particularly for unfamiliar faces.

The literature demonstrates that the importance of *Internal* and *External Features* is dependent on the familiarity of the face. For recognition of familiar faces, there appears to be a consensus that *Internal Features* are more vital to recognition than *External Features* (Bruce et al., 1999; Ellis et al., 1978; Young et al., 1985; Bruce & Young, 1998). However, for recognition of unfamiliar faces, there is some evidence which suggests that *External Features* are most important (Bruce et al., 1999; Kramer et al., 2017; O'Donnell & Bruce, 2001), with other evidence suggesting that there is no difference in the importance of *Internal* and *External Features* (Ellis et al., 1978).

In an earlier version of EvoFIT, face images containing *Internal and External Features* (whole composite faces) were displayed for selection over *Four Screens*. However, blurring or removing the external features during face selection resulted in internal features which were deemed more accurate, improving the likelihood of a composite being identified correctly (Frowd, Skelton et al., 2012). This finding is supported by Havard (2021), who demonstrated that, for UK participants, faces matched based on the whole face were the most accurate, but that faces matched based on the *Internal Features* were slightly more accurate than those matched based

on the *External Features*. However, this paper also demonstrated cultural differences. Chinese participants matched faces equally for accuracy based on the *Whole Face* and the *Internal Features*, with faces matched based on the *External Features* rated significantly less accurately. Contradictory literature suggests that *Internal* and *External Features* are equally as important for the recognition of unfamiliar faces (see, Andrews et al., 2010; O'Donnell & Bruce, 2001). However, the increase in composite likeness when the focus was shifted to internal features during the construction process suggests that, for EvoFIT composites, internal features are the most important (Frowd, Skelton et al., 2012). As participants will view fewer face images when selecting the internal features during the construction procedure, it is important to assess whether the accuracy of the individual facial regions is not impaired as a result of the new procedure.

Editing facial composite images to display only the *Internal* or *External Features* and inviting participants to rate the likeness of the composite based solely on the features presented provides important information about the accuracy of said features. For example, if composites constructed using *Four Screens* are rated highly based on the internal features, but composites constructed using *One Screen* are rated very low based on the internal features, it may be inferred that reducing the number of screens is problematic for these features. In this case, changing the composite construction procedure is unlikely to be beneficial for increasing the number of criminal perpetrators identified through facial composite images.

In addition, if reducing the number of screens from *Four Screens* to *One Screen* reduced the likeness of composite *External Features*, it would be inferred that reducing the number of screens during composite construction is problematic. This thesis does not anticipate a reduction in likeness for the *Internal* or *External Features*

when reducing the number of screens during EvoFIT composite construction. Yet, this measure acts as an important check to ensure that neither *Internal* nor *External Features* are negatively impacted by reducing the number of screens during composite construction, which would reduce overall composite accuracy.

Stage of Construction

EvoFIT construction starts with witnesses viewing various arrays of faces (typically four arrays, each containing 18 faces) and selecting the six faces from this array which look most like the target face (e.g., the perpetrator of crime or, in the case of this experiment, a photograph of a face viewed 24 hours previous). Once six faces that best represent the target's face shape have been selected, six more faces are selected to represent the target's face texture. Next, the singular best face is selected. In the current experiment, an image of this face is saved as *First Generation*. This process is repeated exactly; however, the faces should appear more accurate due to the nature of the genetic algorithm used. Again, the singular best face is selected, and this is saved as *Second Generation*. This face is then edited to improve the likeness (i.e., how much the face resembles the target).

During EvoFIT composite construction, eyewitnesses go through several stages to create the final face image. However, there is currently no understanding as to how important each stage is for the construction of an accurate composite, and how a composite image changes as it moves through the various stages of construction. Furthermore, there is no understanding of how the number of screens used during the initial face selection stage of EvoFIT construction may impact the ability of a participant to utilise these further stages of construction, such as *Holistic Tools*, the

Shape Tool and the selection of *External Features*. Each of these stages are described in more detail below.

Experimental Aims

This experiment sought to understand the impact of cognitive load on witness memory. Specifically, the first aim was to determine the optimum population size during EvoFIT Online construction. Frowd and Grieve (2019) demonstrated that reducing the number of screens from *Four* to *Two* during face selection at the start of EvoFIT composite construction improved the accuracy of the resulting composites. The current experiment was designed to further reduce the number of screens used incrementally (from *Four Screens* to *One Screen*) as a between-groups factor. The accuracy of facial composites was measured by the percentage of correctly named composites and the likeness ratings of facial composites against target images on a Likert (1932) scale of one to seven.

The second aim of Experiment 1 was to understand whether composite *Internal* or *External Features*, or the eye region, are negatively affected (indicated by a low rating score) by reducing the number of screens during composite construction. The *Internal* and *External Features* should both accurately depict the target face to increase the likelihood of a composite image being identified correctly. Moreover, as the eye region is deemed to be important for face recognition (see, Royer et al., 2018), it is important to assess the accuracy of this region for composites construction using fewer screens. Therefore, the third aim of Experiment 1 was to understand how cognitive load impacts participants' ability to utilise each stage of composite construction (face selection, *Holistic Tools* and the *Shape Tool*) effectively. Specifically, it is hypothesised that:

H1: Composites constructed using fewer screens will be more accurate and will therefore be identified more frequently and receive higher likeness ratings than composites constructed using the typical *Four Screens*.

H2: Reducing cognitive load will increase the composite likeness after the use of image-enhancing tools towards the end of the construction procedure. This hypothesis is based on the proposal that reducing witness cognitive load in the early stages of composite construction will allow participants to enhance the face more accurately in the later stages of composite construction.

Method

Part 1- Composite construction

Design. Part 1 employed a between-subjects design wherein the number of screens used for composite construction was manipulated (*One Screen* vs *Two Screens* vs *Three Screens* vs *Four Screens*). A nominal 24-hour delay was implemented between participants viewing a target photograph and subsequent composite construction using EvoFIT Online. This reflects the typical composite construction procedure.

Participants. Participants were 40 adults (29 females, 11 males) between 18 and 60 years (M = 31.23, SD = 11.28). All participants were recruited to be unfamiliar with the target images (England Footballers). Participants were recruited using the participant recruitment websites SONA and Call for Participants and were rewarded course credit or a £5 online shopping voucher for their participation. An equal number of participants were allocated to each level of *Screens (N = 10)*.

Materials. The target stimuli were 10 photographs of Footballers who have played internationally for England between 2010 and 2020. Images were displayed in colour and were approximately 8cm wide by 10cm high. None of the targets had

distinctive characteristics that would make them easy to identify (such as facial tattoos). Composites were constructed using EvoFIT Online, a self-administered version of the EvoFIT App.

Procedure. Participants were tested individually. Part 1 occurred over two days. On the first day, participants received a briefing sheet via email containing details about the experimental process. Participants also received an email containing the target photograph and were asked to view the image for 30 seconds and to delete the email after viewing.

Twenty to twenty-eight hours later, participants were asked whether they had viewed the photograph in accordance with the instructions. No participant stated that they had viewed the photograph for longer or shorter than the required time, or that they had reopened the email containing the photograph after the 30 seconds. As there was no way to control for participants re-viewing the photograph or viewing the photograph for a different time length, the participants were trusted to be honest. Participants then received a link to the EvoFIT Online website. The typical procedure for EvoFIT Online took place whereby participants viewed a video explaining the construction process before being guided through EvoFIT in a self-administered style (Martin et al., 2018). The construction procedure took approximately 40-60 minutes. Composite construction was the same for each participant, other than the experimental condition (i.e., the number of screens viewed). After construction, participants received a further email with the debriefing information for the experiment, containing information relating to the experimental aims.

Part 2a - Composite Naming

Design. Part 2a employed a mixed design in which the between-subjects independent variable (IV) was *Screens* with four levels: *One Screen, Two Screens, Three Screens* or *Four Screens*, and the within-subjects IV was *Naming* with two levels: *Spontaneous* and *Cued*. The dependent variable (DV) was the accuracy of the composites, which was measured by the percentage of composites named correctly by participants familiar with the targets.

Each composite image was named twice by each participant. The first time a composite is named is referred to as *Spontaneous Naming* because participants do not know who the targets are, only that they are part of a group, such as England Footballers or actors from a particular soap. *Spontaneous Naming* replicates a scenario whereby an individual sees a composite image of a perpetrator of crime in the media and attempts to recognise them.

After *Spontaneous Naming* is complete, participants view each of the original target photographs and attempt to name the individual based on their photograph. As participants should be familiar with the target pool (for example, fans of England International football), it is expected that participants name at least eight of the 10 targets correctly. The second time a composite is named is referred to as *Cued Naming* because participants have already viewed (and attempted to name) the targets, and so should be able to identify the composite images more accurately, assuming that the composites are accurate representations of the targets. This step in the experiment does not replicate a real-life scenario as closely as *Spontaneous Naming* but may represent a scenario whereby somebody has an idea that they know who committed a crime and views the composite aiming to confirm or deny this. This step in the experiment is useful for providing rich data, particularly if naming rates were low for

Spontaneous Naming, as this would make analysis difficult due to a floor effect (Michalos, 2014).

Participants. Participants were 24 adults (4 females, 20 males), aged between 22 and 65 years (M = 25.75, SD = 9.17), familiar with England International footballers. Participants were recruited using the participant recruitment websites SONA and Call for Participants and were rewarded course credit or £1 for their participation.

Participants were recruited to be familiar with the targets. As per the *a priori* rule*,* participants had to successfully name 80% of the targets to be deemed familiar with them. Two participants failed the *a priori* rule, so their data were removed, and new participants were recruited to ensure consistent numbers of participants across conditions. An equal number of participants were allocated to each level of *Screens* (N = 6).

Materials. Composites constructed during Part 1 were displayed via a PowerPoint presentation. Four presentations were created (one for each level of *Screens*). Each PowerPoint contained 10 composite images presented in a different order for each participant. A fifth presentation containing the target photographs was created, displayed in a random order for each participant and was used to ensure familiarity with the targets. An example slide containing a facial composite image of Jack Wilshere is depicted in Figure 8.

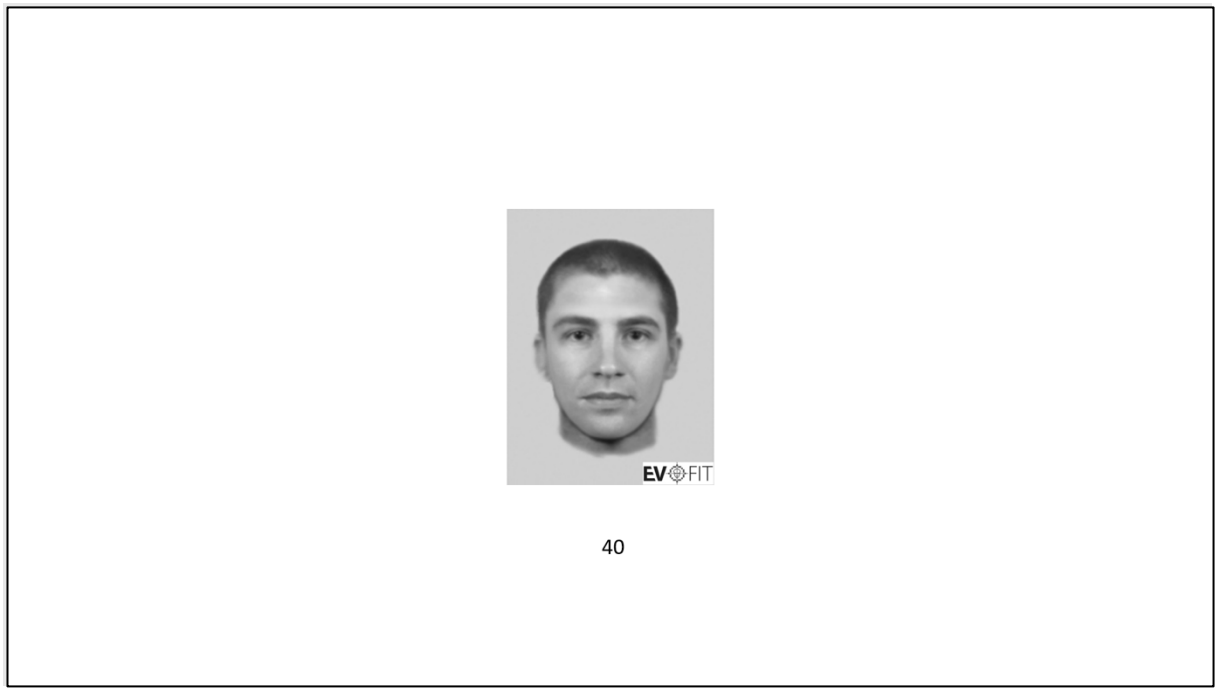


Figure 8. Example Slide for Part 2a: Composite Naming (to scale).

Procedure. Following the briefing, participants viewed the PowerPoint presentation via the 'screen share' feature on Skype. Participants were asked to name the footballer depicted in each composite image and were invited to guess if unsure *(Spontaneous naming)*. After viewing each image, participants were asked to name the 10 footballers from their target photographs. Participants then viewed the same composite images and attempted to name the targets for a second time *(Cued naming)*. The procedure took approximately 10 to 15 minutes, including debriefing.

Part 2b - Composite Final Image Rating

Design. A mixed design was used in which the within-subjects IV was *Screens (One, Two, Three* or *Four)* and the between-subjects IV was *Task*; the type of composites participants were asked to rate (*Internal Features, External Features or Whole Faces)*. The DV was the accuracy of the composites, rated for similarity to the target

face using a Likert scale (1 = very poor likeness, 7 = very good likeness) against the target image.

Participants. Participants were 45 adults (29 females, 16 males) between 18 and 60 years (M = 31.18, SD = 11.54). Participants were recruited using the participant recruitment websites SONA and Call for Participants and were rewarded course credit or £1 for their participation. Participants were recruited on the basis that they were not familiar with the targets. To be deemed unfamiliar, participants must have been unable to name 80% of the targets (England Footballers). No participants failed this *a priori* rule. An equal number of participants were randomly allocated to each level of *Task* (N = 15).

Materials. Composite *Internal Features*, *External Features* and whole composite images constructed in Part 1 were displayed in three separate PowerPoint presentations. Each slide contained a target image and four composites, one from each level of *Screens*. Multiple versions of each PowerPoint presentation were created, displaying the order of the slides and the order of the composites on each slide randomly. Example slides for rating of composite *Internal Features*, *External Features* and whole composite images are depicted in Figures 9-11.

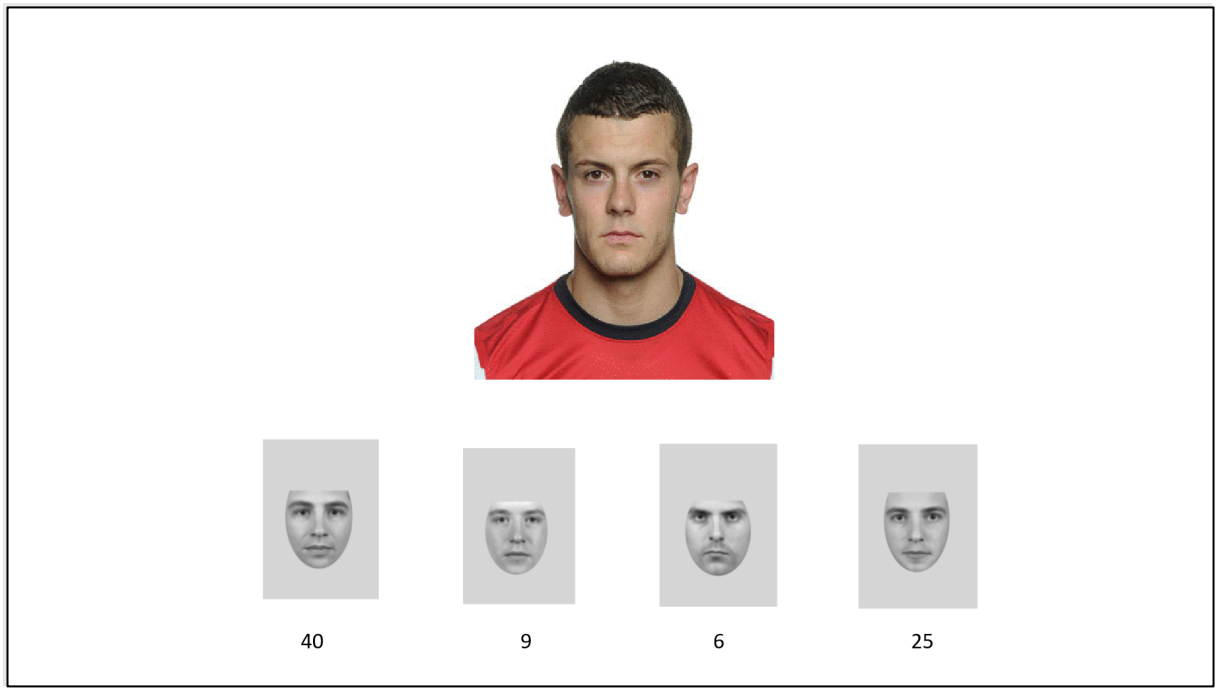


Figure 9. Example Slide for Part 2b: Composite Final Image Rating (*Internal Features*). Note, from left to write, *Internal Features* of composites constructed using *One Screen, Two Screens, Three Screens, Four Screens*.

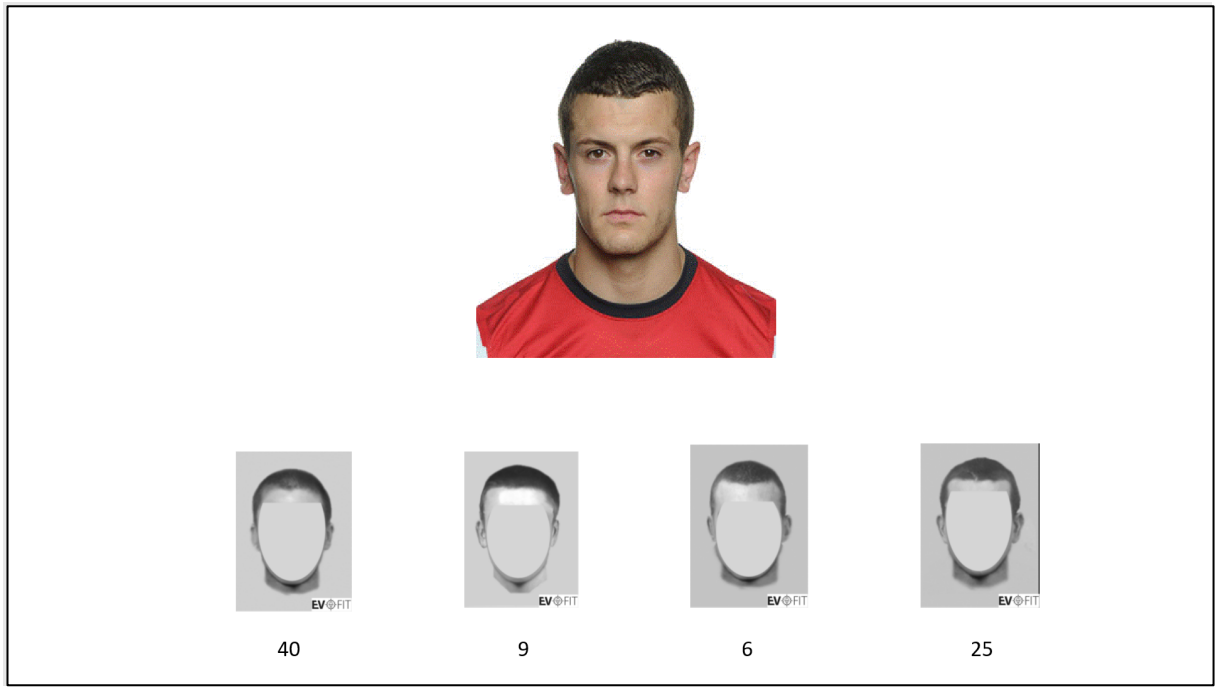


Figure 10. Example Slide for Part 2b: Composite Final Image Rating (*External Features*). Note, from left to write, *External Features* of composites constructed using *One Screen, Two Screens, Three Screens, Four Screens*.

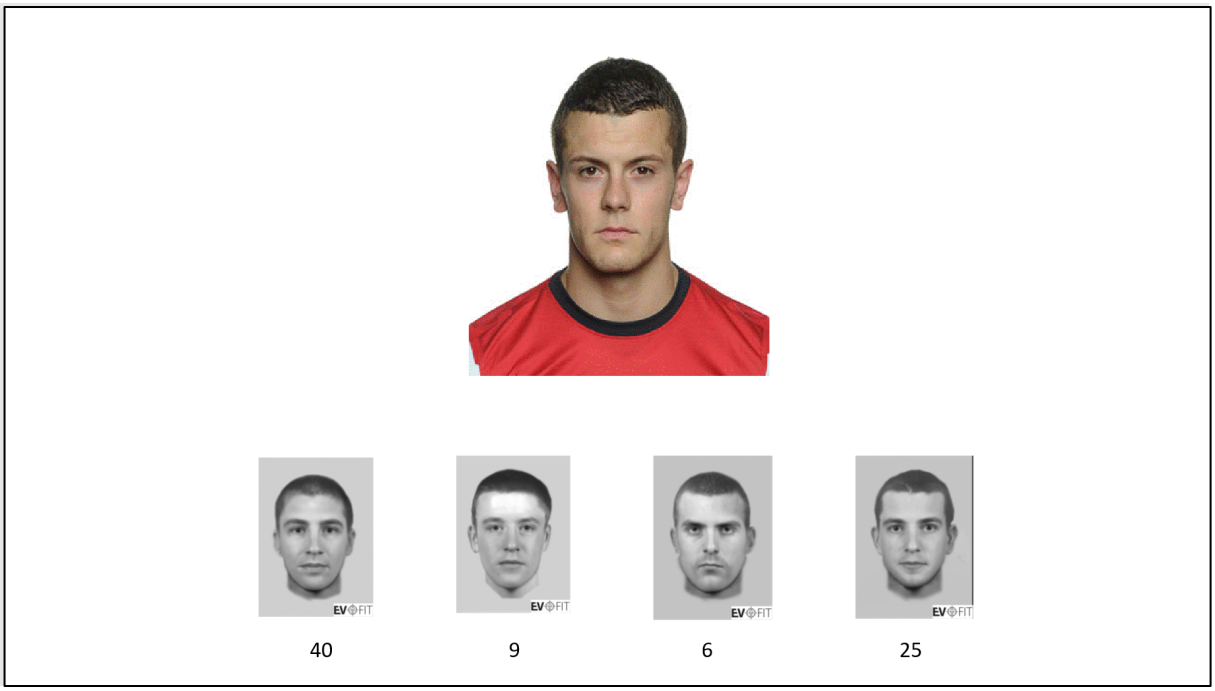


Figure 11. Example Slide for Part 2b: Composite Final Image Rating (*Whole Composites*). Note, from left to write, facial composites constructed using *One Screen, Two Screens, Three Screens, Four Screens*.

Procedure. Participants viewed the PowerPoint presentation via the 'screen share' feature on Skype. They verbally rated each composite image using the seven-point Likert scale, and the number given was recorded by the researcher. Participants were instructed to inform the researcher if they recognised any target images. If participants recognised two of the 10 targets, they were not deemed unfamiliar with the targets and their data was not included in the study. However, no participants were able to name two or more of the targets, so this *a priori* rule was not utilised. The procedure took approximately 10 to 15 minutes, including debriefing.

Part 2c - Intermediate Composite Rating

Design. Part 2c employed a mixed design. The between-subjects IV was the face region: *whole face* or *eye region* referred to as *Region,* and the within-subjects IVs were *Screens: One, Two, Three* or *Four Screens, and Stage: First Generation, Second Generation, Holistic Tools* or *Final Image.* The DV was the accuracy of the composites, measured by how closely participants judged them to resemble the target image. Composites were rated for similarity to the target face using a seven-point Likert scale (1 = very bad likeness, 7 = very good likeness).

Participants. Participants were 24 adults (19 females, 5 males) between 18 and 57 years (M = 25.75, SD = 9.17). Participants were recruited using the participant recruitment websites SONA and Call for Participants and were rewarded course credit or £1 for their participation. Participants were recruited based on being unfamiliar with the targets. To be deemed unfamiliar, participants must have been unable to name 80% of the targets; no participants failed this *a priori* rule. An equal number of participants were randomly allocated to each level of *Type* (N = 12).

Materials. Composite images constructed in Part 1 were displayed via a PowerPoint presentation. Each slide contained a target image and four composites, one from each level of *Screens*. Four versions of each PowerPoint presentation were created, displaying the order of the slides and the composites in a different order. An example slide is depicted in Figure 12.

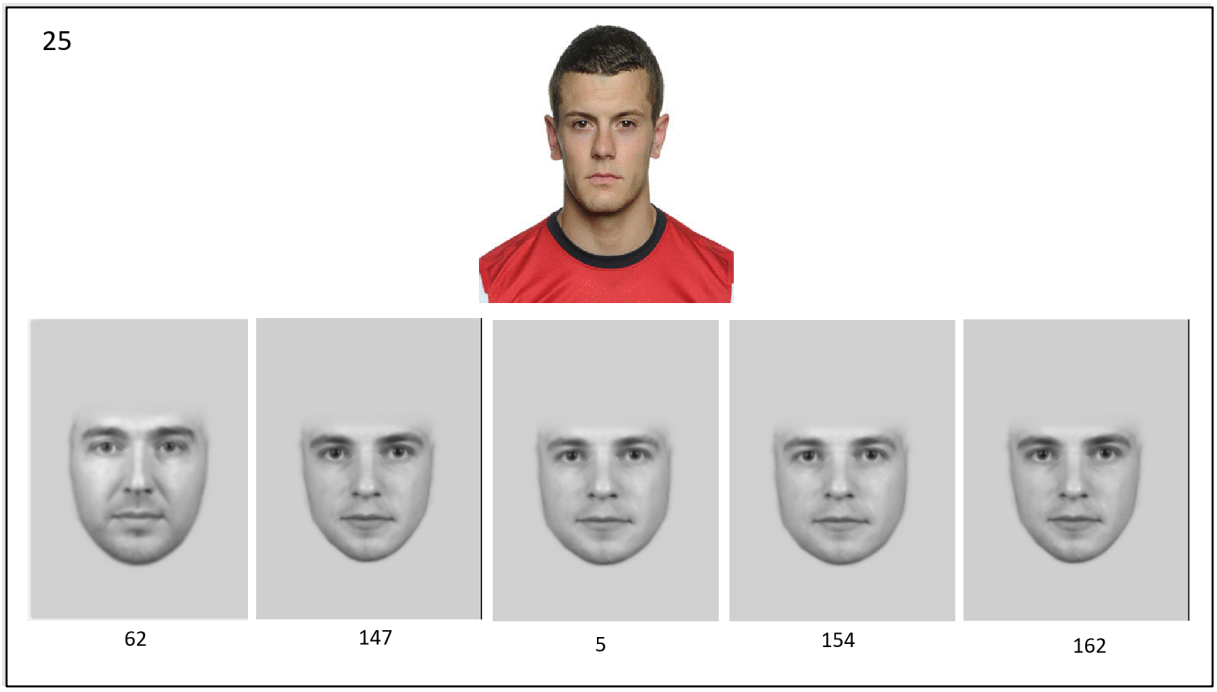


Figure 12. Example Slide for Part 2c: Intermediate Composite Rating

Note. From left to write, composite images from *Random Face, First Generation, Second Generation, After Holistic Tools, Final Image*. The order of composites presented on screen was randomised for participants.

Procedure. Participants viewed the PowerPoint presentation via the 'screen share' feature on Skype. Participants were asked to rate the composite based on the whole face or the eye region. Using the seven-point Likert scale, they verbally rated each composite image compared to the photograph. Participants were instructed to inform the researcher if they recognised any target images. The same *a priori* rule as Part 2b was implemented but was not utilised as no participant was able to name two or more of the targets. The procedure took approximately 10 to 15 minutes, including debriefing.

Results

Part 2a - Composite Naming

Responses to facial composites and target pictures were scored for accuracy. Responses were coded as correct and assigned a value of 1 when participants gave the correct name for the target image and composite image. Responses were coded as incorrect and assigned a value of 0 when a wrong name or "don't know" response was given for the composite , the target was identified correctly. Responses were assigned a value of 2 when the target was not identified, and these data were not included in the analysis. Incorrect names for target pictures occurred 18 times (by 13 participants), across the four conditions (*One Screen, Two Screens, Three Screens, Four Screens)*. As such, the mean correct naming for target pictures was very high (M = 92.50%, SD = 7.94%).

Correct responses were much lower overall for Spontaneous naming of facial composites (M = 3.60%, SD = 18.68%), compared to Cued naming (M = 26.13%, SD = 11.21%). Although it is typical for composites to be named less frequently than target photographs as they are error-prone stimuli, composite naming in this experiment was unusually low compared to past research, which typically demonstrates a naming rate of 60% (Frowd, 2002). Table 1 displays the mean naming results for composites at each level of *Screens* and *Naming*.

Table 1. Spontaneous and Cued Correct Naming of Composites for Each Level of Screens

Number of Screens	4	3	2	1	Total
Spontaneous naming	0.00 (0 / 56)	1.85 (1 / 54)	7.02 (4 / 57)	5.45 (3 / 55)	3.60 (8 / 222)
Cued naming	17.86 (10 / 56)	18.52 (10 / 54)	31.58 (19 / 57)	36.36 (20 / 55)	26.13 (59 / 222)
Total	8.93 (10 / 112)	10.18 (11 / 108)	19.30 (23 / 114)	20.91 (23 / 110)	14.87 (67 / 444)

Note. Figures are expressed in percentage and calculated from participant responses in parentheses: summed correct responses (numerator) and total (correct and incorrect) responses (denominator). Data are presented for composites for which participants correctly named the relevant target photographs (N = 222 out of 240).

These results indicate that composites constructed using *Two Screens* are the most accurate for *Spontaneous naming*, but composites constructed using *One Screen* are the most accurate for *Cued naming*. Further, the result for *Cued naming* indicates that, as composite accuracy increases, the number of screens used during composite construction decreases. This pattern of results supports the hypothesis that lessening the cognitive load by reducing the number of screens used for face selection is beneficial for accurate composite construction.

Due to the uncharacteristically low naming rates, testing ceased as it was inferred that composite construction did not meet the minimum standards found in past research (Stephan et al., 2019) or when used by the police (Frowd et al., 2011). As such, it was deemed unethical to continue collecting data from participants if it would not be helpful for research. Due to the low sample size, no inferential statistics were carried out as the results were unlikely to be reliable (Binu et al., 2014).

Based on the mean rates of naming, these data support the hypothesis that composites constructed using fewer screens are more accurate than those constructed using *Four Screens*. For *Spontaneous naming*, composites constructed using *Two Screens* were named most frequently, followed closely by *One Screen*, and then *Three* and *Four Screens*. For *Cued naming*, composites constructed using *One Screen* were named most frequently, followed by *Two, Three* and *Four Screens* consecutively.

Part 2b - Composite Final Image Rating

To further understand the impact of cognitive load on composite accuracy, *Internal* and *External Features* of composites as well as *Complete Composites* constructed using *One, Two, Three* or *Four Screens* were rated for accuracy using a Likert scale of 1-7. Table 2 presents the mean and standard error for rating of composites at each level of *Task* constructed at each level of *Screens.*

Table 2. Mean (and Standard Error) for Rating of Composites Constructed at Each Level of Screens

Task	Number of Screens				
	4	3	2	1	Mean
Internal Features	2.88	2.95	2.75	3.19	2.95
	(0.15)	(0.15)	(0.15)	(0.16)	(0.08)
External Features	3.05	3.41	3.12	3.12	3.18
	(0.13)	(0.14)	(0.14)	(0.14)	(0.07)
Whole Faces	3.07	3.07	2.79	3.43	3.09
	(0.14)	(0.14)	(0.13)	(0.16)	(0.07)
Mean	3.00	3.14	2.89	3.25[a]	3.07
	(0.08)	(0.08)	(0.08)	(0.09)	(0.04)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant-item SE of the mean. [a] $p < .1$

Based on the pattern of data in Table 2, it is inferred that composites constructed using *One Screen* were rated as being the most accurate, followed by composites constructed using *Three, Four* and *Two Screens*. Table 2 also illustrates that composite *External Features* are rated the most accurate compared to the target photograph, followed by *Whole Faces* and *Internal Features*.

To gain an understanding of the differences between the composite rating in each condition, individual ratings of composite items from participants were analysed using the regression technique, Generalised Linear Mixed Models (GLMM), in SPSS. GLMM models IVs (*predictors* or *fixed effects*) in the context of *random effects*. In this case, the random effects are (i) the participants and (ii) the composite items. This stage of the experiment involved two predictors: *Screens* (coded as *1 = One Screen*, *2 = Two Screens*, *3 = Three Screens* and *4 = Four Screens*) and *Task* (coded as *1 = Whole face, 2 = Internal Features, 3 = External Features).* The DV was individual rating on a Likert scale of 1-7, with the model set to accommodate ordinal responses using multinomial logistic regression.

Estimation of parameters for GLMM is Penalised Quasi-Likelihood (PQL) in SPSS (IBM, 2020), a standard iterative-fitting method. As the current sample has balanced data and is sufficiently large (by design), the residual method (e.g., cf. Satterthwaite approx.) was selected as degrees of freedom for computing tests of significance. Default settings for convergence criteria were used: parameter convergence with an absolute difference of 1E-6 and a maximum of 100 iterations for the algorithm's inner loop. For both fixed and random effects models that were conducted, Beta (slope) coefficients (*B*), standard errors of *B* [*SE(B)*], effect sizes [*Exp(B)*] and confidence intervals (*CI*, all reported at 95%) were checked to be within

sensible limits, neither too low nor too high, that might otherwise indicate an issue with the fit of the model.

The analysis considered the most appropriate method to compute parameter estimates. As *Task* is between subjects, each model for this IV contained the best combination of random intercepts. As *Screens* each model for this IV contained the best combination of random intercepts and random slopes. There are two methods available in SPSS, a Model-based method and a Robust method, which is sometimes preferable if standard errors for the Model-based analysis are unusually high. Model-based and Robust models gave the same pattern of significant and non-significant differences. However, Model-based was selected for presenting the results in the final model since the resulting standard errors for coefficients [*SE(B)*] of the interaction term were substantially lower (cf. Robust), thus providing a better fit of the data. The analysis initially assessed the composition of random effects. This assessment followed the 'gold' standard statistical procedure of Barr et al. (2013).

For the mixed design, estimates of variance for both random intercepts, participants and items were computed. GLMM were duly conducted for each model, each containing the best combination of random intercepts and random slopes. A hypothesis-testing (confirmatory) approach was conducted that comprised three models, each specified with different fixed effects (predictors) along with appropriate random effects (as described above). One model contained *Screens* with a random intercept for participants, a second model contained *Task,* and a third, full-factorial model contained the two predictors and the interaction between them. The model for *Screens* was significant [F (3, 1791) = 4.36, p = .005], but the model for *Type* was not significant [F (2, 1792) = 0.34, p = .71], nor was the interaction [F (6, 1783) = 1.35, p

= .23]. Therefore, the model for *Screens* only was taken as the final model. Fixed coefficients are used to explore this significant result (Table 3).

Table 3. Model parameters from Likeness Ratings for Composites Constructed at Each Level of Screens. Comparisons are presented with reference to the lowest category (Four Screens); negative values of B indicate lower ratings of likeness with respect to the reference.

	B	*SE(B)*	*t (1791)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
Screens							
Three Screens vs. Four Screens	-0.18	0.13	-1.37	.17	0.84	0.65	1.08
Two Screens vs. Four Screens	0.18	0.13	1.39	.165	1.20	0.93	1.54
One Screen vs. Four Screens	-0.25	0.13	-1.92	.056	0.78	0.61	1.01

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix)] final, Model-based full factorial Corrected model [F (3, 1791) = 4.36, p = .005]. The model was specified with the lowest category of categorical predictors as reference (*Screens*; Four), and predictors were sorted in an ascending order. Information criteria are based on -2 log likelihood (*AICC* = 36451.55, *BIC* = 36462.52). Variance of random effects' intercept of participants for *Screens* [1.82, *SE* = 0.42, Z = 4.34, p < .001, *CI* (1.16, 2.85)].

This analysis demonstrated that composites constructed using *One Screen* were significantly more accurate than those constructed using *Four Screens.* However, there was no significant difference between composites constructed using *Four Screens* and those constructed using *Two* or *Three Screens.* To understand the mathematical pattern of these results, polynomial contrasts were simulated using

GLMM. As it was hypothesised that composite accuracy would increase as the number of screens used during composite construction decreased, a significant linear contrast was anticipated. However, the analysis demonstrated a non-significant linear contrast ($p = .70$, *Exp(B)* = 1.04) and a non-significant quadratic contrast ($p = .21$, *Exp(B)* = 0.95). Nonetheless, there was a significant cubic pattern in the data ($p < .001$, *Exp(B)* = 1.26), indicating that composite accuracy decreases between *One Screen* and *Two Screens,* then increases between *Two Screens* and *Three Screens* before decreasing once more between *Three Screens* and *Four Screens.*

Part 2c - Intermediate Composite Rating

To develop an understanding of the impact that number of screens has on the participant's ability to utilise each stage of construction, composite ratings based on the whole face, or the eye region, were analysed at each of the four stages of composite construction (*First Generation, Second Generation, Holistic* Tools and *Final Image*) as well as a random composite image (*Random*). As Region had little effect on the likeness ratings, Table 4 presents overall ratings of composites at each level of *Screens.*

Table 4. Mean (and Standard Errors) of Composites at Each Level of Screens and Stage

Stage	Screens				Mean
	4	3	2	1	
Random	2.47	3.02	3.14	3.00	2.91
	(0.10)	(0.12)	(0.11)	(0.10)	(0.05)
1st	2.76	2.93 [a, b]	2.59 [c]	2.56 [c]	2.71
Generation	(0.11)	(0.11)	(0.11)	(0.10)	(0.05)
2nd	2.81 [a, b]	3.04 [a, b]	2.62 [b, c]	2.77	2.81
Generation	(0.11)	(0.12)	(0.11)	(0.11)	(0.05)
After Holistic	3.03 [a, b]	2.93 [a, b]	2.92 [a]	3.16 [a, b, d]	3.01
Tools	(0.11)	(0.11)	(0.11)	(0.12)	(0.06)
Final Image	3.51 [a, b, c, d]	3.42 [a, b, c, d]	3.13 [a, b, d]	2.90 [a, b]	3.24
	(0.12)	(0.11)	(0.11)	(0.11)	(0.06)
Mean	2.91	3.07	2.88	2.88	2.93
	(0.05)	(0.05)	(0.05)	(0.05)	(0.02)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant-item SE of the mean. [a] $p < .05$ compared to *Screens:1, Stage:1,* [b] $p < .05$ compared to *Screens:2, Stage:1,* [c] $p < .05$ compared to *Screens:3, Stage:1,* [d] $p < .05$ compared to *Screens:4, Stage:1.*

It is clear from the table that a composite image which is irrelevant to the target and is chosen at random was rated higher than facial composite images at some stages of composite construction and, in some cases, higher than the *Final Image.* This result indicates that composite likeness was very poor in this experiment. To gain a better understanding of the pattern of composite likeness throughout the stages of composite construction, data for *Random Faces* has been removed from the following analysis.

To understand the impact that *Screens, Stage* and *Region* have on composite likeness, Generalised Linear Mixed Models (GLMM) were run in SPSS. This stage of the experiment involved three predictors, *Screens* (coded as *1 = One Screen*, *2 = Two Screens*, *3 = Three Screens* and *4 = Four Screens*), *Stage* (coded as *0 = Random Face*, *1 = First Generation*, *2 Second Generation*, *3 = Holistic Tools* and *4 = Final Image*) and *Region* (coded as *0 = Whole Face* and *1 = Eye Region*). The DV was individual ratings of composite likeness compared to the target photograph on a Likert scale of 1-7, with the model set to accommodate ordinal responses using multinomial logistic regression.

A hypothesis-testing (confirmatory) approach was conducted that comprised one model for each predictor: *Screens, Stage, Region*, as well as one full-factorial model for each interaction: *Screens*Stage, Screens*Region, Stage*Region, Screens*Stage*Region.* The model for *Screens* was significant [F (3, 4791) = 2.81, p = .038], as was the model for *Stage* [F (3, 3831) = 16.55, $p < .001$], but the model for *Region* was not significant [F (1, 4793) = 0.00, p = 1.00]. There was also a significant interaction between *Screens* and *Stage* [F (9, 3819) = 2.39), p = .011] but no interactions between *Screens* and *Region* [F (3, 4787) = 0.00, p = 1.00], *Stage* and *Region* [F (3, 3827) = 0.00, p = 1.00] and *Screens, Stage* and *Region* [F (9, 3803) = 0.00, p = 1.00]. Therefore, the final model contained *Screens* and *Stage* as well as the interaction between them.

To explore the interaction between *Screens* and *Stage*, four models were run to investigate *Stage* at each level of *Screens*. The analysis for *One Screen* demonstrated no significant effect of *Stage* [F (4, 1190) = 0.65, p = .62], the analysis for *Two Screens* had a significant effect of *Stage* [F (4, 1990) = 3.06, p = .016], the analysis for *Three Screens* did not have a significant effect of *Stage* [F (4, 1190) = 0.51, p =

.73] and the analysis for *Four Screens* had a significant effect of *Stage* [F (4, 1190) = 3.82, p = .010].

To further understand the data and explore the mathematical pattern of *Stage,* polynomial contrasts were simulated using GLMM in SPSS. The analysis demonstrated a significant linear pattern in the data (p = .004, *Exp(B)* = 1.22), and non-significant quadratic (p = .23, *Exp(B)* = 1.04) and cubic patterns (p = .72, *Exp(B)* = 1.02). This result demonstrated that a linear pattern is the best fit for the data, and composite likeness generally increases throughout the stages of composite construction.

The pattern of results in Part 2c of this experiment indicate that composite accuracy increases throughout the stages of composite construction; however, the *Screens*Stage* interaction demonstrates that the increase in accuracy throughout construction is different depending on the number of screens used.

Experiment 1 Discussion

The current experiment was designed to develop an understanding of the impact that cognitive load has on the construction of a facial composite image. More specifically, the aims of the current experiment were to determine the optimum population size during EvoFIT composite construction and to understand whether cognitive load negatively impacts participants' abilities to utilise each stage of composite construction.

The operationalisation of cognitive load in this thesis is based on Sweller's (1988, 2010) Cognitive Load Theory and focuses on one specific aspect of cognitive load, intrinsic load. Intrinsic cognitive load relates to the complexity of a task and, in this experiment, refers to the number of interacting elements that must be processed in

working memory during the task of selecting faces from a set number of face arrays (Sweller, 2010). Cognitive load was manipulated by altering the number of face arrays that are displayed to participants, either *One, Two*, *Three* or *Four Screens* of faces, with each screen containing 18 faces (the number of faces on each screen does not change throughout the experiment). This thesis tested the proposition that reducing the number of screens used for selecting face images at the beginning of composite construction decreases the number of interactive elements and therefore reduces the intrinsic cognitive load of the task.

The results demonstrated that Spontaneous naming was highest for composites constructed using *Two Screens*, followed by *One, Three* and *Four Screens*. Moreover, Cued naming was highest for composites constructed using *One Screen*, followed by *Two, Three* and *Four Screens*. Overall, these naming results demonstrate that composites constructed using fewer screens are more accurate than composites constructed using the typical *Four Screens*. Composite naming rates in the current experiment were lower than those typically achieved in past research (Frowd et al., 2015); however, the pattern of naming is still useful for gaining an understanding of composite accuracy for each level of *Screens.*

This pattern of correct naming, whereby correct composite naming increases as the number of screens used during construction decreases, replicates Frowd and Grieve (2019), who found that composites constructed using *Two* or *Three Screens* were more accurate than those constructed using *Four Screens*. This pattern indicates that reducing the number of screens used during composite construction is beneficial for the creation of an accurate composite image. One explanation for this result is that the intrinsic cognitive load of the task is reduced when the number of screens used to select face images is decreased. Based on the Theory of Cognitive Load (Sweller,

1988, 2010), tasks which include many interactive elements and therefore require a large amount of working memory capacity to complete result in diminished memory and decision-making ability. The current results support the proposition that reducing the number of interactive elements by decreasing the number of face images viewed during composite construction may improve participants' memories of the target face.

The findings from Part 2b demonstrated that composites constructed using *One Screen* were the most accurate, followed by those constructed using *Three, Four* and *Two Screens*. This finding somewhat supports the results from Frowd and Grieve (2019) and Part 2a of the current experiment, as composites constructed using the least number of screens were rated as the most accurate compared to the target. However, in Frowd and Grieve (2019), composites constructed using *Two Screens* were deemed the most accurate, followed by those constructed using *Three* and *Four Screens* (*One Screen* was not included), whereas composites constructed using *Two Screens* were found to be the least accurate in the current experiment. Similarly, composite naming in Part 2a of the current experiment was the most accurate for *One Screen*, followed by *Two, Three* and then *Four Screens*, following the same pattern as Frowd and Grieve (2019).

Frowd and Grieve (2019) and Part 2a of the current experiment used composite naming to determine the accuracy of composites instead of composite rating, whereas Part 2b used composite likeness rating compared to the target image. It may be inferred that composites constructed using *Two Screens* are identifiable by individuals familiar with the targets, resulting in high composite naming. However, they are deemed less accurate by individuals unfamiliar with the target. Thus, it may be theorised that there are inaccuracies in the face image that are deemed important for individuals rating the composite images but do not impede recognition. Generally,

likeness ratings are a proxy to naming and so differences between the two results are understandable.

One reason why composites constructed using *Two Screens* are less accurate than those constructed using *One, Three* or *Four Screens* may be due to the fact that the benefit of reducing the cognitive load during composite construction did not outweigh the benefit of selecting face images from more screens. When participants select faces from fewer screens, they miss out on selecting faces which may more accurately resemble the target. However, it is also conceivable that participants also benefit from improved memory of the target and improved decision-making ability. When composites are constructed using *Two Screens*, the cognitive load may not be reduced enough to improve participants' memories of the target or their ability to select the most accurate faces, but they nevertheless benefit from viewing more face images that may more closely resemble the target. Ergo, composites constructed using *Two Screens* are rated as being the least accurate. Even so, the difference rating for composites constructed at each level of *Screens* was small and so could not be analysed using the regression model.

External Features of the composite were rated higher than *Internal Features*, although this difference was not large enough to produce a statistically significant result. The literature dictates that *External Features* are the most important for the recognition of unfamiliar faces (Bruce et al., 1999; O'Donnell & Bruce, 2001; Kramer et al., 2018); therefore, when rating unfamiliar faces, participants may rate the *External Features* as being more accurate than the *Internal Features*, as this is the part of a face that would normally be the focus during unfamiliar face recognition. While this result is expected based on the literature, the absence of a significant difference between mean rating of *Internal* and *External Features* demonstrates that

neither internal nor *External Features* are impacted by reducing the number of screens during composite construction. This is very important as it is crucial to create a facial composite with accurate *Internal* and *External Features* to increase the likelihood of recognition (Frowd et al., 2007). Furthermore, there was a reasonable possibility that reducing the number of screens viewed during the construction of a facial composite may have negatively impacted the accuracy of the *Internal Features*, as this is the region displayed during face selection. The findings suggest that there was no significant interaction between the number of screens and the type of rating, so there is little difference in the accuracy of face *Internal* and *External Features* between the number of screens used, suggesting there is no detriment to the accuracy of composite internal or *External Features* when reducing the number of screens.

Based on the result of Part 2c, it appears that cognitive load may not have a negative impact on composite accuracy, as there does not appear to be a clear increase or decrease in composite ratings as the number of screens viewed during construction is reduced. However, it is also important to note that there is only a difference of 0.24 between the highest-rated condition (*Three Screens*) and the lowest-rated conditions (*One* and *Two Screens*). Furthermore, there is only a 0.02 difference between *Four Screens* and *One* and *Two Screens*. Due to the low variation in these results, the difference composite rating for each level of *Screens* could not be analysed using the regression model. Consequently, it is hard to determine whether cognitive load had a large impact on the difference in composite accuracy.

Throughout Part 2c, composites were rated based on the whole face or the eye region. Based on the literature highlighting the importance of the eye region for face perception and recognition (see, Nelson & Mondloch, 2014; Peterson et al., 2008; Royer et al., 2018), it was anticipated that composites would be rated as more

accurate based on the eye region. However, the results demonstrated that there was little difference in likeness ratings between composites rated based on the whole face and those rated based on the eye region. One explanation for this finding may be the methodology used to rate the eye region as participants viewed the whole face image, but were asked to rate only the eye region. Inviting participants to rate only one facial region of a whole face image may have reduced the likeness rating scores because there was nothing in place to ensure that participants were only focused on the eye region, such as an eye tracker (Nishida et al., 2009).

It was also hypothesised that the number of screens used during composite construction would impact the participants' abilities to utilise each stage of composite construction. Reducing the number of screens at the beginning of composite construction decreases the interactive elements and, therefore, the complexity of the task, meaning that a participant may be better able to utilise each stage of composite construction as their working memory is not overwhelmed, which would result in weakened memory and poor decision-making ability. The results from Part 2c demonstrated that composites were more accurate at the end of composite construction, followed by after use of *Holistic Tools*, followed by the *First Generation* and then the *Second Generation*. Mean composite accuracy compared to a photograph of the target surpassed that of a *Random Face* after use of *Holistic Tools*. This finding indicates that a *Second Generation* may not be beneficial to the construction of an identifiable composite image.

Nonetheless, in some conditions (*One* and *Three Screens*), random composite faces were rated as more accurate than the final composite image. Overall, findings from Part 2c of the experiment do not support the hypothesis that composites

constructed using fewer screens are most accurate, as composites constructed using *Three Screens* were rated higher than composites constructed using *Four Screens*.

Limitations

Composites in this experiment were constructed using EvoFIT Online, which uses a self-administered interview (SAI). Research demonstrates that composites constructed using an SAI, as in the current study, are typically less accurate than those constructed using the traditional system (Martin et al., 2018). It is suggested that this may be due to the increased difficulty of the task when using an SAI, as participants must follow instructions as well as construct the face, increasing the extraneous load demand (de Jong, 2010). High extraneous load in a task indicates that the individual is investing mental resources into a process irrelevant to the outcome of the task, such as reading and understanding instructions (Taylor et al., 2022). This problem may be improved by using the EvoFIT App, whereby a practitioner or researcher provides instructions and explains each stage of the construction process to the participant. This is supported by, for example, the Engage and Explain phase of the PEACE model in investigative interviewing, whereby the interviewer uses active listening to build rapport and explains the reason for the interview (Walsh & Milne, 2008). Evidence demonstrates that explaining the interview process clearly and in detail results in improved interviewee memory and increased information gain (Clarke et al., 2011). Participants receiving relevant and easy-to-follow instructions at each step of the process do not have to seek out this information on the screen, reducing the extraneous load of the task (de Jong, 2010), hence potentially improving the accuracy of the resulting composite images.

Another aspect of the PEACE model is rapport building (Abbe & Brandon, 2013). It may also be proposed that composite accuracy decreases due to the lack of interviewer-interviewee rapport during construction using the SAI. During typical EvoFIT construction, interviewers aim to build a rapport with the witness prior to composite construction. This helps the witness to feel relaxed and incites a better memory of the target face (Nash et al., 2014). However, if the opportunity to build rapport is not available, witnesses using EvoFIT Online may feel uncomfortable, impeding their ability during the task, and even their memory of the target (Kiekhaefer et al., 2014). This could be improved by introducing a Cognitive Interview (CI) at the beginning of composite construction and ensuring that the researcher makes pleasant conversation with the participant prior to the interview and prior to composite construction. Increasing the amount of contact between the participant and researcher provides a greater opportunity to build rapport, making participants feel more comfortable and relaxed throughout the experiment. Furthermore, introducing the CI prior to composite construction provides an opportunity for the researcher to obtain a verbal description of the target, as opposed to the written description used for EvoFIT Online. A verbal description may trigger new or more detailed memories of the target face, compared to a written description (Kellogg, 2007), which may aid in the construction of a more accurate composite.

Of note, participants were not in a controlled environment when undertaking the SAI, such as a laboratory. As such, there may be an increased number of distractions during the construction process, or a different level of distraction between participants (Richardson et al., 2022). For example, a participant who has the television on in the background or is looking after children while trying to create the composite has a higher number of distractions compared to a participant who

constructs the composite image in a quiet room. Such distractions may impact the accuracy of the resulting composite, becoming a confounding factor in the research. This problem may be improved if the researcher supervises participants during composite construction in an attempt to observe or monitor the environment. A composite constructed by a participant who appears to be very distracted and is unable to create the composite in close to laboratory conditions may not be included in the experiment. Furthermore, participants may be more likely to complete the experiment in such desirable conditions if they feel that they are being observed by the researcher.

Future research

The current experiment demonstrated the ability to create an EvoFIT facial composite using *One Screen* of faces instead of *Four*. However, the low overall accuracy of composites compared to targets meant that inferential statistics could not be carried out for naming data. Consequently, the information gained about the impact of cognitive load during facial composite construction was limited.

The next experiment will replicate Experiment 1 with a critical difference: composites are constructed using the traditional EvoFIT method (Frowd et al., 2012) with the researcher via video conference. This approach will reduce the amount of extraneous cognitive load as the researcher will provide detailed instructions to the participants and will be able to answer any questions throughout the construction process (de Jong, 2010). By reducing the extraneous load of the task, participants should be able to create more accurate composite images as the load on their working memory is reduced.

This approach, whereby the researcher guides the participants through the construction procedure, is ideal for the introduction of a CI prior to composite construction and for building rapport between the researcher and participant. Intentionally building rapport with a witness to crime can increase the quality of the witness's recall by reducing inaccurate or misinformation (Kieckhaefer et al., 2014; Vallano & Compo, 2011). In the current thesis, improvement in witness recall may result in a more accurate description of the target face provided by the witness. Including inaccurate or misinformation during verbal recall of the target may distort a participant's memory of the face (Loftus & Cahill, 2007), resulting in the construction of a composite which does not resemble the target. Using these techniques to improve witness memory, Experiment 2 will further examine the likeness of EvoFIT composites created using *One, Two, Three* or *Four Screens*.

This chapter explored the impact of cognitive load on EvoFIT composite construction by analysing composite naming and likeness ratings of facial composite images created using EvoFIT Online. Facial composite images were constructed using *One*, *Two*, *Three* or Four *Screens* using the self-administered system, EvoFIT Online. These composite images were named by participants familiar with the target identities or rated for likeness by participants unfamiliar with the identities. Overall, composite naming in this experiment was very low, demonstrating that composite construction using EvoFIT Online without any assistance from a researcher does not create a recognisable composite image. However, the pattern of results did demonstrate that composite naming and likeness rating of the composite *Internal Features* and *Whole Faces* was highest for composites constructed using *One Screen*, likeness ratings of composites constructed using *One Screen* did not increase during the construction procedure past use of *Holistic Tools*. As a result, likeness ratings of composites constructed using *One Screen* were not the highest at the *Final Image*. The next chapter will replicate the experiment conducted in Chapter *Three* with one crucial difference; composites will be constructed using the EvoFIT App as opposed to EvoFIT Online. As composites constructed using the face-to-face EvoFIT system compared to EvoFIT Online

(see, Fodarella, 2020; Giannou et al., 2021; Martin et al., 2018), it is expected that overall composite likeness will increase.

4

EXPERIMENT 2

REDUCING THE POPULATION SIZE DURING TRADITIONAL EVOFIT CONSTRUCTION

Abstract

This experiment aimed to further explore the impact of cognitive load on witness memory to determine the optimum number of screens used during EvoFIT composite construction. In this experiment, 40 participants created facial composites using *One, Two, Three* or *Four* face arrays to select the face images during the construction process to manipulate the number of face images viewed in hopes of reducing the intrinsic cognitive load of the task. Composite likeness was measured through composite naming by 48 participants and composite likeness ratings by 63 participants. The results demonstrated that composites constructed using *One Screen*

or *Four Screens* were the most accurate, indicating the participants benefitted from a reduced number of screens *and* having a wider selection of face images to choose from.

The first experiment demonstrated that reducing cognitive load required by the participant during the task by decreasing the number of screens viewed during EvoFIT Online composite construction is beneficial for the accuracy of the resulting composite. Specifically, composites constructed with fewer screens showed improvements in the likeness of the resulting composite images. This result has potential benefits for accurate composite construction both within police practice and for developing the theoretical understanding of the impact that cognitive load has on witnesses during composite construction. However, the composite construction process was self-administered, so it is unclear whether a similar benefit would be found when using the traditional, face-to-face approach to composite construction. In addition, implementing the EvoFIT App in the current experiment will mitigate some of the limitations identified in Experiment 1, specifically: the lack of awareness of the environment during composite construction, which may increase participant distraction; high extraneous load as a result of written instructions on screen, which may increase working memory load; and high germane load as a result of comparing new (face images on screen) and old (the target face) information, which may also increase working memory overload, potentially reducing composite accuracy.

Distraction

The lack of control over the environment in which participants created the facial composite image in Study 1 is an important limitation. During composite construction with the police, an eyewitness sits in a room with a practitioner and focuses on creating the best likeness possible. However, in the first Experiment, there was no supervision by the researcher, although she was available for questions if a participant had any. This design choice means that participants could have had the television on in the background, could have been looking after children while taking part in the experiment, or could have had long breaks during the experiment (Elliott et al., 2022). The environment during composite construction is important for recall, as demonstrated by Fodarella et al. (2021). In this experiment, participants who recalled the environment they had been in while encoding the target face during the cognitive interview (context reinstatement) or created the composite in the same environment in which they encoded the target face produced more accurate images.

Furthermore, the literature clearly demonstrates the negative effects of completing a task in a distracting environment, as opposed to a calm environment (Min, 2017; Rodrigues & Pandeirada, 2015). Thus, it is reasonable to presume that creating a facial composite in a distracting environment may have a negative impact on participant memory and so on the accuracy of the resulting composite image. Therefore, in this experiment, the researcher will be present during the entire composite construction procedure. As a consequence, the researcher is aware of the participant's environment during the construction process as the researcher can hear (or see, for participants who were comfortable with sharing their video via webcam) the participant's environment. If a participant appears to be very distracted or frequently interrupted during the construction process, their data can be discarded.

Cognitive Load

Extraneous, Germane and Intrinsic cognitive load must be considered when aiming to reduce the cognitive load of a task. In Experiment 1, it is proposed that the extraneous load was high due to the instructional information provided during the construction process (Ayres & Paas, 2007). When using the traditional EvoFIT App, a researcher, or practitioner verbally provides the instructions to the participant as needed. However, when using EvoFIT Online, as in Experiment 1, the instructions are written on screen as they become relevant. Extraneous load is increased with the use of written instructions, as individuals must invest mental resources into understanding the written instruction (de Jong, 2010). This research would suggest that understanding written instructions during EvoFIT Online is more difficult than understanding verbal instructions during face-to-face EvoFIT construction, particularly as constructors can ask for clarification if needed. A large extraneous cognitive load due to the written instructions during EvoFIT Online composite construction may overwhelm participant working memory (de Jong, 2010) and therefore impede memory capacity, resulting in less accurate composite images.

It is also proposed that Germane cognitive load was high during Experiment 1. When participants compare new information (the faces on screen) with old information (the target photograph viewed 24 hours previous), the germane load is likely to be reasonably high due to the effortful connection between new and pre-existing information (de Jong, 2010). If a participant finds it challenging to remember the target photograph, it may be fair to theorise that germane load would be increased further, as more effort is needed to compare the new information with this pre-existing information.

The interview procedure prior to composite construction was self-administered and participants were asked to write a description of the target. The aim of this self-administered CI was to aid participants in actively remembering the target face and, in doing so, trigger memories of details about the face, so that the memory of the target is better after the interview than it was before (Martin et al., 2018). However, as the researcher did not supervise participants during Experiment 1, including during the self-administered cognitive interview, it is impossible to know if this step was completed by all participants. If participants did not complete the cognitive interview, their memory of the target face would likely be lower than expected (Wells et al., 2006), and the process of comparing the facial images on screen with the target face may be more effortful, further increasing the task's difficulty.

Furthermore, the self-administered interview invited participants to write down a description of the target, as opposed to providing a verbal description, as they do in a face-to-face cognitive interview. iterature demonstrates a benefit of verbal recall over written recall for increasing the details of a memory (Kellogg, 2007). Therefore, participants receiving a face-to-face cognitive interview should have improved memory of the target compared to those who completed a self-administered interview. Theoretically, this improved memory should make the comparison of the face images on screen to the target easier, reducing germane load during the construction procedure.

This thesis aims to manipulate the intrinsic cognitive load during composite construction and assess its impact on composite accuracy. High intrinsic cognitive load is related to a task having a large number of interacting elements. During EvoFIT composite construction, the relevant interacting elements are the face images that are viewed and compared during the face selection stage of composite construction. In

this thesis, the number of face images that are viewed is manipulated by reducing the number of face arrays viewed by participants during the face selection stage of the construction procedure. Importantly, overwhelming the working memory due to high extraneous or germane cognitive load results in the same impairments as high intrinsic cognitive load (Louw, 2021). Therefore, the extraneous cognitive load must be reduced in future experiments in order to not confound the results and allow accurate measurement of the impact of intrinsic load during the construction procedure.

Cognitive Interview

The CI is a questioning technique used by the police to obtain an accurate, detailed and thorough description of a crime from an eyewitness (i.e., a witness to or a victim of a crime). The original technique (Geieselman et al.,1986) consists of four stages: reinstating the context of the crime, recalling the events in reverse order, reporting everything the eyewitness can remember, and describing the events from another person's point of view. The CI is rooted in two theories, the encoding specificity principle and the multi-component view of memory.

The encoding specificity principle states that mental reinstatement of environmental or personal contexts is beneficial for recall (Tulving & Thompson, 1973). Ergo, recalling the environment one was in, or recalling the emotions one felt during the event, may trigger information or details that were not previously recollected (Fodarella et al., 2021). The multi-component view of memory states that working memory is functionally separated from long-term memory (Baddeley, 1998) and that there are several techniques to retrieve information about the event from long-term memory, including in-depth recall of the event, describing the event in a different order to how it happened, and describing the event from the perspective of a

different person, such as a bystander. These techniques are designed to trigger information or details about the event of the crime which the eyewitness had not previously recollected.

A CI is vital for gaining accurate details about a crime which has taken place (Geiselman & Fisher, 1988). However, as composite construction focuses on the perpetrator of crime as opposed to the event itself, the CI conducted directly before composite construction focuses on the appearance of the perpetrator (Frowd et al., 2008) with a focus on recalling details of the perpetrator's appearance in hopes that this will trigger a more detailed memory. The CI conducted before EvoFIT facial composite construction mostly focuses on just one stage of the CI designed by Geieselman et al. (1986): reporting everything the eyewitness can remember. Each eyewitness is invited to picture the target face in their head in as much detail as possible and to describe the face in detail, uninterrupted, using free recall.

It is sensible to theorise that eyewitnesses who have a more detailed memory of a perpetrator will create a better facial composite image of said perpetrator due to their potential ability to add detailed, unique details to the face image, which makes them more identifiable. This is supported by research which manipulated the time between a participant viewing a target photograph and creating a facial composite image. The Decay Theory (Berman et al., 2009; Ricker et al., 2016), for example, states that memory fades over time. Hence, participants who create a facial composite after a long delay should have a less accurate memory of the target compared to participants who create a facial composite shortly after viewing the target (Frowd et al., 2005). Therefore, improving eyewitness memory of a target face before composite construction is vital for creating an accurate facial composite image.

Familiarisation

Typical naming rates for composites constructed using EvoFIT Online are fairly low, yet naming rates for composites constructed in Experiment 1 were lower than this standard, with composites in one condition named incorrectly by every participant. Such low naming rates indicated that composite images were not accurate representations of the targets. One reason for this may have been that composite construction did not replicate that of previous experiments or the procedure typically carried out by the police. As composites were constructed using EvoFIT Online without supervision by the researcher, it is difficult to know which aspects of composite construction resulted in low accuracy. As discussed, potential explanations include participants not completing the self-assessment CI or participants being distracted or taking breaks during composite construction. Therefore, in the current experiment, the interviewer personally conducted the CI, as opposed to a self-administered interview, and was also present during composite construction so that they would be made aware of any distractions during the construction process.

Alternatively, the low naming rates in Experiment 1 may be explained by a lack of familiarity with the target identities. As per the *a priori* rule, all participants were familiar enough with the targets to name 80% of them based on the target photographs; therefore, this explanation seems unlikely. It is far easier to recognise a somewhat familiar face based on a photograph of the individual than it is to recognise them from an imperfect face image, such as a facial composite. Thus, increasing participants' familiarity with the target pool before they attempt to name the composite images may increase correct composite naming. One method to increase familiarity with the targets is through a familiarisation task, whereby participants are

invited to think about individuals who may be in the target pool (in this case, England Footballers) for one minute before they view any composite images.

This method is similar to describing a target's face before composite construction in a CI. Both tasks are based on the multi-component view of memory, which states that there is a functional difference between working memory and long-term memory (Baddeley, 1998). Just as describing a target's face is designed to trigger more memories about the target's appearance, recalling individuals in the target pool should trigger the memories of other individuals in the same target pool. In the current experiment, participants were asked to think about England International Footballers, which should bring to mind footballers who are most easily remembered, and then trigger the names of footballers who are less memorable but are also in the target pool. Although the memorable and less-memorable footballers will be different for each participant, the ultimate aim of this task is to evoke the memory of potential targets in the experiment, increasing the likelihood of attaining a correct name based on the composite image. Overall, it is estimated that participants who complete this familiarisation task will be more likely to identify the targets based on the composite images (cf. participants who do not complete the task).

Current Experiment

In addition to addressing the limitations outlined above, Experiment 2 will further explore the impact that reducing the number of screens during composite construction has on the accuracy of facial composite images, including the ability of participants to utilise each stage of composite construction. Despite the difference between the construction methods used in Experiments 1 and 2, many aspects of composite construction remain the same, including the selection of composite *Internal Features*

and *External Features* as well as the stages of composite construction. Specifically, the experimental hypothesises are:

H1: Composites constructed using fewer screens will be more accurate and therefore be identified more frequently and receive higher likeness ratings than composites constructed using the typical *Four Screens*.

H2: Cognitive load will impact the ability to utilise each stage of composite construction; reducing the cognitive load in the early stages of composite construction will allow participants to enhance the face more accurately in the later stages.

Method

Part 1- Composite Construction

Design. A between-subjects design was utilised in which the IV was the number of screens shown during composite construction (referred to as *Screens*) with four levels: *One Screen, Two Screens, Three Screens* and *Four Screens.* The procedure used during composite construction reflected that used most commonly by the police, including a 24-hour delay between the participant viewing the target photograph and subsequent composite construction. In order to participate, participants were required to use a PC and have access to a video conferencing platform, such as Skype or Microsoft Teams.

Participants. Participants were 40 adults (10 males, 30 females) aged between 18 and 61 years (M = 24.8, SD = 9.01). Participants were recruited on the basis that they were not familiar with the population from which targets were selected.

Materials. The target stimuli were the same 10 photographs of England Football players used in Experiment 1. Each image was displayed in colour and was approximately 8cm (wide) x 10cm (high).

Procedure. Part 1 of the experiment occurred over two days. On the first day, participants received a password-protected Word document containing a photograph of the target as well as the passcode to unlock the document during a video call and were instructed to view the image for 30 seconds, timed by the researcher. A password-protected Word document was used so that the researcher could send the target photographs to each participant without viewing the target image herself. It was important that the researcher did not know the target identities before facial composite construction was completed with all 40 participants so that she could not subconsciously aid the participant.

Twenty-four hours after viewing the target image, participants took part in a second video call with the researcher. First, a short CI was conducted, during which the participant was asked to freely recall the target's face in as much detail as they could. Once a detailed description of the face had been obtained and recorded by the researcher using a verbal recall sheet (Appendix 6), the researcher's screen was shared so that the participant could view the screen while the researcher had control of the mouse. EvoFIT composite construction took place as described in Chapter 2, pages 64-73.

Part 2a - Composite Naming

Design. Experiment 2 utilised a between-subjects design with two IVs: *Screens* with four levels: *One, Two, Three* or *Four Screens*, and a familiarisation task before the composite naming task (referred to as *Familiarisation)* with two levels: *Familiarisation* or *No Familiarisation.* The DV was correct composite naming: Yes/No.

Participants. Participants were 48 adults (29 males, 19 females) between the ages of 18 and 67 years (M = 33.55, SD = 13.52). Participants were recruited on the basis that they were familiar with the target population. One participant failed the *a priori* rule of familiarity and so their data were not included, and a new participant was recruited in the same condition. An equal number of participants were randomly allocated to view a composite construction from each level of *Screens* (N = 12).

Materials. As in Experiment 1, composites constructed during Part 1 of the experiment were displayed in four PowerPoint presentations (one for each level of *Screens*). A fifth presentation contained the target photographs. This was shown to all participants to check familiarity with the targets.

Procedure. The procedure replicated that of Experiment 1, Part 2a but with one difference. Before attempting to name any of the composites, participants in the *Familiarisation* condition were asked to think about England International footballers for one minute in an attempt to refamiliarize themselves with the target pool.

Part 2b - Composite Final Image Rating

Design. Part 2b of the experiment utilised a mixed design with two IVs. *Screens* with four levels*: One, Two, Three and Four Screens,* was a within-subjects variable, and the rating task (referred to as *Task)* with three levels: *Internal Features, External Features* and *Whole Faces*, was a between-subjects variable.

Participants. Participants were 45 adults (16 males, 21 females) aged between 18 and 58 years (M = 26.98, SD = 9.71). Participants were recruited on the basis that they were not familiar with the target population. An equal number of participants were randomly allocated to each level of *Task* (N = 15).

Materials. As in Experiment 1, the materials were the *Internal Features*, *External Features* and *Whole Composites* that were constructed by the first group of participants during Part 1. Three PowerPoint presentations were created containing images of either the *Internal Features*, *External Features* or complete faces.

Procedure. Participants viewed one of the three PowerPoint presentations via the 'screen share' feature on Skype and verbally rated each composite image using a seven-point Likert scale. Participants were instructed to inform the researcher if they recognised any of the target images, in which case, their data would have been removed from the experiment, and a new participant recruited in the same condition. However, no participants stated that they recognised any of the targets.

Part 2c - Intermediate Composite Rating

Design. Experiment 2 utilised a within-subjects design with two IVs: *Screens* with four levels: *One, Two, Three* or *Four Screens*, and the stage of composite construction (referred to as *Stages*) with five levels: *Random Face, First Generation, Second Generation, After Holistic Tools* and *Final Image*.

Participants. Participants were 18 adults (9 males, 9 females) aged between 18 and 52 years (M = 22.67, SD = 8.33). Participants were recruited on the basis that they were not familiar with the target; the same *a priori* rule as Experiment 1 was implemented.

Materials. As in Experiment 1, images of the composite at each level of *Stages* were displayed in a PowerPoint presentation alongside a random composite image to provide a baseline rating as well as a photograph of the target.

Procedure. The procedure was the same as Part 2- participants viewed the PowerPoint presentation via the 'screen share' feature on Skype and verbally rated each composite image using the seven-point Likert scale.

Results

Part 2a - Composite Naming

To test the hypothesis that composites constructed using fewer screens will be more accurate, responses to facial composites and target pictures were scored for accuracy as in Experiment 1. As per the *a priori* rule, data were only included from participants who correctly named at least eight of the 10 target pictures. Responses were coded as correct and assigned a value of 1 when participants gave the correct name for the target image and composite image. Responses were coded as incorrect and assigned a value of 0 when a wrong name or "don't know" response was given for the composite image, but the target was identified correctly. Responses were assigned a value of 2 when the target was not identified, and these data were not included in the analysis. Incorrect names for target pictures occurred 30 times (by 21 participants), three to eight times by group (number of *Screens*). As such, the mean correct naming for target pictures was high (M = 94.00%, SD = 24.00%).

As expected, correct responses were much lower overall for spontaneous naming of facial composites (M = 12.44%, SD = 33.05%), and correct naming was fairly low overall. Table 5 displays the mean naming at each level of *Screens* (*One, Two, Three* or *Four Screens*) and each level of *Familiarisation* (*Familiarisation* and *No Familiarisation*).

Table 5. Spontaneous Naming of Composites for Each Level of Screens and Familiarisation

Familiarisation Task	Number of Screens				Mean
	4	3	2	1	
No Familiarisation	18.87 (10 / 53)	7.02 (4 / 57)	8.62 (5 / 58)	14.29 (8 / 56)	12.05 (27 / 224)
Familiarisation	20.00 (11 / 53)	8.62 (5 / 57)	9.09 (5 / 58)	13.79 (8 / 56)	12.83 (29 / 224)
Mean	19.44 (21 / 108)	7.83 [a] (9 / 115)	8.85 [a] (10 / 113)	14.04 (16 / 114)	12.44 (56/ 450)

Note. Figures are expressed in percentage and calculated from participant responses in parentheses: summed correct responses (numerator) and total (correct and incorrect) responses (denominator). Data are presented for composites for which participants correctly named the relevant target photographs (N = 450 out of 480). [a] $p < .05$.

Table 5 shows that composites constructed using *Four Screens* were the most accurate, followed by those constructed using *One Screen, Two Screens* and *Three Screens*. The table also shows that correct naming rates were slightly higher after participants had one minute to think about England International Footballers before starting to name the composites. Individual correct naming responses from participants were analysed using GLMM. This stage of the experiment involved two predictors, *Screens* (coded as *1 = One Screen*; *2 = Two Screens*; *3 = Three Screens*; *4 = Four Screens*) and *Familiarisation* (coded as *0* = no familiarisation and *1* = familiarisation). The DV was correct composite naming, with the responses to composites only included in the analysis if the corresponding target was named correctly.

A hypothesis-testing approach comprised three models, one model for *Screens,* a second model for *Familiarisation,* and a third full-factorial model containing *Screens* and *Familiarisation* and the interaction between them. The model for *Screens* was significant [F (3, 446) = 3.14, p = .025]. However, the model for *Familiarisation* was not significant [F (1, 448) = 0.17, p = .69], nor was the *Screens*Familiarisation* interaction [F (3, 442) = 0.03, p = .99]. Therefore, the best-fit model was for *Screens* only. Coefficients in relation to the baseline condition of *Four Screens* were examined to explore the significant result, displayed in Table 6.

Table 6. Model Parameters from Correct Naming of Composites at Each Level of Screens. Comparisons Are Presented with Reference to The Lowest Category- Four Screens; Negative Values of B Indicate Lower Ratings of Likeness with Respect to The Reference.

Screens	*B*	*SE(B)*	*t (446)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
3 Screens vs. 4 Screens	-1.21	0.47	-2.59	.010	0.30	0.12	0.75
2 Screens vs. 4 Screens	-1.09	0.45	-2.39	.017	0.34	0.14	0.83
1 Screen vs. 4 Screens	-0.40	0.41	-0.97	.33	0.67	0.30	1.50

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix)] final, Model-based full factorial Corrected model [F (3, 446) = 3.14, p = .025]. The model was specified with the lowest category of categorical predictors as reference (*4 Screens*), and predictors were sorted in a descending order. Information criteria are based on -2 log likelihood (*AICC* = 2519.36, *BIC* =

2523.45). Variance of random effects' intercept for items [2.25, *SE* = 1.32, *Z* = 1.70, *p* = .09, *CI* (0.71, 7.13)].

Results demonstrated that composites constructed using *Two* or *Three Screens* were significantly less accurate than composites constructed using *Four Screens*. In contrast, *One Screen* and *Four Screens* did not differ significantly in terms of composite accuracy.

To understand the mathematical pattern of the data, polynomial contrasts were simulated using GLMM in SPSS, focusing on linear and quadratic contrasts. The analysis demonstrated that there was no significant linear (*p* = .22, *Exp(B)* = 1.16) or cubic (p = .52, *Exp(B)* = 1.17) pattern in the data, but there was a significant quadratic contrast (*p* = .011, *Exp(B)* = 1.59). The presence of such significant patterns was due to composite likeness ratings, which decreased from *One Screen* to *Two Screens,* remained relatively stable from *Two Screens* to *Three Screens,* and then increase from *Three Screens to Four Screens.*

Part 2b - Composite Face Image Rating

Part 2b tested the hypothesis that composites constructed using fewer screens would be more accurate and aimed to understand the impact that using fewer screens has on likeness of composite *Internal* and *External Features*. To test this hypothesis, composite images were rated for accuracy in comparison to the target photograph using a Likert scale of 1-7 based on the *Internal Features*, *External Features* and whole face images. Table 7 presents the mean and standard error for rating of composites constructed at each level of *Screens (One, Two, Three* or *Four Screens).*

Table 7. Mean (and Standard Error) for Rating of Composite Internal Features, External Features and Whole Composites Constructed Using One, Two, Three or Four Screens

	4 Screens	3 Screens	2 Screens	1 Screen	Mean
Internal	2.72	2.96	2.77	3.34	2.95
Features [ab]	(0.10)	(0.11)	(0.10)	(0.13)	(0.06)
External	3.50	3.20	3.26	3.83	3.45
Features[1]	(0.11)	(0.11)	(0.12)	(0.15)	(0.06)
Whole	2.07	2.15	2.17	2.45	2.21
Composites [ab]	(0.10)	(0.11)	(0.11)	(0.13)	(0.06)
Mean	2.77	2.77	2.73	3.21 [a]	2.87
	(0.07)	(0.07)	(0.07)	(0.08)	(0.04)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant-item SE of the mean. [a]$p < .05$, [b]$p < .001$

Table 7 shows that composites constructed using *One Screen* were rated as the most accurate, followed by *Three* and *Four Screens* simultaneously, and then *Two Screens*. Table 7 also shows that *External Features* of composites were rated as the most accurate compared to the target photograph, followed by *Internal Features* and *Whole Composites*. Individual ratings of composite items from participants were analysed using GLMM. This stage of the experiment involved two predictors, *Screens* (coded as *1 = One Screen*; *2 = Two Screens*; *3 = Three Screens*; *4 = Four Screens*) and *Task* (coded as *0 = Whole Face*, *1 = Internal Features* and *2 = External Features*). The DV was rating of composite accuracy compared to the target photograph on a Likert scale of 1-7. However, since participants gave few ratings of 7, these scores were re-coded to have a value of 6 (i.e., categories of 6 and 7 were merged together).

A hypothesis-testing approach comprised three models, one model for each predictor: *Screens* and *Task,* as well as a full-factorial model, which contained both predictors and the interaction between them. The model for *Screens* only was significant [F (3, 1792) = 10.25, $p < .001$], as was the model for *Task* only [F (2, 1793) = 11.21, $p < .001$]. However, the interaction between *Screens* and *Task* was not significant [F (6, 1793) = 0.99, $p = .43$]. A fourth model was run, containing *Screens* and *Task* without the interaction. As this model was significant for both predictors, it was taken as the final model. Fixed coefficients were examined to explore the significant results for *Screens* and *Task*, displayed in Table 8.

Table 8. Model Parameters from Composite Rating at Each Level of Task. Comparisons Are Presented with Reference to The Lowest Category (underlined). Negative Values of B Indicate Lower Ratings of Likeness with Respect to The Reference.

	B	*SE(B)*	*t (1784)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
Screens							
3 Screens vs 4 Screens	0.55	0.24	2.26	.024	1.73	1.08	2.79
2 Screens vs 4 Screens	0.17	0.25	0.67	.503	1.18	0.73	1.91
1 Screen vs 4 Screens	0.10	0.25	0.39	.696	1.10	0.68	1.79
Task							
Internal vs Whole Face	2.32	0.49	4.77	<.001	10.13	3.91	26.24
External vs Whole Face	1.25	0.48	2.58	.010	3.49	1.35	9.01
Internal vs External	-0.81	0.48	-1.69	.09	0.45	0.18	1.14

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix X)] final, Model-based full factorial Corrected model [F (11, 1784) = 5.32, p < .001]. The model was specified with the lowest category of categorical predictors as reference (*Four* Screens and *Whole Face*) with predictors sorted in an ascending order. Information criteria are based on -2 log likelihood (*AICC* = 31734.12, *BIC* = 31745.08). Variance of random effects' intercept of participants for *Screens* [1.32, *SE* = 0.32, Z = 4.15, p < .001, *CI* (0.83, 2.12)].

These results demonstrated that composites constructed using *One Screen* were rated as significantly more accurate than composites constructed using *Four Screens.* Although not significant, composites constructed using *Two Screens* or *Three Screens* were also rated as marginally less accurate than those constructed using *Four Screens.* These results also demonstrated that composite *Internal Features* and *External Features* were rated as being more accurate than *Whole Faces* composites. There was also a difference between composites rated based on *Internal Features* and *External Features* which approached significance, with those rated on *Internal Features* being less accurate.

To understand the mathematical pattern of the data, polynomial contrasts were simulated using GLMM in SPSS. Based on the hypothesis that composites constructed using fewer screens are more accurate, a linear contrast was predicted. However, the analysis demonstrates a non-significant linear pattern in the data (p = .13, *Exp(B)* = 0.84), a significant quadratic pattern ($p < .001$, *Exp(B)* = 1.17) and a non-significant cubic contrast (p = .62, *Exp(B)* = 0.89). The significant quadratic pattern indicated that the pattern of results fit a quadratic trend whereby the likeness decreased from *One Screen* to *Two Screens* and then increased from *Two Screens* to *Three* and *Four Screens.*

Part 2c - Intermediate Composite Rating

To continue exploring the impact of cognitive load on composite accuracy, composites at four different stages of composite construction (*First Generation, Second Generation, After Holistic Tools, Final Image)* and a *Random Face* for comparison were rated for likeness against the corresponding target face. Part 2c also

tested the hypothesis that reducing the number of screens during composite construction was beneficial for composite accuracy, and that manipulating the number of screens impacted the ability to utilise each stage of construction (shown by reduced composite accuracy in the final stages of construction for composites constructed using more screens).

Ratings of composite images at four stages of composite construction *(First Generation Best Face, Second Generation Best Face, After Holistic Tools, Final Image),* as well as a *Random Face,* were analysed, and means and standard errors of composites created using *One, Two, Three* or *Four Screens* of faces during the face selection stages of construction are presented in Table 10.

Table 10. Mean (and Standard Error) for Composite rating at Four Stages of Construction and Each Level of Screens.

Stage of Construction	Screens				
	Four Screens	Three Screens	Two Screens	One Screen	Mean
Random	1.80	1.83	1.81	2.22	1.92
	(0.07)	(0.07)	(0.07)	(0.08)	(0.04)
First	2.36	2.40	2.53	2.61	2.47
Generation	(0.08)	(0.07)	(0.08)	(0.09)	(0.04)
Second	2.70	2.65	2.86	2.97	2.79
Generation	(0.09)	(0.08)	(0.09)	(0.08)	(0.04)
Holistic Tools	3.05	3.11	2.85	3.32	3.08
	(0.09)	(0.09)	(0.10)	(0.10)	(0.05)
Final Image	3.16	3.15	2.80	3.50	3.15
	(0.10)	(0.10)	(0.10)	(0.11)	(0.05)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean (which gave a clearer pattern of results cf. median) and, in parentheses, by-participant SE of the mean. [a] $p = .001$

Table 10 shows that composite images increased in accuracy throughout the construction process which suggested that composites constructed using *One Screen* were the most accurate, as they were rated highest at the *Final Image*, followed by those constructed using *Four, Three* or *Two Screens*. To further understand the impact that *Screens* and *Stage* have on composite accuracy, the ratings were analysed using GLMM. This stage of the experiment comprised two predictors: *Screens* (coded as *1 = One Screen*; *2 = Two Screens*; *3 = Three Screens 4 = Four Screens*) and *Stage* (coded as *0 = Random Face*, *1 = First Generation*, *2 Second Generation*, *3 = After Holistic Tools* and *4 = Final Image*). The DV was rating of composite accuracy compared to the target photograph on a Likert scale of 1-7. However, since participants gave few ratings of 7, these scores were re-coded to have a value of 6 (i.e., categories of 6 and 7 were merged together).

A hypothesis-testing approach comprised three models, one model for *Screens* and a second model for *Stage*. The third, full-factorial model contained the two predictors as well as the interaction between them. The model for *Screens* only was significant [F (3, 4776) = 11.52, p <.001], as was the model for *Stage* only [F (4, 4776) = 71.69, p <.001] and the interaction between *Screens* and *Stage* [F (12, 4776) = 2.69, p = .001). Therefore, the final model contained *Screens* and *Stage* as well as the interaction between them.

To explore the interaction between *Screens* and *Stage*, four models were run to investigate *Stage* at each level of *Screens*. The analysis demonstrated a significant effect of *Stage* at every level of *Screens: One Screen* [F (4, 1191) = 6.21, p < .001],

Two Screens [F (4, 1991) = 6.44, $p < .001$], *Three Screens* [F (4, 1191) = 39.24, $p < .001$] and *Four Screens* [F (4, 1191) = 8.73, $p < .001$].

To understand the mathematical pattern of the composite accuracy between each level of *Screens*, polynomial contrasts were simulated using GLMM in SPSS, with a focus on linear and quadratic contrasts. The analysis demonstrated significant linear (p = .055, *Exp(B)* = 1.12) and quadratic (p <.001, *Exp(B)* = 0.89) contrasts in the data. However, when a significant cubic contrast (p = .002, *Exp(B)* = 1.13) was added to the model, the linear contrast was no longer significant (p = .18, *Exp(B)* = 0.88). The presence of such significant patterns was due to composite likeness ratings, which decreased from *One Screen* to *Four Screens,* remain relatively stable from *Four Screens* to *Three Screens,* and then decreased from *Three Screens to Two Screens.*

Polynomial contrasts were also run to understand the mathematical pattern of composite accuracy between each level of *Stage*. The analysis demonstrated significant linear (p <.001, *Exp(B)* = 1.58) and quadratic (p <.001, *Exp(B)* = 0.88) contrasts in the data and confirmed the absence of a cubic contrast (p = .30, *Exp(B)* = 1.02). In this final model, including linear, quadratic and cubic contrasts, linear ($p < .001$, *Exp(B)* = 1.50) and quadratic ($p < .001$, *Exp(B)* = 0.85) patterns were significant. This pattern of results indicated a monotonic increase in composite accuracy, whereby the likeness ratings increased at varying amounts between each level of *Stage.*

The interaction between *Screens* and *Stage* indicated that, the amount by which a composite increased in accuracy throughout the various stages of composite construction, is somewhat dependent on the number of screens used during composite construction. In this experiment, composites constructed using *Two Screens* did not increase in accuracy after the *Second Generation*. Composites constructed using *One*

Screen increased in accuracy by 0.28 between the final two stages, whereas composites constructed using *Four Screens* increased in accuracy by 0.11 and those constructed using *Three Screens* only increased by 0.04 between the same two stages. This pattern of results may suggest that participants who construct a facial composite using *One Screen* utilise each stage of construction more effectively than participants who construct a composite using *Two, Three* or *Four Screens.* Therefore, *One Screen* may be optimal for best practise during EvoFIT composite construction.

Discussion

This experiment was designed to develop an understanding of the impact of cognitive load on EvoFIT construction and determine the optimum population size during facial composite construction to better inform best practise. To replicate Experiment 1, cognitive load was manipulated by altering the number of face arrays viewed by participants when selecting the face shape and texture in the early stages of EvoFIT construction. It was hypothesised that facial composites constructed using fewer screens during this process would be the most accurate as the likelihood of participants experiencing cognitive overload is decreased.

This hypothesis was not strictly supported by the results of this experiment. Composite naming demonstrated that composites constructed using *Four Screens* were named correctly the most frequently, indicating that composites in this condition most resemble the target. Nonetheless, composites constructed using *One Screen* were also named frequently, and named more frequently than those constructed using *Two* or *Three Screens*, implying that composites in this condition also somewhat resemble the target. Furthermore, composite rating demonstrated that composites constructed using fewer screens (either *One* or *Two Screens*) are rated the highest, indicating that

composites in these two conditions most resemble the target. This finding somewhat fits with theories outlining a benefit of reducing cognitive load, as outlined below.

Composites constructed using *One Screen* in this experiment had the lowest number of interacting elements (i.e., face images which must be compared) and it is fair to theorise that composites constructed in this condition were done so with the lowest intrinsic cognitive load (Sweller, 1988). Of the three measures of composite accuracy (composite naming, composite face rating and intermediate composite face rating) composites constructed using *One Screen* were named correctly the second most frequently, rated as the most accurate in one measure and as the second most accurate in the second rating measure. Composites constructed in any other condition did not consistently perform as well as those constructed using *One Screen*, indicating that reducing the intrinsic cognitive load during composite construction is somewhat beneficial for composite accuracy. However, composite naming is arguably the most important measure of composite accuracy as it most closely reflects the identification of criminal perpetrators based on composite images released to the public (Martin, Hancock & Frowd, 2017). In this measure, composites constructed using *Four Screens* were named most frequently. This result indicates that composite construction does benefit from a reduced cognitive load during the construction process but may also benefit from the opportunity to view a wide selection of faces images during the process of composite construction.

One explanation for the pattern of results whereby composites constructed using *One* or *Four Screens* are the most accurate is that composites constructed with the lowest cognitive load benefitted from the reduction in load, reducing the impairment in memory capacity and decision making that may occur as a result of high cognitive load (Sweller, 1988). However, composite construction may have also

benefitted from participants having the opportunity to select face images from a wider array of options. It may therefore be important to find a balance between reducing the cognitive load and reducing the number of face options participants are able to view during the composite construction process.

In Part 2a of this experiment, composite naming was lower than expected, based on EvoFIT literature (see, Erikson et al., 2022; Fodarella et al., 2021; Giannou et al., 2021). This result indicates either that participants naming the composites were not familiarised with the target pool, despite meeting the *a priori* rule, or that composite construction did not replicate that of previous research.

Some of these explanations were tested in the current experiment. For example, half of the participants were invited to complete a familiarisation task in which they had one minute to think about England International Footballers (i.e., the target pool). This is not a standard procedure during experimentation involving facial composite identification; however, it is a sensible method based on the theory of working memory (Baddeley, 1998), aiming to rehearse existing information, bringing it from the long-term memory store into the working memory. It was predicted that participants who complete this familiarisation task would remember more individuals from the target pool, and therefore be more likely to recognise and name the identities based on their composite image in the naming task (Bartsch et al., 2018). In fact, this task made little difference to composite naming rates, suggesting that familiarisation with the target pool was not an explanation for low correct naming.

Composite construction in this experiment took place online, with the participant and researcher interacting via video conferencing. Although this procedure does not replicate that carried out in previous experiments and composite construction by the police, care was taken to ensure that all aspects of composite construction were

carried out correctly. Therefore, the results from this experiment can be generalised to composite construction with the police, making it is unlikely that the low naming in this experiment is a result of composite construction not replicating previous research. However, future research may replicate the current experiment using EvoFIT face-to-face. Given this, it is difficult to understand why composite naming rates in the current experiment were lower than those typically reported in past research (Frowd et al., 2019).

Despite the low overall composite accuracy, composite likeness ratings in comparison to the target face demonstrated that composites constructed using *One Screen* were the most accurate, followed by those constructed using *Three* and *Four Screens* simultaneously, and then *Two Screens.* This pattern of results does not support that of Part 2a, as composite naming demonstrated that composites constructed using *Four Screens* were the most accurate. However, in Parts 2a and 2b, composites constructed using *One* or *Four Screens* were either the most accurate or the (joint) second most accurate. Hence, it is difficult to clearly state whether the pattern of results in these two measures support the hypothesis that composites constructed using fewer screens are most accurate.

A second aim of this part of the experiment was to ensure that reducing the number of face options by reducing the number of screens during the construction process did not have a negative impact on the accuracy of identification of composite *Internal Features*. As participants view only the *Internal Features* during the face selection stage of composite construction, viewing fewer options to select this face region may result in *Internal Features* which are vastly reduced in accuracy. To test this, participants viewed either a whole composite face image, only the composite *Internal Features* or only the composite *External Features* and were asked to rate

how alike the target photograph was compared to the face region viewed. The results demonstrated that *External Features* of composite images were deemed to be the most accurate based on likeness ratings. Previous research has found that *External Features* are more important than *Internal Features* for recognising unfamiliar faces (Bruce & Young, 1998; O'Donnell & Bruce, 2001). During the construction process, the target face remains unfamiliar to the participant creating the composite image. As the *External Features* are most important for unfamiliar face recognition, participants may be better at accurately selecting the *External Features* than they are at selecting the *Internal Features*, resulting in more accurate *External Features*, supporting the finding in this part of the experiment.

Furthermore, composite *Internal Features* were rated higher than whole composite images at all levels of *Screens.* This pattern of results suggests that reducing the number of face options during the composite construction process does not have a negative impact on the accuracy of composite *Internal Features*. This result, whereby *Internal Features* were identified as being more accurate than *Whole Faces* does not support literature that faces are recognised more easily as a whole than by isolated parts (Tanaka & Farrah, 2007). One reason for this may be that past research investigating the importance of featural regions for face recognition typically use perfect face images, such as photographs. However, facial composite images are imperfect representations of a face. When a facial composite is viewed as a part (only the *Internal Features* or *External Features*) as opposed to a whole face containing the *Internal* and *External Features* together, the imperfections in the face may be less obvious, resulting in higher ratings of likeness.

The second hypothesis in this experiment was that the number of screens used during composite construction would impact the participants' abilities to utilise each

stage of composite construction. It was theorised that decreasing the number of screens at the beginning of composite construction reduces the complexity of the task and the likelihood of cognitive overload. If this theory is supported, it may be fair to suggest that participants who create a composite using fewer screens are better able to utilise each stage of construction. This theory provides an explanation for why composites constructed using fewer screens of faces are more accurate, as improvements to the composite image continue throughout the construction process, which may not occur when composites are created using more screens.

The results of this experiment loosely support the hypothesis. Composites constructed using *One, Three* and *Four Screens* increase in accuracy between each stage of composite construction. However, the increase in composite accuracy between the final two stages (*Holistic Tools* and the *Final Image*) was larger for composites constructed using *One Screen* than it was for composites constructed using *Three* and *Four Screens.* Such a result indicates that reducing the cognitive load during composite construction does allow participants to better utilise the stages of composite construction.

Nonetheless, composites constructed using *Two Screens* did not increase in accuracy between the *Second Generation* and after use of *Holistic Tools* or between *Holistic Tools* and the *Final Image.* The reasoning for the inability to utilise the image enhancement tools during composite construction using *Two Screens* is unlikely due to a cognitive overload, as composites constructed using more screens, and therefore with a higher cognitive load, demonstrated an increase in accuracy in the late stages of composite construction. Thus, more research is needed to understand the pattern of composite accuracy during the stages of composite construction to assess whether the inability to utilise the image enhancement tools during composite construction using

Two Screens is isolated to this experiment, or whether this is a pattern which requires a deeper understanding.

Limitations and Future Research

Facial composites were constructed using the EvoFIT App, which resulted in more accurate composite images than those produced using EvoFIT Online in Experiment 1. However, although composite accuracy increased in this experiment compared to the first, correct naming rates for composite images constructed in all conditions were still lower than that typically found in recent research (see, Erikson et al., 2022; Fodarella et al., 2021; Frowd, 2021).

One explanation for the reduced composite accuracy may be high germane load during composite construction, resulting in cognitive overload and therefore a lower ability to create an accurate composite image. Germane load is increased when attempting to combine 'old' and 'new' information (Wood & Zivcakova, 2015), it is therefore theorised that improving a participant's memory of the 'old' information, that is, the target face, may reduce the germane cognitive load of the task. This theory is based on knowledge that an increase in germane cognitive load can be overcome when the new information can easily connect and integrate with pre-existing information from the long-term memory (Dirkx et al., 2021). Asking participants to freely recall a description of the target face before composite construction should refresh their memory, making the comparison of faces on screen to their memory of the target easier. However as composite naming was lower than anticipated, it may be theorised that recalling a description of the target did not enhance participants' memories of the face enough to ease the comparison of the target memory to the images on screen.

Recalling a description of the target face may not have been effective in enhancing participants memories because describing the target face is difficult, if not impossible, to do holistically (Nakabayashi et al., 2012). Thus, participants break the face down into individual features, describing each one separately. Once a participant has described each feature separately, they may subconsciously perceive the face as a collection of features instead of a holistic image. Perceiving the face this way may be particularly troublesome when participants are asked to select whole faces while constructing the composite image using a holistic system such as EvoFIT (see, Portch et al., 2017; Richler et al., 2011). If participants perceive the face as a collection of features before selecting whole face images during EvoFIT construction, their ability to accurately select face images may be reduced compared to a participant who perceives the face as a holistic image prior to composite construction. Going forward, a Holistic-Cognitive Interview (H-CI) should be used in place of the CI, as this interview technique encourages holistic processing of the target face.

Recent EvoFIT composite construction has utilised the H-CI, which aims to obtain a description of the target from a participant, but to also promotes holistic face processing by asking participants to make judgements about the character of the target based on their appearance, for example, how friendly they are. As such, it may be easier for a participant to accurately select the best face images during the face selection stage of composite construction after an H-CI compared to after a CI. Furthermore, the H-CI may help participants to remember the target face in more detail than through the CI alone because it involves a more in-depth analysis of the face. Hence, participants may have a better memory of the target face, making the connection of new (face images on screen) and old information (target face) less

taxing. Therefore, the germane cognitive load of the task should be reduced (Dirkx et al., 2021; Wood & Zivcakova, 2015), lessening the likelihood of cognitive overload.

Experiment 3 will replicate Experiment 2 with a crucial difference: an H-CI is used in place of the CI of the current experiment. This experiment will seek to demonstrate the impact of cognitive load on composite construction when the participant likely has a better memory of the target (Skelton et al., 2020). As the police use an H-CI prior to EvoFIT construction, Experiment 3 will more closely reflect the impact of cognitive load during the construction of a composite in a real case, which allows the results to be generalised more easily to composite construction by the police.

This chapter analysed the impact of cognitive load on EvoFIT composite construction using the EvoFIT App by analysing composite naming by participants familiar with the target identities and likeness ratings by participants unfamiliar with the target identities. Overall, composite likeness was higher than in the previous experiment; however, it still did not reach that achieved in recent EvoFIT literature (Fodarella et al., 2021; Skelton et al., 2020). The pattern of composite likeness did not replicate that of Experiment 1, or that predicted based on the literature indicating that cognitive load may have a negative impact on the ability to complete the task to a high standard (Sweller, 2010). Overall, composites constructed using *One* or *Four Screens* to create the facial composite images were the most accurate, indicating that there is a benefit of reducing the cognitive load during EvoFIT composite construction, but that there is also a benefit of viewing more face options to select the face shape and texture. It was theorised that this pattern of results is due to high germane cognitive load during the construction process, as a result of participants having a poor memory of the target face. Therefore, the next chapter will discuss Experiment 3, which attempts to improve participants' memories of the target by conducting a more detailed interview prior to the composite construction procedure.

5

EXPERIMENT 3

REDUCING THE POPULATION SIZE DURING EVOFIT CONSTRUCTION AFTER AN H-CI

Abstract

This experiment aims to understand the impact of cognitive load on witness memory to determine the optimum number of screens used during traditional EvoFIT composite construction. In this experiment, Forty participants created facial composites using *One, Two, Three* or *Four* face arrays to select the face images during the construction process to manipulate the number of face images viewed in hopes of altering the intrinsic cognitive load of the task. These composite images were judged for accuracy through composite naming by 40 participants and likeness ratings by 48 participants compared to the original target's photograph. The results demonstrated that reducing the number of screens from *Four* to *One Screen* during

the 'face selection' stage of composite construction was beneficial for composite accuracy, compared to reducing the number of screens to *Two* or *Three*.

The results of Experiment 2 demonstrated that composites constructed using *One* or *Four Screens* of faces during EvoFIT construction produced the most accurate resulting composite image. This contradictory pattern of results indicated that participants may benefit from reduced cognitive load during composite construction, but that participants may also benefit from selecting faces from a broader range of options from which to make a face selection. This experiment has potential benefits for improving composite construction procedures that are used by the police and for developing a theoretical understanding of the impact of cognitive load during composite construction. However, construction in this experiment occurred after a CI, as opposed to the more modern H-CI that is used by the police today (Fodarella et al., 2015). Therefore, it was unclear whether the findings from Experiment 2 would generalise to composite construction with the police after an H-CI. Additionally, implementing the H-CI may mitigate against some of the limitations identified in the previous experiment, specifically- the potential for high germane cognitive load and possible poor memory of the target, both factors which may have contributed to composite naming rates which were lower than expected compared to current EvoFIT research (Fodarella et al., 2021; Skelton et al., 2020).

Cognitive Load

Sweller's Theory of Cognitive load (1988) states that 'load' comes from three sources: intrinsic, extraneous and germane. This thesis aims to measure the impact of

intrinsic cognitive load on composite construction. However, as the effect of high cognitive load, that is, cognitive overload, is the same for all three types of load identified by Sweller, the extraneous and germane load must be reduced as to not confound the results. Specifically, if the extraneous and germane load is high during composite construction, changes in the likeness and naming rates may be due to extraneous and germane load, as opposed to intrinsic load, which is manipulated in this thesis. The findings from Experiment 2 indicated that germane cognitive load may have been high due to low composite naming in all four conditions. It was theorised that, if participants have a poor initial memory of the target face, the comparison between new and pre-existing information (i.e., face images on screen and the target face) may be more effortful, increasing the germane load of the task (Kalyuga, 2011). Therefore, it was important to enhance participants' memories of the target in this experiment to reduce the germane load of the task and more accurately assess the impact of intrinsic cognitive load on composite construction.

Holistic-Cognitive Interview

The reliance on a CI to obtain a description of the target in Experiment 2 may not have been adequate to properly refresh a participant's memory of the target (Frowd et al., 2008, 2015). Therefore, this experiment utilised a more in-depth interview technique, the Holistic-Cognitive Interview (H-CI). The H-CI is an adaptation of the CI, originally designed by Geiselman and colleagues (1986). Like the CI, the H-CI is a questioning technique used by the police to obtain a description of a crime from an eyewitness. However, a major difference between the two interview techniques is that an H-CI is specifically designed for use before the construction of a facial composite. The H-CI has been shown to increase the accuracy of the composites. Frowd et al.

(2008) demonstrated that facial composites constructed after an H-CI were four times more likely to be named correctly than composites constructed after a CI.

During a CI, eyewitnesses are invited to provide a detailed description of the criminal perpetrator. However, describing a face in this way, feature by feature, may encourage eyewitnesses to perceive the target face as a collection of features as opposed to a holistic image (Frowd et al., 2013). During composite construction using EvoFIT, eyewitnesses view and select whole faces. Therefore, perceiving a face as a collection of features during whole face selection may be detrimental to the accuracy of the final composite image. Consequently, the H-CI was designed to reinstate eyewitnesses' perception of the target face as a holistic image by inviting them to make character judgements about the face.

After asking eyewitnesses to freely recall a description of the target face, interviewers administering an H-CI then ask eyewitnesses to silently contemplate the personality of the perpetrator based on their appearance. The interviewer then sequentially reads a list of seven characteristics and asks the eyewitness to rate how much each characteristic resonates with the target, rating each characteristic as "low", "medium", or "high". In a real case, seven characteristics would be carefully selected from a list of 22; for example, if a perpetrator committed a violent crime, the eyewitness would not be asked how threatening or angry the perpetrator is based on their appearance. In this thesis, the following characteristics were selected: intelligent, friendly, kind, selfish, arrogant, distinctive/unusual looking and aggressive, as these characteristics have been used successfully in past research. The benefits of making character judgements about a face is not a new concept, with the value of character attribution for accurate composite construction previously evidenced in Shepherd et al. (1978), Wells and Hryciw (1984), and Davies and Oldman (1999).

The effectiveness of the H-CI for the construction of an identifiable composite image was first tested using the system PRO-fit. In this experiment, participants viewed a target and, three-to-four hours later, received either a CI or an H-CI before constructing a facial composite of the target. Composite images created after an H-CI were four times more likely to be identified compared to those constructed after a CI (Frowd et al., 2008). PRO-fit is a featural system, meaning that each feature (such as eyes, nose, mouth) is selected individually and combined to create a face image. Although the H-CI encourages holistic processing, and PRO-fit requires individual feature selection, the use of H-CI before composite construction is clearly beneficial. The effectiveness of an H-CI has also been tested for the construction of a facial composite using the holistic system, EvoFIT and, in this case, composites constructed after an H-CI were named correctly almost twice as frequently as those constructed after a CI, with composites constructed after a HI or Hair-I named less frequently than those constructed after a CI (Frowd et al., 2011). This experiment demonstrates that both parts of the H-CI (the verbal recall and the whole face judgements) are important for the creation of an identifiable likeness, as composites constructed after only the verbal recall or only the whole face judgements were named less frequently than those constructed after the full H-CI.

After a shift in construction procedure from asking eyewitnesses to focus on the whole face to asking them to focus on only the eye region, the benefits of the H-CI were reduced (Portch et al., 2017). Therefore, the H-CI was updated to include character judgements based on the whole face, and then based only on the eye region. Skelton et al. (2020) examined the likeness of composites created with focus on the eye region during the face selection stages of construction, as they are in this thesis, and an H-CI including whole face and eye region judgements (as they are in the

current experiment). Participants viewed the target face and, the next day received one of four interviews: a CI, an H-CI (character judgements based on the whole face and the eye region), a CI with character judgements based on the whole face, or a CI with character judgements based on the eye region. The results demonstrated that composites constructed after a full H-CI (character judgements based on the whole face and the eye region) were named approximately 10% more frequently than those constructed in any other condition.

In this thesis, composite naming rates in Experiments 1 and 2 were lower than those predicted based on the current EvoFIT literature (see, Erikson et al., 2022; Fodarella et al., 2021; Giannou et al., 2021). This reduced naming indicates that composite construction in these experiments may not be reflective of that in previous research or that carried out by the police. Implementing an H-CI before composite construction may increase the accuracy of composites constructed (Frowd et al., 2008; Portch et al., 2017; and Skelton et al., 2020), resulting in composite naming rates which are comparable to current EvoFIT research and are therefore more likely to be reliable and generalisable.

Internal and External Features

In addition to developing an understanding of the impact of cognitive load during composite construction after an H-CI, Experiment 3 will continue to explore the impact that reducing the number of screens used during composite construction has on the accuracy of composite *Internal* and *External Features*. Experiment 2 demonstrated that composite *Internal Features* were deemed more accurate than composite whole face images. This pattern of results indicated that reducing the number of screens used during composite construction does not have a negative

impact on the accuracy of the internal or *External Features*. However, it is important to ensure that the accuracy of composite internal or *External Features* does not reduce with the addition of the H-CI in place of the CI. Therefore, the current experiment will replicate Part 2b of Experiments 1 and 2, measuring likeness of *Internal* and *External Features* as well as whole faces to ensure that changing the interview procedure does not negatively impact the ability to select accurate internal or *External Features* for facial composites, which would ultimately result in a less accurate composite image.

Stages of Construction

To further address the limitations of Experiment 2, this experiment will continue to explore the impact that reducing the number of screens during composite construction has on participants' abilities to utilise each stage of composite construction. Experiment 2 demonstrated that composites constructed using more screens do not increase in accuracy as much towards the end of the construction process as they do at the start of the process. Yet, composites constructed using fewer screens continue to increase in accuracy at a similar rate throughout the composite construction process. This pattern of results suggested that increased cognitive load during composite construction impacted participants' abilities to fully utilise each stage in the construction process.

As the H-CI should refresh participants' memories of the target face more effectively than the CI (Frowd et al., 2008), it will be important to understand whether participants with an improved memory of the target face will experience cognitive load in the same manner as Experiment 2, whereby higher cognitive load during

composite construction resulted in a lower increase in composite likeness between each stage of construction towards the end of the construction process.

Experimental Aims

This experiment was designed to understand the impact of intrinsic cognitive load on witness memory during EvoFIT composite construction with an H-CI. The aims of this experiment replicate those of Experiments 1 and 2; to determine the optimum population size during EvoFIT construction, to understand if composite internal or *External Features* were adversely affected by reducing the population size and to understand whether cognitive load impacts participant's ability to utilise each stage of composite construction after an H-CI. Specifically, the hypotheses are:

H1: Composites constructed using fewer screens will be more accurate than those constructed using the typical *Four Screens.*

H2: Composites created using fewer screens will increase in accuracy more after the use of image enhancement tools towards the end of the construction procedure.

Method

Part 1 - Composite Construction

Design. A between-subjects design was utilised in which the IV was the number of screens shown during composite construction (referred to as *Screens*), and the four levels were: *One, Two, Three* or *Four Screens.* Procedures used during composite construction reflected those used by the police, including a 24-hour delay between the participant viewing the target photograph and composite construction. In order to

participate, participants were required to use a PC and have access to a video conferencing platform, such as Skype or Microsoft Teams.

Participants. Participants were 40 adults (24 females, 16 males) between 18 and 60 years (M = 32.55, SD = 14.67). Participants were recruited on the basis that they were not familiar with the target images.

Materials. The target stimuli were 10 photographs of EastEnders characters. Five of the characters were male, and five were female. Each image was displayed in colour and was approximately 8cm (wide) x 10cm (high). None of the targets had particularly distinctive characteristics that would make them easier to identify, such as a face tattoo or scar.

Procedure. Part 1 of the experiment occurred over two days. On the first day, participants received a photograph of the target during a video call and were instructed to view the image for 30 seconds, as timed by the researcher. Twenty-four hours later, participants took part in a second video call. A holistic-cognitive interview was conducted (as outlined in detail in Chapter 2, pages 62-73), during which the participant freely recalled a description of the target face in as much detail as they remembered. The participant was then invited to consider the personality traits and characteristics that the target may have based on their face. Each participant was instructed to rate the target for having low, medium or high levels of each trait read out by the researcher based on the whole face and then based on the target's eye region. Once the H-CI had been conducted, the researcher's screen was shared so that the participant could view the screen and the researcher had control of the mouse. EvoFIT composite construction took place.

Part 2a - Composite Naming

Design. A between-subjects design was employed in which the IV was *Screens* with four levels: *One, Two, Three* or *Four Screens,* and the DV was *Naming* with two levels: *Spontaneous* and *Cued.*

Participants. Participants were 40 adults (23 females, 17 males) between 18 and 69 years (M = 30.08, SD = 13.83). Participants were recruited on the basis that they were familiar with the targets. An equal number of participants were randomly allocated to each level of *Screens* (N = 10).

Materials. As in the previous experiments, composites constructed during Part 1 of the experiment were displayed in four PowerPoint presentations (one for each level of *Screens*). A fifth presentation was created containing the target photographs. This was shown to all participants to check familiarity with the targets.

Procedure. The procedure replicated that of Experiment 1, Part 2a.

Part 2b - Composite Final Image Rating

Design. Part 2b of the experiment utilised a mixed design with two IVs. The within-subjects variable was *Screens* with four levels*: One, Two, Three and Four Screens,* and the between-subjects variable was *Task* with three levels: *Internal Features, External Features* and *Whole Faces.* The DV was the accuracy of the composites, as measured by a Likert scale of 1-7.

Participants. Participants were 30 adults (10 females, 20 males) between 19 and 43 years (M = 25.4, SD = 5.43). Participants were recruited on the basis that they were not familiar with the targets. An equal number of participants were randomly allocated to each level of *Task* (N = 10).

Materials. As in the previous experiments, the materials were the *Internal Features*, *External Features* and *Whole Composites* constructed in Part 1. Again, three PowerPoint presentations were created containing images of either the *Internal Features*, *External Features* or complete faces.

Procedure. As in Experiments 1 and 2, participants viewed the PowerPoint presentation via the 'screen share' feature on Skype and verbally rated each composite image using the seven-point Likert scale. Participants were instructed to inform the researcher if they recognised any of the target images, in which case, their data would have been removed from the experiment, and a new participant recruited in the same condition.

Part 2c - Intermediate Composite Rating

Design. A within-subjects design was utilised with two IVs: *Screens* with four levels: *One, Two, Three* or *Four Screens*, and the stage of composite construction (referred to as *Stages*) with five levels: *Random Face, First Generation, Second Generation, after Holistic Tools* and *Final Image*. The DV was the accuracy of the composites, measured using the same Likert scale as Part 2b.

Participants. Participants were 18 adults (8 female, 10 male) between 20 and 56 years (M = 28.94, SD = 9.30). Participants were recruited on the basis that they were not familiar with the targets.

Materials. As in Experiments 1 and 2, images of the composite at each level of *Stages* were displayed in a PowerPoint presentation alongside a random composite image and a photograph of the target.

Procedure. This procedure replicated that from Part 2a.

Results

Part 2a – Composite Naming

To test the hypothesis that composites constructed using fewer screens are more accurate than those constructed using the typical *Four Screens*, responses to facial composites and target pictures were scored for accuracy, as in the previous experiments. As per the *a priori* rule, data were only included from participants who correctly named at least 8 of the 10 target pictures. Responses were coded as in Experiments 1 and 2: they were assigned a value of 1 when a correct name was given for the target image *and* composite photograph; 0 when a wrong name or "don't know" response was given for the composite image, but the target was identified correctly; and 2 when the target was not identified. Incorrect names for target pictures occurred 36 times (by 25 participants), five to eight times by group. As such, the mean correct naming for target pictures was high (M = 91.00%, SD = 8.10%). Where a target had not been named correctly (data assigned a value of 2), the associated composites also could not be named correctly, so responses to these composites were removed prior to analysis.

As expected, correct responses were much lower overall for spontaneous naming of facial composites compared to target images (M = 50.27%, SD = 50.01%). Table 11 displays the mean correct naming at each level of *Screens*: *One, Two, Three or Four Screens.*

Table 11. Spontaneous Naming of Composites for Each Level of Screens

Number of Screens	4	3	2	1
Correct naming	39.78 (37 / 93)	44.94 (40 / 89)	55.91[1] (52 / 93)	60.67[1] (54 / 89)

Note. Figures are expressed in percentage and calculated from participant responses in parentheses: summed correct responses (numerator) and total (correct and incorrect) responses (denominator). Data are presented for composites for which participants correctly named the relevant target photographs (N = 364 out of 400). [1]$p < .05$.

Table 11 demonstrates that composite accuracy increases as the number of screens used during composite construction decreases. Therefore, composites constructed using *One Screen* were named most frequently, followed by those constructed using *Two*, *Three* and *Four Screens.* Individual responses were analysed using GLMM.

This stage of the experiment involved one predictor: *Screens* (coded as *1 = One Screen*; *2 = Two Screens*; *3 = Three Screens*; *4 = Four Screens*). The DV was correct composite naming; in cases where the target was named incorrectly, the corresponding composite for that target was not included in the analysis. A hypothesis-testing approach comprised one model for *Screens,* which was significant [F (3,39) = 3.60, p = .016]. To explore the significant result, fixed coefficients were examined (Table 12).

Table 12. Model Parameters for Effect of Number of Screens on Correct Composite Naming. Comparisons Are Presented with Reference to The Lowest Category- Four Screens; Positive Values of B Indicate Higher Naming with Respect to The Reference.

	B	*SE(B)*	*t (360)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
Screens							
Intercept	-5.51	0.43	-1.29	.26	0.58	0.25	1.33
3 vs. 4	0.26	0.33	0.80	.43	1.30	0.68	2.49
2 vs. 4	0.74	0.33	2.27	.024	2.10	1.10	3.99
1 vs. 4	0.96	0.33	2.89	.004	2.62	1.36	5.03

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix)] final, Model-based full factorial Corrected model [F (3,360) = 3.49, p = .016]. The model was specified with the lowest category of categorical predicators as reference (*4 Screens*), and predictors were sorted in descending order. Information criteria are based on -2 log likelihood (*AICC* = 1631.71, *BIC* = 1635.59). Variance of random slopes intercept for item [1.26, *SE* = 0.70, *Z* = 1.79, p = .07, *CI* (0.42, 3.76)].

This analysis indicated that composites constructed using fewer screens were successively more accurate than those constructed using the current procedure, *Four Screens*. It revealed a sizeable significant increase in composite accuracy of reducing the number of screens during construction from *Four* to *Two,* and a further benefit when reduced to *One Screen*. Reducing the number of screens from *Four* to *Three* was also somewhat beneficial, but this difference was not significant.

To understand the mathematical pattern of the data, polynomial contrasts were simulated using GLMM. Specifically of interest was the presence of a significant linear trend, which would indicate that composite accuracy increases at a similar rate from *Four*, to *Three*, to *Two,* to *One Screen*, as was predicted based on the results

displayed in the table of means. A polynomial contrast indeed demonstrated a significant linear trend by Screens (p = .002) and non-significant quadratic (p = .92) and cubic (p = .65) trend.

Part 2b - Composite Final Image Rating

Composite *Internal and External Features,* as well as *Whole Composites,* were analysed for likeness to assess the impact of reducing the number of screens during EvoFIT composite construction on composite accuracy. Hence, composites constructed using *One, Two, Three* or *Four Screens* were rated for likeness in comparison to the target photograph using a Likert scale of 1-7. Table 13 presents the mean (and standard error) for composite rating at each level of *Task (Internal Features, External Features* and *Whole Composites)* and *Screens (One, Two, Three* or *Four Screens).*

Table 13. Mean (and Standard Error) for Rating Internal and External Features as well as Whole Composites Constructed Using One, Two, Three or Four Screens

Task [a]	Screens [a]				
	4	3[a]	2	1	Mean
Internal	3.52	2.91	3.34	3.71	3.37
Features [b]	(0.14)	(0.12)	(0.13)	(0.13)	(0.07)
External	4.53	4.17	4.22	4.68	4.4
Features	(0.14)	(0.12)	(0.13)	(0.13)	(0.07)
Whole	3.35	2.94	2.99	3.43	3.18
Composites [b]	(0.17)	(0.14)	(0.14)	(0.19)	(0.09)
Mean	3.80	3.34	3.52	3.94	3.65
	(0.09)	(0.08)	(0.08)	(0.09)	(0.04)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant SE of the mean. [a] $p < .05$, [b] $p < .001$.

Table 13 showed that composites constructed using *One Screen* were rated the most accurate, followed by those constructed using *Four, Two* and *Three Screens.* This pattern of results is consistent for rating of *Internal* Features, *External Features* and *Whole Composites*. Individual ratings of composite items from participants were analysed using GLMM. This stage of the experiment involved two predictors, *Screens* (coded as *1 = One Screen; 2 = Two Screens; 3 = Three Screens; 4 = Four Screens*) and *Task* (coded as *0 = Whole Face, 1 = Internal Features and 2 = External Features*). The DV was rating of composite accuracy compared to the target photograph on a Likert scale of 1-7. However, very few images were rated as 7 in some of the conditions; therefore, ratings for six and seven were combined into one category (six).

A hypothesis-testing approach comprised three models. The first model was for *Screens* only, and the second model was for *Task* only. The third was a full-factorial model comprising both predictors (*Screens* and *Task*) and the interaction between them. The model for *Screens* was significant [F (3, 1193) = 3.59, p = .013], the model for *Type* was also significant [F (2, 1194) = 73.50, p < .001], but the interaction between *Screens* and *Type* was not significant [F (6, 1185) = 0.88, p = .51]. Therefore, a fourth model was run containing the two predictors' *Screens* and *Type* without the interaction. This model was significant for *Screens* [F (3, 1191) = 3.53, p = .015] and *Type* [F (2, 1191] = 76.10, p < .001] and so became the final model. Fixed coefficients were examined to explore the two significant predictors in the final model (Table 14).

Table 14. Model Parameters from Composite Rating at Each Level of Facetype. Comparisons Are Presented with Reference to The Lowest Category (Underlined).

	B	*SE(B)*	*t (1193)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
Screens							
3 vs. 4	-0.59	0.25	-2.33	.020	0.55	0.34	0.91
2 vs. 4	-0.35	0.25	-1.37	.17	0.71	0.43	1.16
1 vs. 4	0.16	0.25	0.61	.54	1.17	0.71	1.92
Facetype							
Internal vs Whole Face	0.22	0.16	1.33	0.18	1.24	0.90	1.72
External vs Whole Face	1.58	0.13	11.84	< .001	4.86	3.74	6.31
External vs Internal	1.30	0.17	7.58	<.001	3.66	2.62	5.13

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix)] final, Model-based full factorial Corrected model [F (5, 1789) = 10.49, p < .001]. The model was specified with the lowest category of categorical predictors as reference (underlined) and predictors were sorted in an ascending order. Information criteria are based on -2 log likelihood (*AICC* = 41385.74, *BIC* = 41396.71. Variance of random slopes intercept of participants for *Screens* [1.33, *SE* = 0.32, *Z* = 1.14, p < .001, *CI* (0.83, 2.13)].

These results demonstrate that composites constructed using *Four Screens* were significantly more accurate than those constructed using *Three Screens*. However, the difference in likeness ratings between composites constructed using *Four Screens* and those constructed using *One* or *Two Screens* was not significant. Furthermore, composite *External Features* were rated as significantly more accurate

than *Whole Faces* and *Internal Features,* but there was not a significant difference between *Whole Faces* and *Internal Features.*

To understand the mathematical pattern of these results, polynomial contrasts were simulated using GLMM in SPSS, with a focus on linear and quadratic patterns. This analysis demonstrated that a linear pattern between the levels of *Screens* was not significant ($p = .21$, *Exp(B)* = 0.94) but a quadratic pattern was significant ($p < .001$, *Exp(B)* = 1.28). When the full model was run, including the linear, quadratic and cubic contrast, cubic was not significant ($p = .25$, *Exp(B)* = 1.10), and the quadratic contrast remained significant ($p < .001$, *Exp(B)* = 1.32) indicating that the pattern of data fits a trend where composite accuracy changes between each level of screens by a varying amount, and not necessarily in the same direction, for example, increasing from *Four Screens* to *One Screen* as it would with a significant linear trend.

Part 2c - Intermediate Composite Rating

Part 2c tests the hypothesis that reducing the number of screens during composite construction is beneficial for composite accuracy because cognitive load impacts the ability to utilise each stage of composite construction. Ratings of composite *Internal Features* were analysed at four stages of composite construction (*First Generation*, *Second Generation*, After *Holistic Tools*, *Final Image*) and a random composite image for comparison. As composite *Internal Features* from each stage of construction were displayed simultaneously, ratings push the scores apart across *Stage*, leaving the mean relatively unchanged. This means that the analysis expects a difference in rating scores between the stages of construction across *Screens, i.e.*, an interaction between *Screens* and *Stage.* Table 15 presents the means (and standard error) of likeness ratings for composite *Internal Features* created using *One*, *Two*,

Three or *Four Screens* during face selection and at each level of *Stage.* Table 15 demonstrated that composite *Internal Features* generally increased in accuracy throughout the construction process, suggesting that composites constructed using *One Screen* were the most accurate in the *Final Image*.

Table 15. Mean (and Standard Error) for Composite Rating at Each Level of Stages and Screens.

Stage of Construction [a]	Screens				
	Four	Three	Two	One	Mean
Random Face	2.02	2.01	2.10	2.17	2.07
	(011)	(0.10)	(0.11)	(0.11)	(0.05)
First	3.12	2.89	2.89	2.91	2.95
Generation	(0.11)	(0.11)	(0.10)	(0.11)	(0.05)
Second	3.31	3.17	3.21	3.11	3.20
Generation	(0.11)	(0.10)	(0.10)	(0.11)	(0.05)
After Holistic	3.39	3.33	3.64	3.57	3.48
Tools	(0.10)	(0.11)	(0.11)	(0.11)	(0.05)
Final Image	3.53	3.42	3.63	3.66	3.56
	(0.11)	(0.11)	(0.11)	(0.12)	(0.06)
Mean	3.07	2.96	3.10	3.09	3.05
	(0.05)	(0.05)	(0.05)	(0.05)	(0.03)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant SE of the mean. [a] $p < .001$.

To further understand the impact that *Screens* and *Stage* have on composite accuracy, the results were analysed using GLMM. This stage of the experiment comprised two predictors: *Screens* (coded as *1 = One Screen*; *2 = Two Screens*; *3 = Three Screens, 4 = Four Screens*) and *Stage* (coded as *0 = Random Face*, *1 = First*

Generation best face, *2 Second Generation* best face, *3* = after use of *Holistic Tools* and *4* = *Final Image*). The DV was rating of composite accuracy compared to the target photograph on a Likert scale of 1-7. However, since participants gave few ratings of 7, these scores were re-coded to have a value of 6 (i.e., categories of 6 and 7 were merged together).

A hypothesis-testing (confirmatory) approach was followed that comprised three models, each specified with different fixed effects (predictors) along with appropriate random intercepts (as described above). One model contained *Screens* only, and a second contained *Stage* only. A third model contained the interaction between these two fixed effects; as it is standard practice to include the individual predictors in a model that contains their interaction (e.g., Field, 2013), this third model was full factorial. Based on the usual alpha of .1 for regression analyses, the model for *Stage* only was significant [F (4, 3591) = 17.01, $p < .001$], but the model for *Screens* only was not significant [F (3, 3592) = 0.29, $p = .83$]. The interaction between *Screens* and *Stage* was also not significant [F (12, 3576) = 0.82, $p = .63$]. Therefore, the final model was for *Stage* only. Fixed coefficients were examined to explore the significant result, displayed in Table 16.

Table 16. Model Parameters from Composite Likeness Ratings at Each Level of Stage. Comparisons Are Presented with Reference to The Lowest Category (Random face).

Stage	*B*	*SE(B)*	*t (3591)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
Final Image vs Random Face	2.47	0.34	7.38	< .001	11.87	6.15	22.90
Holistic Tools vs Random Face	2.38	0.34	7.09	< .001	10.76	5.58	20.77
2nd Generation vs Random Face	1.94	0.33	5.78	< .001	6.92	3.59	13.34
1st Generation vs Random Face	1.55	0.33	4.65	< .001	4.73	2.46	9.11

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix)] final, Model-based full factorial Corrected model [F (4, 3591) = 17.71, p <001]. The model was specified with the lowest category of categorical predictors as reference (*Random Face*), and predictors were sorted in an ascending order. Information criteria are based on -2 log likelihood (*AICC* = 64633.39, *BIC* = 64658.12). Variance of random slopes intercept of participants [0.83, *SE* = 0.32, *Z* = 2.57, *p* = .010, *CI* (0.39, 1.79)] and items for *Stage* [0.07, *SE* = .06, *Z* = 1.13, *p* = .26, *CI* (0.01, 0.41)].

The results demonstrate a significant increase in ratings of composite accuracy from *Random Faces* to *First Generation*, *Second Generation*, *Holistic Tools* and *Final Image*. The effect size, as indicated by the odds ratio (Exp(B)), between the *Random Face* and the composite image increases appreciably successively for each stage of construction.

To gain a further understanding of the mathematical pattern of the data, polynomial contrasts were conducted using GLMM for *Stages*. The analysis demonstrated a significant linear (p <.001, *Exp(B)* = 0.77) and quadratic contrast (p <.001, *Exp(B) = 0.82*). When cubic and quartic contrasts were added to the model, the linear contrast was still significant (p = .014, *Exp(B)* = 1.40) but the quadratic contrast was not (p = .39, *Exp(B)* = 1.10); additionally, cubic (p = .007, *Exp(B)* = 1.08) and quartic (p = .006, *Exp(B)* = 0.93) contrasts were significant. As linear, cubic and quartic contrasts were all significant, it is likely that the mathematical pattern of the data for *Stages* is not consistent for all levels of *Screens*, indicating that there is a *Screens*Stages* interaction. Furthermore, the effect size is somewhat larger for the linear contrast compared to the quadratic, cubic and quartic contrasts. This indicates that a linear pattern is a better fit for the data.

Discussion

This experiment aimed to develop an understanding of the impact that cognitive load has on EvoFIT composite construction after an H-CI. This experiment demonstrated that reducing the number of screens used during composite construction to *One Screen* was beneficial to composite accuracy. This finding indicates that high intrinsic cognitive load during EvoFIT composite construction has a negative effect on participants' abilities to create an accurate composite image.

The first hypothesis, that composites constructed using fewer screens would be the most accurate, was supported. Composite naming demonstrated a linear trend whereby correct naming increased as the number of screens viewed during the construction procedure decreased. Moreover, composites constructed using *One Screen* were rated as the most accurate in both rating measures. In Part 2b, composites

constructed using *One Screen* were rated as the most accurate based on the *Internal Features, External Features* and *Whole Composites*. In Part 2c, there was no significant difference in composite likeness between the levels of *Screens.* However, mean likeness demonstrates that composites constructed using *One* or *Two Screens* were the most accurate in the *Final Image*, indicating a benefit of reducing the number of screens during composite construction.

In Part 2b, although composites constructed using *One Screen* were the most accurate, composites constructed using *Four Screens* were rated as the second most accurate, followed by those constructed using *Two* and *Three Screens*. This pattern of results indicates that viewing a wider selection of face images is somewhat beneficial for composite accuracy, a pattern that was also found in Experiment 2. In Part 2c, the composite *Final Image* was rated the highest for composites constructed using *One Screen,* followed by those constructed using *Two, Four* and *Three Screens.* This pattern of results supports the hypothesis, as composites constructed using fewer screens (*One* or *Two Screens*) were rated as more accurate than composites constructed using more screens (*Three* or *Four Screens*). However, the difference in composite accuracy in this stage of the experiment was very similar, demonstrating no significant difference between *Screens.*

Overall, the findings from this experiment support the theory of cognitive load (Sweller, 1988), as the simpler a task is (i.e., fewer elements are involved), the lower the intrinsic load is, and thus the likelihood of cognitive overload. Therefore, simplifying the task of selecting face images during EvoFIT composite construction by reducing the number of face arrays viewed from *Four* to *One*, should reduce the likelihood of cognitive overload (Sweller et al., 2011). As cognitive overload results in impairments of memory capacity, participants who experience cognitive overload

may perform poorly, creating a facial composite image that does not resemble the target (Sewell et al., 2020).

In all three measures of composite accuracy, composites constructed using *One Screen* were named the most frequently or rated as the most alike the target photograph. This pattern of results indicates that composite construction using *One Screen* is optimal. Furthermore, in two of the three measures (Composite Naming and Intermediate Composite Rating), composites constructed using *Two Screens* were deemed the second most accurate. This finding demonstrates that reducing the intrinsic cognitive load during EvoFIT construction by reducing the number of screens used is beneficial to the accuracy of the resulting composite images.

An alternative explanation for the pattern of results in this experiment is that the attention span of participants is not long enough to adequately complete composite construction using more screens. Composites constructed using *One Screen* may be the most accurate because composite construction using fewer screens is likely to be quicker than construction using more screens. If a participant's attention span does not last the whole procedure, they may become fatigued and distracted, no longer paying attention to the construction process. However, there is a lack of solid evidence demonstrating how long the average individual's attention span is. Some theories determine an individual's attention span based on their age (see, Fortenbaugh et al., 2015; Valessi et al., 2021), while other theories suggest that the average attention span ranges from 8-seconds (quoted in Customer Insights, Microsoft Canada, 2015) to 6 hours (Cornish & Dukette, 2009), depending on the research. Therefore, it can be difficult to comprehend whether the time difference between composites constructed using *One* or *Four Screens* is large enough to be explained by a participant's attention span. Furthermore, if the attention span of participants was

the reason for composites constructed using *One Screen* being more accurate than those constructed using *Four Screens,* one would expect to see this same pattern of results in Experiments 1 and 2. However, in the two previous experiments, the results did not clearly indicate that composites constructed using fewer screens were the most accurate. Specifically, Experiment 2 demonstrated that composites constructed using *One* or *Four Screens* were the most accurate, which cannot be explained by attention span.

A second alternative explanation for the pattern of results is Perceptual Load Theory (Lavie & Russell, 2003). This theory states that, although attentional resources are limited in capacity, all of the attentional resources must be used at all times, though task-relevant stimuli are processed before task-irrelevant stimuli. Put simply, if a task uses all of the attentional resources on stimuli relevant to the task, no irrelevant stimuli will be processed. However, if a task is less complex and does not use all of the attentional resources on stimuli relevant to the task, irrelevant stimuli will be processed, which may be distracting to the goals of the task. During EvoFIT composite construction, all of the information on the screen is relevant to the task, as the interface has been designed with knowledge of theories of distraction and attention (see, Broadbent, 1958; Lavie, 1995; Triesman, 1964). Therefore, none of the information *on-screen* should be considered irrelevant stimuli, though distractions in a participant's environment may be notable.

According to Lavie's (1995) Perceptual Load Theory, composite construction using *One Screen* may use fewer attentional resources on stimuli relevant to the task, which would actually increase the number of irrelevant stimuli (distractions) processed in the participants' environment, potentially resulting in a less accurate composite. Yet, composite construction using *Four Screens* may use more attentional

resources on stimuli relevant to the task, reducing the number of distractions processed and resulting in a more accurate composite image. However, as there are only a limited number of stimuli, participants may be unable to process all of the face images during the face selection process. In this case, there may be no distractions processed, but participants may be unable to process face images which best resembled the target, which would again result in the construction of a less accurate composite image. Alternatively, as there are 18 face images on each screen, as opposed to 72 face images shown simultaneously (18 face images on *Four Screens*), participants may attend to each face image as they would for composite construction using *One Screen*, processing the same number of distractions per screen. Hence, there would be little difference between composite construction using *One Screen* and that using *Four Screens.* However, in this experiment, correct naming of composites constructed using *One Screen (M* = 60.67) is almost double that of composites constructed using *Four Screens (M* = 39.79). Therefore, Lavie's Perceptual Load Theory does not explain the findings and Cognitive Load Theory, as originally proposed, provides a more robust explanation of the results.

In Part 2a of this experiment, composite naming was higher than in Experiments 1 and 2 and was similar to naming achieved in EvoFIT literature more generally (see, Erikson et al., 2022; Fodarella et al., 2021; Giannou et al., 2021). This increase in composite naming indicates that implementing an H-CI in place of a CI before composite construction is beneficial for increasing the accuracy of composites constructed. One reason for the increased naming rate for composites in this experiment compared to Experiments 1 and 2 may be that participants had a better memory of the target face. As such, participants may have been affected less by high germane load as the comparison between the target face and the images on screen was

less effortful (Kalyuga, 2011). Moreover, composites constructed using *Four Screens* are less accurate because they are affected by high intrinsic load, and not because they are affected by germane load as they may have been in Experiment 2.

To further test the hypothesis that composites constructed using fewer screens would be the most accurate, composite images were rated in comparison to the target face based on how alike the composite image was to the target photograph. The result of Part 2b somewhat supported the hypothesis, demonstrating that composite images constructed using *One Screen* were the most accurate, supporting the finding in Part 2a, where composites constructed using *One Screen* were also deemed the most accurate. However, unlike Part 2a, where a linear trend demonstrated an increase in correct composite naming as the number of screens decreased, this part of the experiment found that composites constructed using *Four Screens* were the next most accurate. Overall, this result indicates some benefit to reducing the cognitive load during EvoFIT composite construction, as composite images created using the fewest screens (*One Screen*) were rated higher than those constructed in any other condition. However, this part of the experiment also indicates that there is a benefit of participants having the opportunity to select face images from a wider array of options. It may therefore be important to find a balance between reducing the cognitive load and maintaining an optimal number of face options participants can view during the composite construction process. In future research, one method to reduce the cognitive load is this way may be to reduce the number of screens used for the selection of face images while maintaining the use of *Four Screens* for the selection of face textures, or vice versa. Alternately, it may be important to assess the impact of reducing the number of face images on each screen while maintaining the use of *Four Screens* for selection of the face shape and texture overall.

As well as demonstrating the benefit of reducing the number of screens during composite construction, Part 2b also aimed to ensure that reducing the number of face options available during composite construction did not have a negative impact on the accuracy of composite *Internal Features*. As participants only viewed the *Internal Features* during the face selection stage of composite construction, viewing fewer options to select this face region may result in *Internal Features* which are vastly reduced in accuracy. To test this, participants viewed either a whole composite face image, only the composite *Internal Features* or only the composite *External Features* and were asked to rate how alike the target photograph was compared to the face region viewed. The results demonstrated that *External Features* were rated the most accurate, followed by *Internal Features* and then *Whole Composites*. This result is unsurprising as the *External Features* are deemed to be more important than *Internal Features* for the recognition of unfamiliar faces (Bruce & Young, 1998; O'Donnell & Bruce, 2001). During composite construction, the target face remains unfamiliar to participants as they only view the face for 30 seconds; therefore, it is sensible to find that participants are better at selecting accurate *External Features* for the composite than they are at selecting *Internal Features*.

Moreover, composite *Internal Features* were rated as more accurate than *Whole Composites* in comparison to the target face. This pattern of results replicates that of Experiment 2, demonstrating that composite *Internal Features* are frequently deemed more accurate than the whole face. Overall, the results from Part 2b demonstrate that reducing the number of screens viewed during EvoFIT composite construction does not have a negative impact on the accuracy of composite internal or *External Features*, as both of these facial regions are deemed more accurate in isolation than they are as a whole composite.

The second hypothesis was that the number of screens used during composite construction would impact the participant's abilities to utilise each stage of construction. The results of Part 2c demonstrated that there is no clear interaction between the number of screens used during composite construction and the ability to utilise each stage of composite construction. For composites constructed using *One*, *Three* or *Four Screens*, composite accuracy increased at each stage of composite construction. However, the increase between each stage did not appear to be related to the number of screens. Interestingly, the increase between each stage of composite construction when composite images were created using *Three Screens* reduced at each subsequent level. This is the result that is expected if a participant experiences cognitive overload, as their ability to utilise each stage of construction reduces at each level. However, this same pattern was not found for composites constructed using *Four Screens*, where the cognitive load is expected to be highest. Further research is needed to better understand the pattern of composite accuracy throughout the stages of composite construction.

Limitations and Future Research

A limitation of the current experiment, and the previous two experiments, is that there was no direct measure of cognitive load. Future research could use subjective measures of cognitive load, such as Paas' (1992) questionnaire. However, this questionnaire faced heavy criticism due to the inconsistency between the labels used throughout the questionnaire (i.e., whether the phrase "task difficulty" or "mental effort" should be used) as well as the inconsistency in timing and frequency of measurement. More reliable measures of cognitive load include physiological measures of stress, such as electrodermal activity (EDA), electroencephalography

(EEG) and electrooculography (EOG). However, such physiological measures would not be possible in the current thesis as all testing was remote. Furthermore, conducting such tests during EvoFIT composite construction may induce stress in participants, and would not replicate composite construction with the police. Future research should seek to identify a more reliable, suitable measure of cognitive load during EvoFIT composite construction in order to more reliably determine the cognitive load during composite construction.

A further limitation of the current experiment is that the number of screens used for selection of the face shape and the face texture is reduced at an equal rate. However, face perception literature has questioned the importance of both face *Shape* and Texture for recognition of familiar and unfamiliar faces, indicating a disparity in their importance (Rogers et al., 2022). On balance, it is also important to note that a facial composite image is an imperfect image of a face, as opposed to a photograph which is a perfect image. Hence, the roles that face *Shape* and *Texture* play in the recognition of a facial composite may differ from those in the recognition of a face photograph.

Research using a variety of methods, including modifying face photographs or head models and caricaturing, demonstrates an increased importance for facial shape in recognition of an unfamiliar face and an increased important for facial texture for the recognition of a familiar face. Eyewitnesses create a facial composite of an unfamiliar face, indicating that the face shape may be more important during the construction procedure (Bruce et al., 1991; Knight & Johnston, 1997). Yet, individuals identifying the composites are familiar with the target identities, indicating that the face texture may be more important (Bruce et al., 1991; Lee & Perrett, 1997; Rogers et al., 2022; Russell & Sinha, 2007). Therefore, the next

experiment will manipulate the number of screens used for the selection of the face *Shape* and *Texture* to understand the importance of *Shape* and *Texture* for the creation of an identifiable composite.

This chapter analysed the impact of cognitive load on participants' abilities to create an accurate facial composite image following a Holistic-Cognitive Interview (H-CI). As in previous experiments, composites were created using either *One*, *Two*, *Three* or *Four Screens* to select the face shape and texture. The pattern of results in this experiment replicated Experiment 1, whereby composite likeness increased as the number of screens used during the construction procedure decreased. This result clearly demonstrated the negative impact that viewing many screens during the construction procedure has on composite likeness.

The first three experiments in this thesis assumed that reducing the number of screens used to select the face shape and the face texture at the same rate would be optimal. However, as face shape is typically most important for unfamiliar face recognition (Bruce et al., 1991; Kaufmann et al., 2013; Limbach et al., 2022), fewer screens may be needed to accurately select the face shape during the construction procedure than those needed to accurately select the face texture. Therefore, the next chapter will explore the impact of cognitive load on facial composite construction by manipulating the number of screens viewed for selection of face *Shape* and *Texture* individually.

6

EXPERIMENT 4

REDUCING THE POPULATION SIZE DURING EVOFIT CONSTRUCTION FOR FACE SHAPE AND TEXTURE INDEPENDENTLY

Abstract

The current experiment aims to reduce the number of screens individually for face *Shape* and *Texture* to determine the optimum number of screens used for the selection of face *Shape* and *Texture*. In this experiment, forty participants constructed composites using *Two* or *Four Screens* to select the face *Shape* and *Texture* during composite construction. As in the previous experiments, composites were judged for likeness through composite naming by 60 participants and composite likeness rating by 48 participants. The results demonstrated that reducing the number of screens

overall was beneficial for composite likeness, and that reducing the number of screens for selection of the face *Shape* was more effective than it was for selection of the face *Texture*.

The results of the previous experiments demonstrated that reducing the number of screens used to select the face images during EvoFIT composite construction is beneficial for the creation of an identifiable composite image. During EvoFIT construction, face *Shape* and face *Texture* are selected individually. Yet, in the previous experiments, the number of screens used to display the face *Shape* and the face *Texture* options were reduced at the same rate. This equal reduction in the number of screens assumes that the *Shape* and the *Texture* of the composite face are of equal importance. However, there is evidence to suggest a disparity in the importance of face *Shape* and face *Texture* for facial recognition, particularly linked to how familiar an individual is with the face in question (Burton et al., 2015). Such evidence indicates that it may not be optimal to view the same number of faces for selection of the face *Shape* and *Texture*. Therefore, composite construction may be further optimised by reducing the number of screens displaying face *Shape* and *Texture* independently. Furthermore, this experiment will offer a better theoretical understanding of the importance of face *Shape* and *Texture* for facial composite recognition.

Facial Shape and Texture

The shape and texture of a face both play an important role in the perception and recognition of face images (Bruce et al., 1991). However, there are some instances

where the face shape or the face texture is considered to be more crucial, for example, during familiar and unfamiliar face recognition as well as age estimation (Lu & Tan, 2011; Rhodes, 2009). Face shape is typically viewed as being most important for the recognition of unfamiliar faces. Head models of unfamiliar faces were displayed either static or moving, as though the head was looking in various directions. Displaying the head model moving in this way provides more information about the face shape compared to when the head is still. Participants were invited to match the unfamiliar face in the head model with the correct photograph from a line-up. Faces in head models which were displayed in motion were identified more frequently than head models displayed stationary (Knight & Johnston, 1997). Such evidence demonstrates that information about the shape of a face is more important than information about the texture of a face for unfamiliar face recognition.

On the other hand, face texture is deemed more important than face shape for recognition of familiar faces (Burton et al., 2015; Rogers et al., 2022). When familiar face images were modified to change the configuration of features, modifying the shape of the face, participants were still able to recognise the individuals from the images (Burton et al., 2015). This finding supports the understanding that face texture is more important than face shape for familiar face recognition. Similarly, in a face-matching task, images of familiar faces with morphed texture were more difficult to recognise than those with morphed shape, supporting a superiority effect of textual information of a shape for familiar face recognition (Itz et al., 2017).

Other research has shown that hybrid faces created using the shape of one familiar face and the texture of a different familiar face were identified most frequently based on the face texture than the face shape (Rogers et al., 2022). Furthermore, 3-D head models of familiar faces were named more frequently when

only textural information was available than when only shape information was available, although naming was optimal when shape and textural information were both displayed (Bruce et al., 1991). This same result was found when participants attempted to name individuals known to them based on photographs which had been modified to include only shape or only textural information (Russell & Sinha, 2007). There is evidence, therefore, using a variety of methods that demonstrates the importance of face texture over face shape for familiar face recognition.

However, familiarity with a face may not be the only factor which dictates the importance of shape and texture for its recognition. Another element which may play a role in determining the importance of face shape and texture in face recognition is the skill of the individual attempting to recognise the face. In Kaufmann et al. (2013), the Bielefelder Famous Faces Test (BFFT) was used to determine individuals who are good at recognising faces and those who are poor at recognising faces. Kaufman et al. then assessed the speed and accuracy at which participants in both groups (*good recognisers* and *poor recognisers*) could identify an unedited familiar face, or a familiar face with the facial shape caricatured. Poor recognisers were better able to recognise the caricatured face than the unedited face, whereas good recognisers performed equally with the caricatured and the unedited face. This finding indicates that better recognisers may rely on texture more than shape for recognition of a familiar face (Kaufmann et al., 2013). Therefore, perhaps the importance of face shape and texture for facial composite recognition is more individualistic, with a different procedure required depending on recognition ability.

The importance of face shape and texture may also rely on the angle at which an individual is facing (Bruce et al., 1991; Hill et al., 1997). Head models of unfamiliar faces were modified to contain no textural information and were displayed

at various angles, from face-on to profile. The results demonstrated that head models of unfamiliar faces displayed at a ¾ view, where most shape information about the face is available, were identified most frequently (Bruce et al., 1991). In Hill et al. (1997), learned face images were recognised equally well based on the front, a three-quarter view, and the side profile but were recognised poorly when inverted (i.e., the Thatcher Effect, see Thompson, 1980). The angle at which a face image was encoded dictated the angle at which the face image was best recognised, with performance reducing with increasing angle of rotation. This finding was stronger when only the face shape was viewed. When textural information was included, viewpoint dependence was reduced, meaning that the angle at which the face was encoded was less important.

During EvoFIT composite construction, eyewitnesses create a face image by selecting the face shape and then the face texture, spending approximately the same time on each. However, the literature indicates that there are some differences in the importance of face shape and texture for face recognition, which could influence the overall accuracy of a composite. The current experiment will reduce the number of screens used to select the face shape and texture by varying amounts in an attempt to optimise composite construction and produce more accurate facial composite images.

Experimental Aims

This experiment aimed to optimise EvoFIT composite construction by reducing the number of screens used to select the face *Shape* and the face *Texture*, with the aim of reducing the intrinsic cognitive load during composite construction in a targeted way. This experiment hypothesises that:

H1: Composites constructed using fewer screens to select the face *Shape* will be the most accurate, as face *Shape* is deemed most important for the recognition of unfamiliar faces, meaning that fewer face options will be needed for accurate construction.

H2: Composites constructed using fewer screens overall will be more accurate than those constructed using more screens.

H3: Composites constructed using fewer screens will increase in likeness after the use of image-enhancing tools towards the end of the construction procedure more than composites created using more screens.

Method

Experiment 4 is separated into four parts. In Part 1, composites were constructed via EvoFIT using *Two* or *Four Screens* of face *Shapes* and *Textures*. *Four Screens* was used as the baseline because this allows a comparison to be made between published EvoFIT research, which usually implements composite construction using *Four Four*. Although composite construction using *One Screen* was optimal previously, it is reasonable to predict that participants may become confused when constructing a composite using *Four Screens* for face *Shape* but only *One Screen* to select face *Texture* (or vice versa). Consequently, a decision was made to reduce the number of screens to only *Two Screens* (cf. *One Screen*), as selecting face *Shape* from *Four Screens* and *Texture* from *Two Screens* (and vice versa) may be more reasonable for participants.

Part 2a involved composite naming. In Part 2b, composite face *Shape*, face *Texture* or *Whole Faces* were rated for likeness against target images, and in Part 2c, composites were rated for likeness against the target at different stages in the

composite construction process. In Parts 1, 2b and 2c, participants were recruited on the basis that they were not familiar with the targets. In Part 2a, participants were recruited on the basis that they were familiar with the targets. As an *a priori* rule, participants who could identify 80% of the target photographs were deemed familiar with the targets.

Part 1- Composite Construction

Design. A 2x2 factorial design was used whereby the two factors were *Shape* and *Texture,* and the two levels were *Two Screens* and *Four Screens*. Therefore, composites were constructed using *Two Screens* of face *Shapes* and face *Textures* (2 x 2 condition), *Two Screens* of face *Shapes* and *Four Screens* of face *Textures* (2 x 4 condition), *Four Screens* of face *Shapes* and *Two Screens* of face *Textures* (4 x 2 condition), or *Four Screens* of face *Shapes* and *Textures* (4 x 4 condition). The experiment was designed to reflect the procedures used by the police, including a 24-hour delay between the participant viewing the target photograph and composite construction. To participate, participants were required to use a PC and have access to a video conferencing platform, such as Skype or Microsoft Teams.

Participants. Participants were 40 adults (29 females, 11 males) between 18 and 52 years (M = 27.45, SD = 8.55). Participants were recruited on the basis that they were not familiar with the targets.

Materials. The target stimuli were 10 photographs of Coronation Street characters. Five of the characters were male, and five were female. Each image was displayed in colour and was approximately 8cm (wide) x 10cm (high). None of the targets had particularly distinctive characteristics that would make them easier to identify, potentially reducing experimental power.

Procedure. Part 1 of the experiment occurred over two days. On the first day, participants received a photograph of the target during a video call and were instructed to view the image for 30 seconds, timed by the researcher. Twenty-four hours later, participants took part in a second video call. As in the previous experiment, rapport building occurred, and an H-CI was conducted, after which the researcher shared her screen so that the participant could view the screen and the researcher had control of the mouse. EvoFIT composite construction took place in line with the previous experiments; however, the number of screens viewed containing face shape and *Texture* options varied by condition.

Part 2a- Composite naming

Design. The same 2x2 factorial design was used as Part 1, with *Shape* and *Texture* as factors and the two levels: *Two Screens* and *Four Screens*. The DV was *Naming* with two levels: *Spontaneous* and *Cued.*

Participants. Participants were 60 adults (37 females and 23 males) aged between 18 and 58 years (M = 34.48, SD = 12.82). An equal number of participants were randomly allocated to each level of *Screens* (N = 15). Participants were recruited on the basis that they were not familiar with the targets. To be deemed unfamiliar, participants must be unable to name 80% of the targets. No participants failed this a priori rule.

Materials. As in the previous experiments, composites constructed during Part 1 of the experiment were displayed using four PowerPoint presentations (one for each condition). A fifth presentation was created containing the target photographs.

Procedure. The procedure replicated that of Experiment 1, Part 2a.

Part 2b - Composite Final Image Rating

Design. A mixed design was utilised with two IVs. *Screens* was a within-subjects variable with four levels: *Two Screens* for *Shape* and *Texture; Two Screens* for *Shape* and *Four Screens* for *Texture; Four Screens* for *Shape* and *Two Screens* for *Texture; Four Screens* for *Shape* and *Texture. Task* was a between-subjects variable with three levels: rating of face *Shape, Texture* or *Whole Face*. The DV was the accuracy of the composites, as measured by a Likert scale of 1-7.

Participants. Participants were 30 adults (12 females, 18 males) between the ages of 18 and 59 (M = 30.13, SD = 12.56). An equal number of participants were randomly allocated to each level of *Task* (N = 10). Participants were recruited on the basis that they were not familiar with the targets. To be deemed unfamiliar, participants must be unable to name 80% of the targets. No participants failed this a priori rule.

Materials. Facial composites constructed in Part 1 were displayed using a PowerPoint presentation. Each slide contained a target image and four composites constructed of the target, one from each condition. Multiple versions of the PowerPoint presentation were created, displaying the order of the slides and the order of the composites on each slide randomly.

Procedure. Participants viewed the PowerPoint presentation via the 'screen share' feature on Skype. For participants rating the face *Shape*, the researcher explained that face shape meant the shape of the features and the head shape as well as the distance between the features. There was no mention of the face texture for participants in this condition to avoid bringing participants' attention to the texture of the face, increasing the likelihood that participants focus on the face shape. For participants rating the face texture, the researcher explained that face texture referred

to the colour tones of the face. There was no mention of the face shape for participants in this condition to avoid bringing participants' attention to the face shape. For participants rating the whole face, there was no explanation for face shape or texture. Participants in each condition verbally rated each composite image using the seven-point Likert scale. Participants were instructed to inform the researcher if they recognised any of the target images.

Part 2c - Intermediate Composite Rating

Design. Experiment 4 utilised a within-subjects design with two IVs: One IV was *Screens* with four levels: *Two Screens* for *Shape* and *Texture; Two Screens* for *Shape* and *Four Screens* for *Texture; Four Screens* for *Shape* and *Two Screens* for *Texture; Four Screens* for *Shape* and *Texture.* The second IV was the stage of composite construction (referred to as *Stages*) with five levels: *Random Face, First Generation, Second Generation, after Holistic Tools* and *Final Image*. The DV was the accuracy of the composites, measured using the same Likert scale as Part 2b.

Participants. Participants were 18 adults (8 female, 10 male) between 19 and 48 years (M = 29.94, SD = 8.19). Participants were recruited on the basis that they were not familiar with the targets.

Materials. As in the previous experiments, images of the composite at each level of *Stages* were displayed using a PowerPoint presentation alongside a random composite image and a photograph of the target. Multiple versions of the PowerPoint were created, displaying the order of the images and the order of the slides randomly.

Procedure. This procedure replicated that from Part 2b, only using different materials.

Results

Part 2a – Composite Naming

To test the hypothesis that face *Shape* is more important than face *Texture* during composite construction, responses to facial composites and target pictures were scored for accuracy as in the previous experiments. As per the *a priori* rule, data were only included when participants correctly named at least eight of the 10 target pictures. Incorrect names for target pictures occurred 31 times (by 22 participants), four to 10 times by group. As such, the mean correct naming for target pictures was very high (M = 93.00%, SD = 8.30%). Where a target had not been named correctly, the associated composites also could not be named correctly, so responses to these composites were removed prior to analysis.

As expected, correct responses were much lower overall for spontaneous naming of facial composites (M = 40.42%, SD = 49.11%) compared to naming of target photographs. Table 16 displays the mean correct naming of composites constructed using *Two* or *Four Screens* for selection of face *Shape* and *Texture*.

Table 16. Correct Naming of Composites by Shape and Texture.

		Texture	
	Screens	2	4
Shape [a]	2	54.39 (78 / 142)	45.89 (67 / 146)
	4	33.33 (47 / 141)	27.14 (38 / 140)

Note. Figures are expressed in percentage and calculated from participant responses in parentheses: summed correct responses (numerator) and total (correct and incorrect) responses (denominator). Data are presented for composites for which participants correctly named the relevant target photographs (N = 569 out of 600). [a] $p < .05$.

The pattern of results in Table 17 demonstrates a large difference in correct composite naming between each level of *Four*. Overall, composites constructed using fewer screens were the most accurate, with composites constructed using *Two Screens* for *Shape* and *Texture* named correctly twice as frequently as those constructed using *Four Screens* for *Shape* and *Texture.* Individual responses were analysed using GLMM. This experiment involved two fixed effects, *Shape* and *Texture* (coded as 2 = *Two Screens*; 4 = *Four Screens*), both specified to have an ascending sorting order, plus an intercept. The method of analysis replicated that of Experiment 1.

A hypothesis-testing (confirmatory) approach was followed that comprised three models, each specified with different fixed effects (predictors) along with appropriate random intercepts (as described above). One model contained *Shape* only and a second contained *Texture* only. A third model contained the interaction between these two fixed effects. As it is standard practice to include the individual predictors in a model that contains their interaction (e.g., Field, 2018), this third model was full factorial. Based on the usual alpha of .1 for regression analyses, the model for *Shape* only was significant [F (1, 567) = 9.75, p = .002], while neither the model for *Texture* only [F (1, 567) = 0.78, p = .38] nor the interaction between *Shape* and *Texture* was significant [F (1, 565) = 0.17, p = .68]. Therefore, the model for *Shape* only became the final model and was the focus of analysis.

To explore the significant effects, fixed coefficients were examined (Table 18).

Table 18. Model Parameters for Effects of Number of Shape Screens on Composite Naming. Comparisons are Presented with Reference to the Lowest Category- Four Screens; Positive Values of B Indicate Higher Naming with Respect to The Reference.

	B	*SE(B)*	*t (567)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
Shape							
2 vs. 4	0.99	0.32	3.12	.002	2.68	1.44	5.01

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix)] final, robust-based full factorial Corrected model [F (1, 567) = 9.75, p = .002]. The model was specified with the lowest category of categorical predictors as reference (*Shape*; *4 Screens*), and predictors were sorted in an ascending order. Information criteria are based on the -2log likelihood (*AICc* = 2516.18, *BIC* = 2524.84). Variance of random slopes intercept of item for *Screens* [0.30, *SE* = 0.29, *Z* = 1.04, *p* = .30, *CI* (0.05, 1.97)].

This analysis demonstrated a significant difference in composite accuracy depending on the number of screens used to select the face *Shape*, whereby reducing the number of *Shape Four* from 4 to 2 is beneficial for composite construction.

Part 2b - Composite Final Image Rating

Part 2b tests the hypothesis that manipulating the number of screens used to select the face *Shape* and *Texture* has an impact on the accuracy of the composite *Shape* and *Texture.* To test this hypothesis, composite images were rated for accuracy in comparison to the target photograph using a Likert scale of 1-7 based on the *Shape, Texture,* or the *Whole Face.* Table 19 presents the mean and standard errors for likeness ratings of composites constructed using *Two* or *Four Screens* of *Shape* and *Texture*, rated based on the *Face Shape*, *Texture* or the *Whole Face*.

Table 19. Mean Composite Likeness Ratings of Face Shape, Texture and Whole Face for Each Level of Shape and Texture

Screens		Task			
Shape	Texture	Face Shape	Face Texture	Whole Composite	Mean
2	2	4.36	4.15	4.37	4.29
		(0.15)	(0.13)	(0.15)	(0.08)
	4	3.89	4.01a	4.29	4.06
		(0.14)	(0.12)	(0.15)	(0.08)
4	2	3.92	3.91[a]	4.24	4.02
		(0.14)	(0.13)	(0.15)	(0.08)
	4	3.97	4.60 [a]	4.15	4.24
		(0.13)	(0.13)	(0.16)	(0.08)
Mean		4.04	4.17	4.26	4.16
		(0.07)	(0.07)	(0.08)	(0.04)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant SE of the mean. [a]$p = < .05$.

Table 19 showed that face *Shape*, *Texture* and whole composite accuracy differ at each level of *Shape* and *Texture*. Mean likeness ratings for composites constructed using the same number of screens for selection of *Shape* and *Texture* are very similar, as are likeness ratings for composites constructed using a different number of screens to select the *Shape* and *Texture.* Furthermore, composites constructed using the same number of screens for selection of *Shape* and *Texture* are rated considerably higher than those constructed using a different number of screens. To further understand the impact that *Shape* and *Texture* have on composite accuracy, the data were analysed using GLMM. As in previous experiments, a hypothesis testing approach was followed, which comprised separate models for each predictor:

Shape, Texture and *Task,* as well as full-factorial models for each interaction: *Shape*Texture, Shape*Task, Texture*Task and Shape*Texture*Task*.

The model for *Shape* only [F (1, 1193) = 0.05, p = .83], *Texture* only [F (1, 1193) = 0.08, p = .78] and *Task* only [F (2, 1192) = 0.25, p = .78] were not significant. The interactions between *Shape* and *Texture* [F (1, 1191) = 0.07, p = .79], *Shape* and *Task* [F (2, 1189) = 0.26, p = .77] and *Texture* and *Task* interaction [F (2, 1189) = 0.43, p = .65] were also not significant. Nonetheless, the three-way interaction between *Shape, Texture* and *Task* was significant [F (2, 1183) = 3.20, p = .041], making this the final model. Fixed coefficients were examined to explore the significant three-way interaction.

As the three-way interaction between *Shape, Texture* and *Task* was significant, it is important to analyse the likeness ratings of composites constructed using *Two* or *Four Screens* to select the face *Shape* and *Texture* when rated based on the *Shape,* the *Texture* or the *Whole Face.* The aim of this thesis is to determine the optimum number of screens during EvoFIT composite construction. Therefore, the analysis did not compare composites at each level of *Task,* as this would not be beneficial in achieving the experimental aim but would instead demonstrate whether composites are rated higher based on the *Shape*, the *Texture* or the whole face.

For composites rated based on the face *Shape,* there was no significant interaction between *Shape* and *Texture* [F (1, 391) = 1.32, p = .25]. For composites rated based on the face *Texture*, there was a significant *Shape*Texture* interaction [F (1, 391) = 3.84, p = .051], as composites constructed using *Four Screens* to select the face *Shape* were more accurate when *Four Screens* were also used for *Face Texture, compared* to *Two Screens* (p = .019, *Exp(B)* = 2.95). However, for composites rated

based on the *Whole Face,* there was also no significant *Shape*Texture* interaction [$F(1, 391) = 0.00$, $p = .98$].

This pattern of results indicates that there is little difference in likeness between composite images created using a different number of screens for selection of the face *Shape* and *Texture*. However, the results do indicate that composites created using the same number of screens may have a more accurate facial *Texture*, indicated by the higher likeness ratings for composites created using *Four Screens* for selection of the *Shape* and *Texture* ($M = 4.60$) compared to those created using *Four Screens* for *Shape* and *Two Screens* for *Texture* ($M = 3.91$) when rated based on the *Texture.*

Part 2c - Intermediate composite rating

To continue exploring the impact of cognitive load on composite accuracy, composites at four different stages of construction (*First Generation, Second Generation, Holistic Tools* and the *Final Image*) as well as a *Random Face,* were compared to the target face and rated for likeness against the corresponding target photograph. As in Experiments 1-3, Part 2c also tests the hypothesis that reducing the number of screens during composite construction is beneficial for composite accuracy, and that manipulating the number of screens impacts the ability to utilise each stage of construction (shown by reduced composite accuracy in the final stages of construction for composites constructed using more screens). The mean likeness ratings of composites constructed using *Two* or *Four Screens* of *Shape* and *Texture*, and at each stage of composite construction are presented in Table 20.

Table 20. Composite Likeness Ratings at each Level of Screens and Stage

Shape	Stage[1]	Texture[1]		Mean
		2	4	
2	Random	2.75[a]	2.61[d]	2.68[g]
		(0.12)	(0.10)	(0.08)
	1st Generation	3.01	2.76	2.88
		(0.11)	(0.09)	(0.07)
	2nd Generation	3.45	3.19	3.32
		(0.10)	(0.09)	(0.07)
	Holistic Tools	3.98	3.68	3.83
		(0.10)	(0.09)	(0.07)
	Final Image	4.15[a]	4.06[d]	4.11[g]
		(0.11)	(0.10)	(0.08)
4	Random	2.37[b]	2.38[e]	2.38[h]
		(0.09)	(0.10)	(0.07)
	1st Generation	3.06	2.81	2.93
		(0.10)	(0.10)	(0.07)
	2nd Generation	3.38	3.21	3.29
		(0.10)	(0.10)	(0.07)
	Holistic Tools	3.58	3.67	3.63
		(0.10)	(0.10)	(0.07)
	Final Image	4.01[b]	3.86[e]	3.93[h]
		(0.11)	(0.11)	(0.08)
Mean	Random	2.56[c]	2.50[f]	2.53[i]
		(0.08)	(0.07)	(0.05)
	1st Generation	3.03	2.78	2.91
		(0.07)	(0.07)	(0.05)
	2nd Generation	3.41	3.20	3.31
		(0.07)	(0.07)	(0.05)

Holistic Tools	3.78	3.68	3.73
	(0.07)	(0.07)	(0.05)
Final Image	4.08[c]	3.96[f]	4.02[i]
	(0.08)	(0.08)	(0.02)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant-item SE of the mean. Exp(B): [a] = 8.96, [b] = 21.05, [c] = 12.81, [d] = 12.42, [e] = 17.34, [f] = 13.34, [g] = 10.57, [h] = 19.32, [i] = 13.67. [l] $p < .001$.

Table 20 demonstrated that composite internal features increased at each stage of composite construction, despite the number of screens used to select the face *Shape* and *Texture.* Mean composite ratings show a considerable, and very similar increase in likeness ratings between each level of *Stage*, indicating that each stage of construction may be equally important for creating an accurate composite image. To further understand the impact that *Shape* and *Texture* have on the accuracy of composite internal features throughout the stages of composite construction, the data were analysed using GLMM. As in previous experiments, a hypothesis testing approach was followed, which comprised separate models for each predictor: *Shape, Texture* and *Stage,* as well as a full-factorial model for each interaction: *Shape*Texture, Shape*Stage, Texture*Stage and Shape*Texture*Stage*.

The model for *Shape* only was not significant [F (1, 3594) = 1.38, p = .24], but the models for *Texture* only [F (1, 3594) = 11.81, $p < .001$] and *Stage* only [F (4, 3591) = 22.15, $p < .001$] were significant. Furthermore, neither the interactions between *Shape* and *Texture* [F (1, 3592) = 1.68, p = .20], *Shape* and *Stage* [F (4, 3586) = 1.75, p = .14] and *Texture* and *Stage* [F (4, 3586) = 0.84, p = .50] were significant, nor was the three-way *Shape*Texture*Stage* interaction [F (4, 3576) = 0.91, p = .46]. Therefore, a further model was run, containing the significant predictors: *Texture* and *Stage.* As this model was significant for both predictors, it

was taken as the final model. To explore the significant predictors, fixed coefficients were examined (Table 21).

Table 21. Model Parameters from Composite Likeness Ratings from the Texture and Stage.

	B	*SE(B)*	*t (3590)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
Texture Screens							
Texture: 2 vs Texture: 4	0.24	0.07	3.26	< .001	1.27	1.10	1.47
Stage							
Final Image vs Random Face	2.62	0.31	8.35	< .001	13.79	7.45	25.53
Holistic Tools vs Random Face	2.10	0.31	6.71	< .001	8.19	4.43	15.15
2nd Generation vs Random Face	1.40	0.31	4.47	< .001	4.05	2.19	7.47
1st Generation vs Random Face	0.74	0.31	2.36	.018	2.09	1.13	3.85

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix)] final, robust-based full factorial Corrected model [F (5, 3590) = 19.77, p < .001]. The model was specified with the lowest category of categorical predictors as reference and predictors were sorted in an ascending order. Information criteria are based on the -2log likelihood (*AICI* = 69467.20, *BIC* = 69504.29). Variance of random slopes intercept of participants [1.00, *SE* = 0.39, *Z* = 2.54, *p* = .011, *CI* (0.46, 2.16)] and item [0.17, *SE* = 0.09, *Z* = 1.75, *p* = .08, *CI* (0.05, 0.51)] for *Texture* and *Stage.*

GLMM demonstrated a considerable significant increase in ratings of composite internal features from composite images constructed using *Four Screens*

for *Texture* to *Two Screens*, indicating that fewer screens were needed to accurately select the face *Texture* for a facial composite image. There was also a substantial, significant increase in ratings of composite internal features from *Random Faces* to the *First Generation, Second Generation, After Holistic Tools* and *Internal Features.* Furthermore, the effect size between *Random Faces* and each level of *Stage* increases as composite images move through the construction process.

To further understand the mathematical pattern of the data, polynomial contrasts were conducted, focusing on linear and quadratic patterns. A single model exploring a linear pattern in the data was significant ($p < .001$, *Exp(B)* = 1.92), demonstrating that composite accuracy increases between each level of construction. A single model exploring a quadratic pattern was not significant ($p = .65$, *Exp(B)* = 0.98), demonstrating that composite accuracy increases at a similar rate between each stage of construction. The final model contained analysis for a linear, quadratic, cubic and quartic pattern, with only a linear pattern being significant ($p < .001$, *Exp(B)* = 1.97).

As face images were rated within the context of other faces, i.e., composites from each stage of construction were displayed on the screen simultaneously, the 'best' composites were those with the largest difference in rating between the least accurate point (*Random Face)* and the most accurate point (*Final Image).* Therefore, GLMM was also used to compare the difference in composite rating between *Random Faces* and the *Final Image* for composites constructed at each level of *Shape* and *Texture*. The *Final Image* was significantly more accurate than the *Random Face* for composites constructed using *Two Screens* for *Shape* and *Texture* ($p < .001$, *Exp(B)* = 8.96), *Two Screens* for *Shape* and *Four Screens* for *Texture* ($p < .001$, *Exp(B)* = 12.42), *Four Screens* for *Shape* and *Two Screens* for *Texture* ($p < .001$, *Exp(B)* =

21.05), and *Four Screens* for *Shape* and *Texture* ($p < .001$, *Exp(B)* = 14.61). Although there was a significant difference between the *Random Face* and *Final Image* in all conditions, the largest effect size was found for composites constructed using *Four Screens* for *Shape* and *Two Screens* for *Texture.* This result may indicate that composites in this condition were deemed to be the most accurate in comparison to the target photograph.

Discussion

This experiment examined the importance of face *Shape* and *Texture* during the creation of an EvoFIT composite. The first hypothesis was that composites constructed using fewer screens to select the face *Shape* would be the most accurate, as face *Shape* is deemed most important for the recognition of unfamiliar faces (Johnston & Edmonds, 2009; Hancock et al., 2000), meaning that fewer face options will be needed for accurate selection of the face *Shape*. This hypothesis was supported by the results of this experiment, which demonstrate that facial composites constructed using *Two Screens* to select the face *Shape* are more accurate than composites constructed using *Four Screens* to select the face *Shape*. This result indicates that few screens are needed to select the face *Shape* to produce an accurate facial composite image.

Composite naming in Part 2a of this experiment demonstrated that composites constructed using *Two Screens* to select the face *Shape* were named more frequently than composites constructed using *Four Screens* to select the face *Shape*, despite the number of screens used to select the face *Texture*. Furthermore, the difference in correct naming rate between composites constructed using *Two* or *Four Screens* to select the face *Shape* was significant, whereas the difference in the correct naming

rate between composites constructed using *Two* or *Four Screens* to select the face *Texture* was not significant. This finding highlights the importance of face *Shape* over face *Texture* for unfamiliar face recognition, as manipulating the number of face arrays viewed had a larger impact on the face *Shape* than it did on the face *Texture*.

In Part 2b, participants were invited to rate how alike an unfamiliar facial composite was to the target based on the face *Shape*, the face *Texture*, and the whole face. The aim of this measure was to assess whether the number of screens used to select the face *Shape* and *Texture* has an impact on the accuracy of the composite face *Shape* and *Texture*. The results demonstrated that there was little difference in the likeness between composites constructed using *Two* or *Four Screens* to select the face *Shape* and *Texture.* A significant three-way interaction between *Shape, Texture* and *Task* indicated that, when *Four Screens* are used for selection of the face *Shape*, the face *Texture* is more accurate when created using *Four Screens* as opposed to *Two Screens.* This pattern of results indicates that a heavy cognitive load may have less impact on the accuracy of the face *Texture* than it does on the face *Shape*. This finding may be explained by the use of the *Shape Tool* and *Holistic Tools* towards the end of the construction procedure, both of which focus on changing the shape of the face, allowing for alteration of the size, shape and position of facial features. When cognitive load is heavier during the construction procedure, the ability to accurately enhance the face shape is lower, which explains the higher face shape likeness ratings for composites constructed using *Two Screens* than *Four* Screens. However, as the face *Texture* is altered less during use of the *Shape Tool* and *Holistic Tools*, when the effects of cognitive overload may inhibit the ability to make accurate changes to the face, likeness ratings based on the face *Texture* are higher for composites constructed using *Four Screens* than *Two* Screens.

The second hypothesis was that composites constructed using fewer screens overall would be more accurate than those constructed using more screens. In Part 2a, composites constructed using *Two Screens* to select the face *Shape* and *Texture* were named twice as frequently as composites constructed using *Four Screens* to select the face *Shape* and *Texture*. Of note, composites constructed using *Two Screens* to select face *Shape* and *Texture*, which is the lowest number of screens used in this experiment, were also named more frequently than composites constructed using either *Two Screens* to select the *Shape* and *Four Screens* to select the *Texture*, or *Four Screens* to select the *Shape* and *Two Screens* to select the *Texture*. This finding indicates that facial composite construction may be optimal when the same number of screens are viewed for selection of the face *Shape* and *Texture*, as opposed to viewing a different number of screens for selecting the *Shape* and *Texture*.

In Part 2b, composites constructed using *Two Screens* to select the face *Shape* and *Texture* were rated higher than composites constructed using *Four Screens* to select the face *Shape* and *Texture*. As in Part 2a, composites constructed using *Two Screens* to select the face *Shape* and *Texture* were also rated higher than composites constructed using either *Two Screens* to select the *Shape* and *Four Screens* to select the *Texture* or *Four Screens* to select the *Shape* and *Two Screens* to select the *Texture*. In this experiment, composites constructed using *Two Screens* for selection of the face *Shape* and *Texture* have the lowest number of interacting elements and therefore should have the lowest intrinsic cognitive load (Sweller, 1988). It may be fair to assume that composite construction utilising more than *Two Screens* for the selection of face *Shape* or *Texture* may result in cognitive overload, reducing the abilities of participants to accurately create a facial composite image.

The third hypothesis was that cognitive load would impact the ability to utilise each stage of composite construction. This hypothesis is based on the assumption that reducing the cognitive load in the early stages of composite construction will allow participants to enhance the face more accurately in the later stages. The pattern of results in Part 2c demonstrated that, in all conditions, composite accuracy increased between each stage of construction, supported by the significant linear trend for *Stage.* Furthermore, there was no interaction between the number of screens used and the stage of construction, indicating that cognitive load may not have an impact on participants' ability to use each stage of composite construction. This pattern of results does not support the hypothesis that cognitive load will impact the ability to utilise each stage of composite construction. One explanation for this result may be that the cognitive load experienced by participants during the construction procedure was not so great that they were unable to utilise the image enhancement tools. Alternatively, the memory capacity that is needed to utilise the image enhancement tools may not be so great that participants who *are* experiencing cognitive load cannot complete this stage of the procedure well. However, as the experiments in this thesis are the first to explore the likeness of EvoFIT facial composites at each stage of the construction procedure, there is no literature to support these findings. To continue to develop an understanding of the impact of cognitive load throughout the construction procedure, this measure will also be carried out in the next experiment.

Limitations and Future Research

A limitation of this experiment is that the construction method, that is, viewing a different number of face images to select the face *Shape* and *Texture*, does not generalise to other composite systems. The closest system to EvoFIT is EFIT-6, an

evolutionary facial composite system that aims to create realistic and accurate facial composites based on the witness' ability to perform various facial processing tasks (George et al., 2008). However, one major difference between the two systems mentioned is their representation of face *Shape* and *Texture*. As discussed, EvoFIT invites eyewitnesses to select the face *Shape* and face *Texture* independently, whereas E-FIT6 combines the face *Shape* and *Texture* into a single representation referred to as the global appearance model (Valentine et al., 2010). Therefore, any findings that depend on viewing the face *Shape* and *Texture* independently cannot be generalised to EFIT-6. Although all experiments in this thesis are conducted using EvoFIT, it is still important to recognise that contributing to the general facial composite literature is superior to contributing only to the literature on EvoFIT. Therefore, although the findings of Experiments 1-3 can be generalised to other composite systems, the findings of the current experiment cannot.

A further limitation of this experiment is that it is unlikely to determine the optimum number of screens during composite construction. The first three experiments in this thesis demonstrated that reducing the number of screens to just *One Screen* produced the most accurate composite images. However, the lowest number of screens used in this experiment was *Two Screens.* The choice to reduce the number of screens to only *Two* was made so that the difference between the highest number of screens used (*Four*) and the lowest number of screens used (*Two*) would not be so huge that participants constructing a composite using a different number of screens to select the face *Shape* and face *Texture* would not presume that either face *Shape* or face *Texture* was most important to composite construction. If participants created a facial composite using *Two Screens* to select the face *Shape* and *Four Screens* to select the face *Texture* or *Four Screens* to select face *Shape* and *Two*

Screens to select the face *Texture*, they may have presumed that *Shape* and *Texture* were not of equal importance. Participants may therefore concentrate on selecting one element of the face (either *Shape* or *Texture*) more accurately than the other.

Although it was sensible to display *Two* and *Four Screens* instead of *One* and *Four Screens*, it also means that the findings from this experiment are unlikely to determine the optimum number of screens during composite construction, which is the overall aim of this thesis. Therefore, the next experiment will further reduce the number of screens during EvoFIT construction so that participants view *One Screen* to select the *Shape* or *Texture*, or *Two Screens* to select the *Shape* or *Texture.* This experiment will further determine the importance of face *Shape* and *Texture* during the construction of an EvoFIT composite, with an increased likelihood of determining the optimum number of screens during the construction procedure.

This experiment was crucial to understanding the impact that reducing the number of screens used to select face *Shape* and *Texture* had during EvoFIT composite construction. The results of this experiment demonstrated that reducing the number of screens used to select the face *Shape* was beneficial for composite construction, but that reducing the number of screens used overall was more beneficial. The next chapter will present the method and results for Experiment 5, which aims to replicate the current experiment and extend the current experiment by reducing the number of screens used for composite construction.

7

EXPERIMENT 5

FURTHER REDUCING THE POPULATION SIZE DURING EVOFIT CONSTRUCTION FOR FACE SHAPE AND TEXTURE INDEPENDENTLY

Abstract

This experiment aimed to replicate the fourth experiment, but with one crucial difference, reducing the number of screens used to select the face *Shape* or the face *Texture* to *One Screen*, further reducing the cognitive load in comparison to Experiment 4. Forty participants created facial composites using *One* or *Two Screens* to select the *Shape* and *Texture*. These composite images were judged for accuracy using composite naming by 60 participants and target likeness by 45 participants. The results demonstrated that composites constructed using *One Screen* to select the face

Shape were more accurate than composites constructed using *Two Screens* to select the face *Shape*, supporting the first hypothesis. Furthermore, composites constructed using fewer screens overall were the most accurate.

During the typical EvoFIT composite construction procedure, *Four Screens* are used to select the face *Shape* and *Texture,* yet Experiments 1-3 found that composites were the most accurate when created using just *One Screen*. Composite construction using *One Screen* to select the face *Shape* and *Texture* invites participants to select six face images from one face array. On the other hand, composite constructed using *Four Screens* invites participants to select two face images from *Three Screens* (resulting in six faces selected) and to swap any face images that are selected on the fourth screen.

Experiment 4 manipulated the number of screens viewed during selection of the face *Shape* and *Texture* individually, with the number of screens used to select face *Shape* or *Texture* reduced from *Four* to *Two*, so that participants may view *Four Screens* to select the face *Shape* (selecting two faces from *Three Screens* and swapping any faces on the fourth screen) and *Two Screens* to select the face *Texture* (selecting three faces from *Two Screens*). As Experiments 1-3 demonstrated that the most accurate facial composites were constructed using *One Screen,* it is crucial to reduce the number of screens used during EvoFIT composite construction to *One Screen* for selection of face *Shape* or face *Texture* to understand the optimal number of screens during EvoFIT composite construction.

However, if participants viewed *One Screen* to select the face *Shape* but *Four Screens* to select the face *Texture*, they may have presumed that there was a difference in the importance of face *Shape* and *Texture*, which may have influenced

their choices, for example, selecting faces for *Texture* more carefully. Therefore, in the current experiment, participants will be invited to create a facial composite by selecting the face *Shape* from *One* or *Two Screens* and the face *Texture* from *One* or *Two Screens*. So, in this experiment, participants view *One Screen* for face *Shape* and *Two Screens* for face *Texture*, *Two Screens* for face *Shape* and *One Screen* for face *Texture*, *One Screen* for face *Shape* and *One Screen* for face *Texture*, or *Two Screens* for face *Shape* and *Two Screens* for face *Texture*.

Experimental Aims

Following on from Experiment 4, this experiment aimed to optimise EvoFIT composite construction by reducing the number of screens used to select the face *Shape* and *Texture*. In line with the results from Experiment 4, this experiment hypothesises that:

H1: Composites constructed using fewer screens to select the face *Shape* will be the most accurate, as face *Shape* is deemed most important for the recognition of unfamiliar faces, meaning that fewer face options will be needed for accurate construction.

H2: Composites constructed using fewer screens overall are more accurate than those constructed using more screens.

H3: Composites constructed using fewer screens will increase in accuracy more after the use of image enhancement tools towards the end of the construction procedure.

Method

As in the previous experiments, Experiment 5 is separated into four parts. In Part 1, composites were constructed via EvoFIT using *One* or *Two Screens* of face *Shape* and *Texture*, replicating the procedure of Experiment 4 but with a reduced number of screens. Furthermore, construction using *Two Screens* is utilised as the baseline as opposed to *Four Screens*. If composite construction uses *One Screen* to select the face *Shape* but *Four Screens* to select the face *Texture*, participants may assume that there is a difference in the importance of face *Shape* and face *Texture* for composite construction, which may invalidate the results. Consequently, a decision was made to use *Two Screens* as the baseline (cf. *Four Screens*), as this may be viewed as a reasonable change by participants. As in previous experiments, in Part 2a, composite naming is attempted, in Part 2b, composite face *Shape*, *Texture* or *Whole Faces* were rated for likeness against target images, and in Part 2c, composites were rated for likeness against the target at different stages in the composite construction process. In Parts 1, 2b and 2c, participants were recruited on the basis that they were not familiar with the targets. In Part 2a, participants were recruited on the basis that they were familiar with the targets. As an a priori rule, participants who could identify 80% of the target photographs were deemed familiar with the targets.

Part 1- Composite Construction

Design. As in the previous experiment, a 2x2 factorial design was used whereby the two factors were *Shape* and *Texture,* and the two levels were *One Screen* and *Two Screens*. Therefore, composites were constructed using *One Screen* to select the face *Shape* and face *Texture* (1 x 1 condition), *One Screen* of face *Shapes* and *Two Screens* of face *Textures* (1 x 2 condition), *Two Screens* of face *Shapes* and *One Screen* of

Face Textures (2 x 1 condition), or *Two Screens* of face *Shapes* and *Textures* (2 x 2 condition). The experiment was designed to reflect the procedures used by the police, including a 24-hour delay between the participant viewing the target photograph and composite construction. Participants were required to use a PC and have access to a video conferencing platform, such as Skype or Microsoft Teams.

Participants. Participants were 40 adults (27 females, 13 males) aged between 18 and 43 years (M = 26.03, SD = 7.28). Participants were recruited on the basis that they were not familiar with the targets.

Materials. The target stimuli were 10 photographs of Emmerdale characters. Five of the characters were male, and five were female. Each image was displayed in colour and was approximately 8cm (wide) x 10cm (high). None of the targets had particularly distinctive characteristics that would make them easier to identify, potentially reducing experimental power.

Procedure. The procedure replicated that of Experiment 4; however, the number of screens viewed containing face *Shape* and face *Texture* was dependent on the condition.

Part 2a- Composite naming

Design. The same factorial design was used as Part 1, with *Shape* and *Texture* as factors and the two levels: *One Screen* and *Two Screens*. The DV was *Naming* with two levels: *Spontaneous* and *Cued.* Participants were recruited on the basis that they were not familiar with the targets. To be deemed unfamiliar, participants must be unable to name 80% of the targets; no participants failed this a priori rule.

Participants. Participants were 60 adults (38 females and 22 males) between 18 and 59 years (M = 40.31, SD = 12.61). An equal number of participants were randomly allocated to each level of *Screens* (N = 15).

Materials. As in the previous experiments, composites constructed during Part 1 of the experiment were displayed using four PowerPoint presentations (one for each condition). A fifth presentation was created containing the target photographs.

Procedure. The procedure replicated that of Experiment 1, Part 2a.

Part 2b - Composite Final Image Rating

Design. A mixed design was utilised with two IVs. *Screens* was a within-subjects variable with four levels: *One Screen* for *Shape* and *Texture; One Screen* for *Shape* and *Two Screens* for *Texture; Two Screens* for *Shape* and *One Screen* for *Texture; and Two Screens* for *Shape* and *Texture. Task* was a between-subjects variable with three levels: rating of face *Shape, Texture* or *Whole Face*. The DV was the accuracy of the composites, as measured by a Likert scale of 1-7.

Participants. Participants were 30 adults (14 females, 17 males) between the ages of 18 and 57 (M = 26.23, SD = 10.82). An equal number of participants were randomly allocated to each level of *Task* (N = 10). Participants were recruited on the basis that they were not familiar with the targets. To be deemed unfamiliar, participants must be unable to name 80% of the targets; no participants failed this a priori rule.

Materials. Facial composites constructed in Part 1 were displayed in a PowerPoint presentation. Each slide contained a target image and four composites constructed of the target, one from each condition. Multiple versions of the

PowerPoint presentation were created, displaying the order of the slides and the order of the composites on each slide randomly.

Procedure. The procedure replicated that of Experiment 4, Part 2b.

Part 2c - Intermediate Composite Rating

Design. Experiment 5 utilised a within-subjects design with two IVs: One IV was *Screens* with four levels: *One Screen* for *Shape* and *Texture; One Screen* for *Shape* and *Two Screens* for *Texture; Two Screens* for *Shape* and *One Screen* for *Texture; and Two Screens* for *Shape* and *Texture.* The second IV was the stage of composite construction (referred to as *Stages*) with five levels: *Random Face, First Generation, Second Generation, after Holistic Tools* and *Final Image*. The DV was the accuracy of the composites, measured using the same Likert scale as Part 2b.

Participants. Participants were 15 adults (6 female, 9 male) between 18 and 53 years (M = 24.53, SD = 10.94). Participants were recruited on the basis that they were not familiar with the targets.

Materials. As in the previous experiments, images of the composite at each level of *Stages* were displayed in a PowerPoint presentation alongside a random composite image and a photograph of the target. Multiple versions of the PowerPoint were created, displaying the order of the images and the order of the slides randomly.

Procedure. This procedure replicated that from Part 2a, only using different materials.

Results

Part 2a- Composite naming

To test the hypothesis that face *Shape* is more important than face *Texture* during composite construction, responses to facial composites were scored for accuracy, as in the previous experiments. As per the *a priori* rule, data were only included from participants who correctly named at least eight of the 10 target pictures. Incorrect names for target pictures occurred 27 times (by 20 participants), three to seven times by group. As such, the mean correct naming for target pictures was very high (M = 95.50%, SD = 20.75%). Where a target had not been named correctly, the associated composites also could not be named correctly, so responses to these composites were removed prior to analysis. As expected, correct responses were much lower overall for spontaneous naming of facial composites (M = 55.23%, SD = 49.76%). Table 22 displays the mean correct naming for composites constructed using *Two* or *Four Screens* for selection of *Face Shape* and *Texture.*

Table 22. Correct Naming of Composites by Shape and Texture.

		Texture	
	Screens	1	2
Shape [a]	1	63.70 (93 / 146)	60.14 (86 / 143)
	2	53.47 (77 / 144)	43.26 (61 / 141)

Note. Figures are expressed in percentage and calculated from participant responses in parentheses: summed correct responses (numerator) and total (correct and incorrect) responses (denominator). Data are presented for composites for which participants correctly named the relevant target photographs (N = 364 out of 400). [a] $p < .05$.

Table 22 demonstrates that composites constructed using *One Screen* for *Shape* and *Texture* are named correctly more frequently than composites in any other condition. However, composites constructed using *One Screen* for *Shape* and *Two Screens* for *Texture* follow closely, with little difference in correct naming between the two conditions. On the other hand, there is a considerable reduction in correct naming between composites constructed using *One Screen* for *Shape* and *Two Screens* for *Texture* and those constructed using *Two Screens* for *Shape* and *One Screen* for *Texture,* with an even larger reduction for *Four Screens* for *Shape* and *Texture.* Overall, composite naming rates were similar between the two levels of *Texture; however,* there was a large difference in composite naming between the two levels of *Shape.* Individual responses were analysed using GLMM. This experiment involved two fixed effects, *Shape* and *Texture* (coded as *1 = One Screen, 2 = Two Screens*), both specified to have an ascending sorting order, plus an intercept. The method of analysis replicated that of Experiment 4.

A hypothesis-testing approach was followed, comprising three models: one model contained *Shape* only, a second contained *Texture* only, and a third model contained the interaction between these two fixed effects. Based on the usual alpha of .1 for regression analyses, the model for *Shape* was significant [F (1, 572) = 8.66, p = .003], while the model for *Texture* [F (1, 572) = 2.61, p = .11] and the *Shape*Texture* interaction were not [F (1, 570) = 0.60, p = .44]. Therefore, the model for *Shape* became the final model, and the focus of analysis. To explore the significant effects, fixed coefficients were examined (Table 23).

Table 23. Model Parameters for Effects of Number of Screens on Composite Naming. Comparisons are presented with reference to the lowest category- Two Screens; Positive values of B indicate higher naming with respect to the reference.

	B	*SE(B)*	*t (572)*	*p*	*Exp(B)*	*95% CI (-)*	*95% CI (+)*
Shape							
1 vs. 2	0.66	0.22	2.94	.003	1.93	1.24	2.98

Note. GLMM [IBM SPSS (Version 28) using the GENLINMIXED procedure (see Appendix)] final, robust-based full factorial Corrected model [F (1, 572) = 8.666, p = .003]. The model was specified with the lowest category of categorical predictors as reference (*Shape; 2 Screens*), and predictors were sorted in an ascending order. Information criteria are based on the -2log likelihood (*AICc* = 2534.20, *BIC* = 2542.88). Variance of random slopes intercept of item for *Shape* [0.64, *SE* = 0.36, *Z* = 1.76, p = .08, *CI* (0.21, 1.95)].

As seen in Table 23, this analysis demonstrated a significant difference in composite accuracy depending on the number of screens used to select the face *Shape*, whereby reducing the number of *Shape Screens* from *Two* to *One* is beneficial for composite construction.

Part 2b - Final Composite Image Rating

To further understand whether manipulating the number of *Shape* and *Texture Screens* viewed during composite construction has a direct impact on composite accuracy, likeness ratings of composite face *Shape* and *Texture* were analysed. Table 24 presents the mean and standard errors for likeness ratings of composites constructed using *One* or *Two Screens* of *Shape* and *Texture*, rated based on the *Face Shape*, *Texture* or the *Whole Face*.

Table 24. Composite Likeness Ratings of Face Shape, Texture and Whole Face for Each Level of Shape and Texture

Screens		Rating Task			
Shape	Texture	Face Shape	Face Texture	Whole Composite	Mean
1	1	3.86	3.73	4.08	3.89
		(0.13)	(0.16)	(0.13)	(0.08)
	2	3.89	3.59	4.04	3.84
		(0.15)	(0.15)	(0.13)	(0.08)
2	1	3.90	3.62	4.08	3.87
		(0.15)	(0.15)	(0.13)	(0.08)
	2	3.52	3.52	3.82	3.62
		(0.16)	(0.16)	(0.15)	(0.09)
Mean		3.79	3.62	4.01	3.80
		(0.07)	(0.08)	(0.07)	(0.04)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant SE of the mean.

This table demonstrated that accuracy of composites constructed using *One* or *Two Screens* during composite construction is quite similar, with very little between the mean of the highest-rated condition (*One Screen* for *Shape* and *Texture*) and the mean of the lowest-rated condition (*Two Screens* for *Shape* and *Texture*). This table also indicates that accuracy of composites based on the *Face Shape, Face Texture* and *Whole Face* is quite similar, with little difference between the mean of the highest-rated condition (*Whole Face*) and the mean of the lowest-rated condition (*Face Texture*).

Individual responses were analysed using GLMM, a hypothesis texting approach comprised one model for each predictor: *Shape, Texture* and *Task* and one full-factorial model for each interaction between the three predictors: *Shape*Texture, Shape*Task, Texture*Task*, as well as a three-way interaction between *Shape, Texture* and *Task.* As all scores were roughly the same, none of the models that were run were significant. In more detail, there was no significant difference between the levels of *Shape* (F (1, 1195) = 0.79, p = .38], *Texture* [F (1, 1195) = 0.57, p = .45] or *Task* [F (2, 1194) = 0.89, p = .41]. There was also no significant *Shape*Texture* interaction [F (1, 11913 = 0.60, p = .44], *Shape*Task* interaction [F (2, 1191) = 0.02, p = .98], *Texture*Task* interaction [F (2, 1191) = 0.12, p = .89] or *Shape*Texture*Task* interaction [F (2, 1185) = 1.03, p = .36].

Part 2c - Intermediate Composite Rating

As in the previous experiments, likeness ratings of composites of four stages of composite construction (*First Generation, Second Generation, Holistic Tools* and the *Final Image*), as well as *Random Faces,* were compared to the target face and rated for likeness on a Likert scale of 1-7. Table 25 presents the mean (and standard errors) for likeness ratings of composites constructed using *One* or *Two Screens* of *Shape* and *Texture*, and at each stage of composite construction.

Table 25. Composite Likeness Ratings at each Level of Screens and Stage

Shape	Stage	Texture		Mean
		1	2	
1	Random	2.27[a]	2.46[b]	2.36[e]
		(0.09)	(0.10)	(0.07)
	1st Generation	2.82	2.99	2.90
		(0.10)	(0.10)	(0.07)
	2nd Generation	3.19	3.26	3.23
		(0.10)	(0.11)	(0.07)
	Holistic Tools	3.45	3.53	3.49
		(0.10)	(0.10)	(0.07)
	Final Image	3.63[a]	3.46[b]	3.54[e]
		(0.11)	(0.10)	(0.07)
2	Random	2.45[c]	2.58[d]	2.51[f]
		(0.10)	(0.10)	(0.07)
	1st Generation	3.17	2.99	3.08
		(0.10)	(0.10)	(0.07)
	2nd Generation	3.33	3.28	3.30
		(0.10)	(0.10)	(0.07)
	Holistic Tools	3.42	3.31	3.36
		(0.10)	(0.10)	(0.07)
	Final Image	3.34[c]	3.54[d]	3.44[f]
		(0.10)	(0.10)	(0.07)
Mean	Random	2.36[g]	2.52[h]	2.44[i]
		(0.07)	(0.07)	(0.05)
	1st Generation	3.00	2.99	2.99
		(0.07)	(0.07)	(0.05)
	2nd Generation	3.26	3.27	3.27
		(0.07)	(0.07)	(0.05)

Holistic Tools	3.44	3.42	3.43
	(0.07)	(0.07)	(0.05)
Final Image	3.48[g]	3.50[h]	3.49[i]
	(0.08)	(0.07)	(0.05)

Note. Rating scale (1 = very poor likeness … 7 = very good likeness). Values are expressed using the mean and, in parentheses, by-participant-item SE of the mean. Exp(B): [a] = 15.15, [b] = 6.80, [c] = 5.22, [d] = 6.85, [e] = 11.28, [f] = 6.41, [g] = 9.61, [h] = 7.14, [I] = 8.10.

Table 25 demonstrated that, in most cases, the accuracy of composite internal features increases at each stage of construction. Mean likeness ratings indicate that the difference in likeness between each stage of construction reduces throughout the construction process. More specifically, the difference in likeness ratings between *Random Faces* and the *First Generation* was relatively large, but the difference between the *First* and *Second Generation* was rather small. The difference between the *Second Generation* and *After Holistic Tools* was smaller still, and the difference between composites *After Holistic Tools* and the *Final Image* was very small. To understand the impact that *Shape* and *Texture* have on the accuracy of composite internal features throughout the stage of composite construction, the data were analysed using GLMM. As in the previous experiments, a hypothesis testing approach was followed, which comprised separate models for each predictor: *Shape, Texture* and *Stage,* as well as a full-factorial model for each interaction: *Shape*Texture, Shape*Stage, Texture*Stage and Shape*Texture*Stage*.

The model for *Shape* only was not significant [F (1, 2993) = 0.01, p = .92], nor was the model for *Texture* only [F (1, 2993) = .013, p = .91], but the model for *Stage* only was significant [F (4, 2990) = 4.48, p < .001]. The interaction between *Shape* and *Texture* was not significant [F (1, 2991) = 0.30, p = .59], neither was the

interaction between *Texture* and *Stage* [F (4, 2985) = 0.55, p = .70], nor the three-way interaction between *Shape, Texture* and *Stage* [F (4, 2977) = 3.25, p = .011]. However, the interaction between *Shape* and *Stage* [F (4, 2985) = 2.37, p = .051] was significant.

GLMM was used to explore the overall improvement of composites as they move through the process of composite construction. A single model for *Stage* [F (4, 2992) = 21.53, p <.001] demonstrated that *Random Faces* are significantly less accurate than composites after the *First Generation* (p <. 001, *Exp(B)* = 3.10), the *Second Generation* (p <. 001, *Exp(B)* = 6.09), *Holistic Tools* (p <. 001, *Exp(B)* = 10.74) and *Final Image* (p <. 001, *Exp(B)* = 16.15).

To understand the mathematical pattern of the data, polynomial contrasts were simulated using GLMM. The result demonstrated a significant linear (p <.001, *Exp(B)* = 1.56) and quadratic pattern (p <.001, *Exp(B)* = 0.88). However, when cubic (p = .16, *Exp(B)* = 1.03) and quartic (p = .67, *Exp(B)* = 0.99) contrasts were added to the model, the quadratic contrast was no longer significant, indicating that a linear pattern is a better fit for the data.

Additionally, GLMM was used to compare the difference between *Random Faces* and the *Final Image* for composites constructed at each level of *Shape* and *Texture.* As face images were rated within the context of other faces, i.e., composite internal features from each stage of construction were displayed on the screen simultaneously, the 'best' composites are those with the largest difference in rating between the least accurate point (*Random Face)* and the most accurate point (*Final Image).* There was a significant difference between *Random Faces* and the *Final Image* for composites constructed using *One Screen* for *Shape* and *Texture* (p <.001, *Exp(B)* = 16.16), composites constructed using *One Screen* for *Shape* and *Two*

Screens for *Texture* (p <.001, *Exp(B)* = 5.22), *Two Screens* for *Shape* and *One Screen* for *Texture* (p <.001, *Exp(B)* = 8.13) and composites constructed using *Two Screens* for *Shape* and *Texture* (p < .001, *Exp(B)* = 6.14). Although there was a significant difference between the two levels for composites in each condition, the large effect size for composites constructed using *One Screen* for *Shape* and *Texture* indicates that this is the optimal condition for composite construction. However, it is important to understand whether the difference in effect size is significant.

To learn whether there was a significant difference between the effect sizes for composites constructed in each condition, GLMM was used to simulate polynomial contrasts for *Stages*, focusing specifically on linear and quadratic patterns. As in the previous experiment, the data were re-coded to compare the accuracy of *Random Faces* and *Final Image* of composites from two conditions. Polynomial contrasts demonstrated a significant linear and quadratic pattern between composites constructed using *One Screen* for *Shape* and *Texture* and those constructed using *One Screen* for *Shape* and *Two Screens* for *Texture (*linear: p <.001, quadratic: p = .049), *Two Screens* for *Shape* and *One Screen* for *Texture (*linear: p <.001, quadratic: p = .013), and *Two Screens* for *Shape* and *Texture (*linear: p <.001, quadratic: p = .030). Polynomial contrasts also demonstrated a significant pattern between composites constructed using *One Screen* for *Shape* and *Two Screens* for *Texture* and those constructed using *Two Screens* for *Shape* and *One Screen* for *Texture (*linear: p <.001, quadratic: p = .088) and *Two Screens* for *Shape* and *Texture (*linear: p <.001, quadratic: p = .12). Finally, polynomial contrasts demonstrated a significant pattern between composites constructed using *Two Screens* for *Shape* and *One Screen* for *Texture* and those constructed using *Two Screens* for *Shape* and *Texture* (linear: p <.001, quadratic: p = .71).

This pattern of results indicates that the monotonic increase previously found between *Stages* is also present between composites constructed using *One Screen* for *Shape* and *Texture* and composites at all other levels of *Shape* and *Texture, as well as between* composites constructed using *One Screen* for *Shape* and *Two Screens* for *Texture* and those constructed using *Two Screens* for *Shape* and *One Screen* for *Texture.* Such a result for composite images constructed using *One Screen* for both *Shape* and *Texture* indicates that composite accuracy increases from any other condition to *One Screen* for *Shape* and *Texture* (shown by the significant linear pattern), but at a different rate for *Shape* and for *Texture* (shown by the significant quadratic pattern). As is clear from Table 29, composites constructed using *One Screen* for *Shape* and *Texture* are superior compared to composites in any other condition. The presence of a significant three-way interaction between *Shape, Texture* and *Stage* and the pattern of means indicates that composites constructed using *One Screen* for *Shape* are better only when *One Screen* is also used for *Texture.*

Discussion

The final experiment in this thesis aimed to further understand the importance of facial *Shape* and *Texture* during EvoFIT composite construction. In line with the findings from Experiment 4, the first hypothesis was that composites constructed using fewer screens to select the face *Shape* would be more accurate. As facial composites are typically constructed of a target unfamiliar to the participant creating the image, it is theorised that eyewitnesses are more able to recognise and select a face *Shape* which accurately resembles the target compared to recognising and selecting the facial *Texture*. This hypothesis was supported by the results of the experiment; composites constructed using *One Screen* to select the face *Shape* were

more accurate than composites constructed using *Two Screens* to select the face *Shape*. This result indicates that fewer screens are needed to select the facial *Shape* accurately during EvoFIT composite construction. The findings from this experiment support the literature, which states that face *Shape* is more important and *Texture* for the recognition of unfamiliar faces (Johnston & Edmonds, 2009), a process which takes place during the construction of an EvoFIT composite.

Composite naming in Part 2a demonstrated that composites constructed using *One Screen* to select the face *Shape* were named more frequently than composites constructed using *Two Screens*, despite the number of screens used to select the face *Texture*. Furthermore, there was a significant difference in correct naming between composites constructed using *One* and *Two Screens* for *Shape*, but there was little difference found for *Texture*. This result is also supported by composite likeness ratings in Part 2b, which also demonstrate increased accuracy for composites constructed using *One Screen* to select the face *Shape*.

In Part 2b, participants who were unfamiliar with the target images were invited to compare images of the facial composites created in Part 1 and rate how alike they were in comparison to the target face based on the face *Shape*, the face *Texture* or the whole composite. This part of the experiment aimed to understand whether manipulating the number of screens used to select the face *Shape* or *Texture* had a direct effect on the accuracy of the *Shape* or *Texture* of the composite. Mean likeness ratings in this part of the experiment were very similar, resulting in no significant differences between the conditions. In all three measures (rating of the face *Shape*, *Texture* and whole composite), those created using *Two Screens* for selection of the *Shape* and *Texture* were rated the lowest. However, the highest rating condition differed between measures.

For composites rated based on the face *Shape*, those created using *Two Screens* for *Shape* and *One Screen* for *Texture* were the most accurate. For composites rated based on the face *Texture* and the whole face, those created using *One Screen* for *Shape* and *Texture* were rated as the most accurate. One explanation for the lack of significant differences may have been that the difference in the number of screens used in each condition was too small, compared to when *One* screen was compared with *Four Screens* in previous experiments, for example.

The second hypothesis was that composites constructed using fewer screens overall would be more accurate than those constructed using more screens. In Part 2a, composites constructed using *One Screen* to select the face *Shape* and *Texture* were named more frequently than composites constructed using *Two Screens* to select the face *Shape* and *Texture*. Furthermore, the difference in correct naming between composites constructed using *One* and *Two Screens* for selection of the *Shape* and *Texture* was approximately 20%, which demonstrates the large benefit of reducing the number of screens during EvoFIT composite construction from *Two* to *One*.

In Part 2b, composites constructed using *One Screen* to select the face *Shape* and *Texture* were the most accurate based on mean rating, and composites constructed using *Two Screens* to select the face *Shape* and *Texture* were the least accurate, based on mean rating. This pattern of results, and that from Part 2a, demonstrate the clear benefits of reducing the number of screens during EvoFIT composite construction. In this experiment, composites that are constructed using *One Screen* to select the face *Shape* and *Texture* have the lowest number of interacting elements and therefore have the lowest intrinsic cognitive load. Consequently, participants are less likely to experience the negative effects of cognitive overload and are better able to create an accurate facial composite image (Ayres, 2006; Paas et al., 2003).

The third hypothesis was that cognitive load would impact the ability to utilise each stage of composite construction. The pattern of results in Part 2c demonstrates that composites constructed using *One Screen* to select the face *Shape* are more accurate overall, but only when *One Screen* is used to select the face *Texture*, as opposed to *Two Screens*. The results demonstrate that composites constructed using *Two Screens* for *Shape* are more accurate than those created using *One Screen* at the *First* and *Second Generation*, indicating that viewing more face options during face selection is beneficial for the accuracy of the composite at this point in the process. However, after use of *Holistic Tools* and the *Shape* tool to enhance the composite likeness, composites constructed using *One Screen* to select the face *Shape* were more accurate than those constructed using *Two Screens* to select the face *Shape*. This pattern of results indicates that intrinsic cognitive load may have an impact on a participant's ability to utilise the stages of composite construction, as composites constructed using the fewest interacting elements improved more during image enhancement than composites constructed with more interacting elements.

Limitations and Future Research

Crucially, the current experiment demonstrated the impact of reducing the number of screens used to select the face *Shape* and *Texture* during EvoFIT composite construction. In this experiment, composites constructed using the fewest number of screens were most accurate, although it was predicted that composites constructed using *One Screen* to select the face *Shape* and *Two Screens* to select the face *Texture* would be the most accurate. The pattern of results indicated that a smaller population size, that is, the number of faces viewed during composite construction, is beneficial. However, decreasing the population size by reducing the number of screens only

provides a coarse understanding of the optimal population size during composite construction, demonstrating that viewing 18 face images is ideal. It is possible that a smaller number of faces, for example, nine face images or a larger number, such as 24 faces, produces an even more identifiable composite. Therefore, future research should alter the number of face images on *One* or *Two Screens* to more accurately understand the optimal population size during EvoFIT construction.

An additional avenue of research may identify whether viewing more or fewer face images is optimal for eyewitnesses with a strong or poor memory of the face. This might include individuals who viewed the target face after a long delay, had only a partial view of the face or viewed the face from a distance (Holland & Tarlow, 1972).

This experiment was fundamental to understanding the impact of further reducing the number of screens used to select the face *Shape* and *Texture* during EvoFIT composite construction. The results of this experiment demonstrated that reducing the number of screens used to select the face *Shape* is beneficial to the creation of an identifiable composite image. However, reducing the number of screens overall (for selection of the *Shape* and *Texture*) has the greatest benefit. The next chapter will explore the findings of all five experiments in relation to the current literature and will discuss the practical and theoretical contributions to research.

8

GENERAL DISCUSSION

The previous chapter demonstrated the importance of face *Shape* for composite construction but ultimately found that reducing the number of screens overall produced an even more identifiable composite image. The current chapter will explore the findings from all five experiments in this thesis in relation to the literature. This chapter will also consider the theoretical and practical contribution of this thesis while identifying the limitations and future research.

The overarching aim of this thesis was to develop an understanding of the impact of cognitive load on witnesses' abilities to construct an accurate facial composite and to use this understanding knowledge to improve the likeness of facial composite images produced during witness recall. This thesis focused on reducing intrinsic cognitive load, as it was theorised that the intrinsic load would be particularly high given the large number of interactive elements during the construction procedure. Moreover,

intrinsic load is increased further when an individual has no prior knowledge of the task, as is the case for composite construction, making the reduction of intrinsic cognitive load potentially beneficial. Intrinsic cognitive load is directly manipulated in this thesis by reducing the number of face arrays utilised during the composite construction procedure. However, other types of cognitive load are likely to be present during a complicated task such as facial composite construction. As a result, extraneous load (related to the instructional material for the task) and germane load (related to the connection of new and pre-existing information) were also high in Experiments 1 and 2. Therefore, the thesis reduced the intrinsic, extraneous and germane cognitive load during EvoFIT composite construction to increase the likeness of the resulting composites.

The theory of cognitive load (Sweller, 1988) has been used in a plethora of research in the field of education (see, Leppink & van den Heuvel, 2015; van Merriënboer & Sweller, 2010; Paas & Ayres, 2009, 2015) where the theoretical framework is used to design task or learning materials efficiently to reduce the likelihood of cognitive overload and facilitate learning. However, it has not yet been applied to understand the impact on memory in other situations, such as composite construction by eyewitnesses to crime. This thesis is the first to apply this theory to the novel and, as yet, unexplored area of facial composite construction. Composite construction can be considered a somewhat complicated task. It is a novel type of task in which we do not have much experience. Therefore, understanding how to best undertake this task to produce a recognisable composite image is crucial.

The impact of reducing the number of screens during EvoFIT composite construction was explored in Frowd and Grieve (2019); however, this experiment only reduced the number of screens during the construction procedure from *Four* to

Two. Frowd and Grieve (2019) demonstrated that composite likeness increased as the number of screens during the procedure decreased. This thesis aimed to replicate and expand the research in Frowd and Grieve (2019) and to provide a theoretical basis for this phenomenon, thus contributing to the theoretical literature on cognitive load and human memory.

Three main hypotheses were tested in this thesis. The first, tested in Experiments 1-3, was that composites constructed using fewer screens would be more accurate than those constructed using the typical *Four Screens* due to reduced participant cognitive load. The second, tested in Experiments 4 and 5, was that composites constructed using fewer screens to select the face *Shape* would be more accurate than those constructed using more screens, given the important role of face *Shape* in unfamiliar face recognition. The third, tested in all five experiments, was that participants creating a facial composite using fewer screens would be better able to effectively utilise the stages of construction (face selection in Generation 1 and 2 as well as use of *Holistic Tools* and the *Shape Tool*) compared to composites that are created using more screens.

The findings from this thesis will contribute to the facial composite literature, proposing a new construction procedure based on knowledge of human memory and cognitive capability to increase composite likeness. This thesis will also contribute to the image processing literature, demonstrating the importance of face *Shape* over face *Texture* for unfamiliar faces such as facial composites. From an applied perspective, this new procedure is already being trialled by the police, the intention being to increase the number of criminal perpetrators identified through facial composite images as well as reduce police time needed for EvoFIT composite construction.

Reducing the Number of Screens during EvoFIT Construction

This thesis theorised that experiencing high cognitive load during EvoFIT composite construction would overwhelm participant working memory, resulting in reduced memory capacity during complex tasks such as face recall (Camos & Portrat, 2015). If a participant has difficulty with memory recall, they are unlikely to produce a sufficiently accurate composite. Moreover, if a participant has low memory capacity, remembering the target face may be more effortful, making composite construction even more difficult, potentially resulting in the construction of an unrecognisable, and so unusable, composite image. Experiments 1-3 reduced the number of face images viewed to better understand the impact of this on the construction process.

EvoFIT Online Composite Construction

In Experiment 1, the number of screens used during EvoFIT Online composite construction was reduced from *Four Screens* to *One Screen* incrementally. Composites constructed were assessed for accuracy through composite naming and rating. EvoFIT Online allows participants to create a facial composite independently (i.e., self-administered), without assistance or supervision by a researcher or police officer, by following the written on-screen instructions. Although this online system is used by the police in some cases (e.g., if the time delay between the incident and the composite construction would be too long using the EvoFIT App), the composites created are typically less accurate than those constructed using the EvoFIT App (see, Fodarella, 2020; Giannou et al., 2021; Martin et al., 2018). The results from Experiment 1 supported these findings. Specifically, composite naming rates were

very low in this experiment, making it hard to draw clear conclusions. The patterns in the data did support the hypothesis, with composites constructed using *Two Screens* named the most frequently and composites constructed using *Four Screens* named consistently incorrectly.

Overall composite naming rate increased for cued naming.[1], as would be expected. In this measure, composite naming rates increased as the number of screens used during the construction procedure reduced. This finding suggested that viewing many screens during the construction procedure reduces the likelihood of composites being identified, and reducing the cognitive load during the procedure by reducing the number of face arrays viewed is beneficial for composite likeness. This pattern of results supports the hypothesis that composites constructed using fewer screens would be more accurate as the composite naming rate increased, alongside a decrease in the number of screens presented.

Although this low naming rate was unsurprising, given that composites created using EvoFIT Online are identified less frequently than those created using the EvoFIT App (Martin et al., 2018), the severity of the reduced naming rate was unexpected. This could be due to differences in the method. For example, previous experiments conducted using EvoFIT Online included some researcher interaction (Fodarella, 2020; Martin et al., 2018). In Fodarella (2020), participants who used the online, self-administered EvoFIT procedure were interviewed by the researcher, just

[1] Cued naming was used to supplement the results of spontaneous naming, which revealed very low naming rates in Experiment 1. Spontaneous naming rates are representative of typical composite identification that would take place when viewing a facial composite image constructed with the police. However, during cued naming, participants have already viewed the target photographs, which is not representative of composite identification based on a police facial composite. However, viewing the target photographs prior to completing the task does make composite naming easier, resulting in composites that are named more accurately and providing richer data when spontaneous naming rates are too low to develop an understanding of the pattern of the data. Although cued naming results were collected in all five experiments, they were only included in the results where necessary, that is, in Experiment 1.

as participants partaking in the face-to-face procedure would be. Although participants in Martin et al. (2018) were not interviewed by the researcher, the researcher was 'on hand' to answer any procedural questions. In both of these publications, the interaction between the participant and researcher was higher than that in Experiment 1 of this thesis, and both experiments demonstrated higher naming rates compared to Experiment 1. Therefore, the lack of interviewer interaction may have impeded the participant's ability to create an identifiable composite image.

In the current experiment, the researcher did not conduct a CI before composite construction, although they were available over email for any questions during the construction procedure. As a result, there was no opportunity for the researcher to build rapport with the participant (Dion et al., 2021). Building rapport during an eyewitness interview is important for obtaining an accurate description of the perpetrator (Abbe & Brandon, 2013) as it has been shown to facilitate the accuracy of details collected from the witness. Therefore, the lack of rapport between the researcher and participant may be one explanation for the low naming rate of composites created.

Alternatively, low naming rates may have been caused by a lack of attention towards the task or distractions in the environment (Turoman & Vergauwe, 2023). During the construction procedure using EvoFIT Online, there was no check to ensure that participants were focused on the task. Self-administered studies are more likely to result in a lack of attention than studies with the researcher present (Berinsky et al., 2014); hence, the low composite naming may have resulted from a lack of attention to the task. Furthermore, as the researcher was unable to view the environment that participants were in while creating the facial composite, there is no way to ensure that participants were sat in a quiet room with few distractions, as eyewitnesses would be

when creating a facial composite in a police station. Therefore, the low composite naming may have been caused by a number of distractions in the participant's environment, such as the television or other people in the house who need attention. These limitations were addressed in Experiments 2 and 3 by increasing the interaction between the researcher and the participant.

Despite the low correct naming rates of composites in this experiment, facial composite rating demonstrated that there was a significant difference between composites constructed using *One* and *Four Screens*.

Composite rating in this experiment indicated that composites constructed using *One Screen* to select faces were rated as the most accurate based on the *Internal Features* and the whole face. During the face selection process, where the number of screens viewed is manipulated, only the composite *Internal Features* are visible. Therefore, this pattern of results may be due to the reduction in cognitive load when selecting the *Internal Features* at the beginning of the construction procedure.

In support of the third hypothesis, composite ratings in this experiment demonstrated that the ability of participants to utilise the image enhancement tools during EvoFIT composite construction (*Holistic Tools* and the *Shape Tool*) varied dependent on the number of screens used to select the face images at the beginning of the construction procedure. Composites constructed using *One* or *Three Screens* did not demonstrate a significant difference between the composite likeness ratings at the various stages of composite construction. However, composites constructed using *Two* or *Four Screens* did demonstrate a significant difference between the composite likeness ratings throughout the stages of construction. Composites constructed using *Two Screens* reduced in accuracy from the *random face* to the *First generation* before increasing in accuracy after the *Second generation, after holistic tools* and the *final*

image. Whereas composites constructed using *Four Screens* increased in accuracy at each stage of the construction process, from the *random face* to the *final image*, in contrast to the hypothesis.

This pattern of results is not supported by Cognitive Load Theory, which states that the cognitive load should increase with the number of screens used (Sweller, 1988). Based on this theory, it was predicted that composites constructed using *Four Screens* would reduce in accuracy towards the end of the construction procedure, whereas composites constructed using *One Screen* should increase in accuracy throughout the construction procedure (Paas et al., 2003). Therefore, this result indicates that cognitive load did not impact the ability of participants to utilise the stages of composite construction.

In this experiment, composite likeness ratings indicated that random composite images were a more accurate representation of the target than the composites created in Part 1 of the experiment using *One* and *Two Screens.* Given that Frowd and Grieve (2019) demonstrated that identifiable composites can be created using *Two Screens* via the EvoFIT App, it is unlikely that reducing the number of screens typically reduces the likeness so much that a random composite image is more accurate. Furthermore, the low composite naming in Part 2a of this experiment may be an indication that the composites constructed in this experiment did not reflect those in prior research (Erikson et al., 2022). Consequently, Experiment 2 repeated this measure in an experiment constructing facial composites with increased researcher-participant interaction and using the EvoFIT App to better reflect composite construction with the police.

Overall, this Experiment loosely supported the hypothesis that composites constructed using fewer screens would be more accurate than those constructed using

the typical *Four Screens*. This hypothesis was supported in Part 2b of the experiment, where composites constructed using *One Screen* were rated as the most accurate based on the *Internal Features* and the whole face images. However, this experiment did not support the hypothesis that participants creating a facial composite using fewer screens were better able to utilise the stages of construction compared to composite construction with more screens.

Face-to-Face EvoFIT Composite Construction after a Cognitive Interview

In Experiment 2, the accuracy of composites constructed using *One*, *Two*, *Three* or *Four Screens* during traditional, face-to-face construction was assessed using composite naming and rating measures. One hypothesis predicted that composites created using fewer screens would result in more accurate composites; however, this pattern of results only occurred for composites constructed using *One*, *Two* or *Three Screens*, with composites created using *Three Screens* named the least frequently, followed by *Two* and then *One Screen*. Yet, composites constructed using *Four Screens* were named the most frequently overall, indicating that any negative impact of cognitive load may have been outweighed by the benefit of viewing a wider variety of options during face selection, ultimately creating a more accurate composite image.

In Part 2b, composites constructed using *One Screen* were rated as the most accurate based on the *Internal Features*, the *External Features* and the whole composite images. This pattern of results demonstrates the benefits of reducing the number of screens during EvoFIT composite construction, as composites constructed using the lowest number of screens (*One Screen*) were rated as the most accurate across all three measures. As previously mentioned, during face selection, where the number of screens viewed is manipulated, only the *Internal Features* of the face are

visible. Removing the *External Features* from a face image during this stage of construction enhances the likeness of the resulting composite because participants are not distracted by the *External Features* (Frowd et al., 2012). This is particularly effective as familiar face recognition, which takes place during composite construction, relies on *Internal Features* as opposed to *External Features* (Latif & Moulson, 2021). Based on composite rating of the *Internal Features*, it appears that composites constructed using *One Screen* are the most accurate, and composites constructed using *Four Screens* are the least accurate.

In addition, the results of this experiment demonstrate a significant interaction between the number of screens used during composite construction and the stage of construction. The results demonstrate that composites constructed using *One Screen* increased in accuracy between the final two stages of composite construction twice as much as composites constructed using *Four Screens*. This finding strongly supports Cognitive Load Theory (Sweller, 1988), demonstrating that participants who completed the task with a smaller intrinsic load were able to execute the task more effectively than participants with a heavier intrinsic load. As cognitive overload (resultant from high cognitive load) can impair memory (Byyny, 2016), it is possible that participants creating a composite using more screens, and therefore experiencing higher cognitive load, were unable to effectively utilise the image enhancement tools towards the end of the construction procedure.

In Part 2a of this experiment, a second condition, familiarisation, aimed to refamiliarise participants with the targets to increase correct composite naming rates. The results demonstrated that there was little difference in correct composite naming between participants who partook in the familiarisation task, and participants who did not. Perhaps one reason why this familiarisation task did not work was because all

participants were sufficiently familiar with the target pool (in this case, England Footballers), and thinking about the potential targets for one minute was not necessary for participants to become familiar with the identities. Alternatively, participants may have refamiliarised themselves with some England Footballers, but not those who were included as targets in the experiment, so the task did not improve the recognition of the targets.

However, inviting participants to recall footballers to trigger the memory of other footballers may not work effectively. Typically, recall is used to trigger more detail about episodic memories (Hassabis & Maguire, 2007), which are summary records of experience, enabling individuals to remember past experiences (Conway, 2009; Tulving, 2002). However, knowledge of past footballers is considered semantic memory, which includes the recollection of ideas, concepts and facts generally (i.e., general knowledge: Tulving, 1972). Importantly, different methods are used to aid the recollection of episodic memories, such as using association (Hills et al., 2012). With this knowledge, an alternative technique could have been used to better refamiliarise participants with the targets. For example, participants may have remembered more footballers if they had been given a cue, such as a list of years (perhaps the last 10 years as footballers selected for targets have all played for England in the last decade) and were invited to recall the players in each team for each year. Having the cue of the year may have increased the number of footballers retrieved from memory, increasing the likelihood of participants remembering the targets.

Despite the limited improvement in composite naming due to the familiarisation task, composites in this experiment were named somewhat more frequently than composites in Experiment 1. This increase in overall correct naming may be due to the increased interaction during the construction procedure, which

enabled the researcher and participant to build rapport (Portch et al., 2017). The positive effect of rapport during forensic interviewing is well documented (Collins et al., 2002; Nash et al., 2016; Wolfs et al., 2022), so the increase in accurate, detailed information recalled by the participant prior to the construction procedure may have enabled participants to create a more accurate composite image.

Overall, the findings of Experiment 2 somewhat supported the hypothesis that composites constructed using fewer screens are more accurate than composites constructed using the typical *Four Screens*. Although composite naming demonstrated that the ability to view more face options outweighed the benefits of reducing the cognitive load, the pattern of results from composite rating in Parts 2b and 2c strongly supported the hypothesis, demonstrating that composite likeness increases as the number of screens reduces during the construction procedure. The findings in Part 2c also supported the second hypothesis that participants creating a facial composite using fewer screams will be better able to utilise the stages of construction compared to composites that are created using more screens.

Face-to-Face EvoFIT Composite Construction after a Holistic-Cognitive Interview

Although composite likeness increased between Experiments 1 and 2, the composite naming rates in Experiment 2 were still lower than those in recently published EvoFIT research (see, Erikson et al., 2022; Fodarella et al., 2021; Frowd, 2021). Experiment 3 used a Holistic-Cognitive Interview (H-CI), which is specifically designed to obtain a description of the criminal perpetrator before the construction of a facial composite image. The literature demonstrates a substantial increase in composite identification when an H-CI is used prior to EvoFIT construction (Frowd,

Nelson et al., 2012; Frowd, Pitchford et al., 2012; Frowd, Portch et al., 2019; Portch et al., 2017).

In Experiment 3, composites constructed using *One Screen* were correctly named most frequently, followed by composites constructed using *Two*, *Three* and *Four Screens*. This pattern of results clearly demonstrated the benefit of reducing the number of screens during EvoFIT composite construction, revealing a linear trend whereby the accuracy of a composite increased as the number of screens viewed during the construction process decreased. This finding is strongly supported by Cognitive Load Theory (Sweller, 1988), which states that, as the cognitive load of a task increases, performance decreases (Sweller, 2010).

This pattern of results, whereby composites constructed using *One Screen* were the most accurate, was supported by likeness ratings in Part 2b of the experiment. In Part 2b, composites constructed using *One Screen* were rated as the most accurate based on the composite *Internal Features*, *External Features* and the whole face. However, unlike composite naming, this pattern of results was not linear, demonstrating that, although composites constructed using *One Sceeen* were the most accurate overall, this was followed by those constructed using *Four*, *Two* and then *Three Screens*. This pattern of results indicates that, although there is a benefit of reducing the number of screens used during EvoFIT composite construction, there is still some benefit of viewing many screens during the construction process. This finding supports findings from Experiment 2, whereby composites constructed using *Four Screens* were named the most frequently followed by those created using *One*, *Two* and *Three Screens*.

Despite the hypothesis that reducing the number of screens during composite construction would allow participants to utilise the image enhancement tools more

effectively, the number of screens used to select face images during the EvoFIT construction procedure did not impact the likeness of composites produced. However, the results did demonstrate that composite likeness is dependent on the stage of construction. Composites created using *One*, *Three* and *Four Screens* increased in accuracy throughout the construction process, whereas composites constructed using *Two Screens* reduced in accuracy very slightly in the *Final Image*. The increase in accuracy between each stage of composite construction when composite images were created using *Three Screens* reduced at each subsequent level. This pattern of results is expected because participants' ability to utilise each stage of the construction procedure is reduced as the cognitive load of the task increases. This same pattern of composite likeness was not found for composites constructed using *Four Screens*, where the cognitive load is expected to be highest. Hence, further research is still needed to understand the impact of cognitive load on composite accuracy through the stages of construction.

Comparison of Experiments 1-3

In Experiments 1 and 3, composite naming was as predicted, with composites constructed using the fewest screens being named the most frequently. However, this differed in Experiment 2, with composites constructed using *Four Screens* being named more frequently than those created using *One*. As the results from Experiment 1 could not be analysed using GLMM, it is sensible to compare the findings of Experiments 2 and 3. The third experiment in this thesis was a replication of the second, with the only change being the interview technique, whereby the CI was replaced with the H-CI. In Experiment 2, composites constructed using *One*, *Two* and *Three Screens* increased in accuracy as the cognitive load was reduced; therefore,

cognitive load likely did have an impact on composite construction. However, as composites constructed using *Four Screens* were the most accurate overall, one may speculate that the cognitive load was not reduced enough, allowing the ability to view more face options to outweigh the benefit of reduced cognitive load. However, the H-CI introduced in Experiment 3 refreshed the participants' memories of the target face more effectively, improving their memories of the targets and therefore making the comparison of new information (the face images on the screen) to existing information (the target face) easier, reducing the germane cognitive load (Kalyuga, 2009). It is theorised that reducing the germane cognitive load by introducing the H-CI in all conditions means that the benefits of viewing more face options during the face selection stage of composite construction no longer outweigh the benefit of reducing the cognitive load.

Composite likeness ratings in Part 2a measured the likeness of composite *Internal Features*, *External Features* and *Whole Faces*. However, it is the *Internal Features* that are most interesting here, as only the *Internal Features* are visible during the face selection stage, which is when the number of screens viewed is manipulated. In all three experiments, composites constructed using *One Screen* are rated as having the most accurate *Internal Features*, which suggests that reducing the number of screens during the face selection stage enables participants to select more accurate face images. As individuals identifying composite images created with the police view an image of the whole face, likeness ratings of composite whole face images are also important. The pattern of results for likeness rating of composite *Whole Faces* follows that of *Internal Features*, with composites constructed using *One* screen rated the highest. This pattern of results indicated that reducing the number of screens during composite construction and therefore reducing the cognitive

load (Sweller, 1988) has a positive impact on the likeness of the resulting composite image.

Intermediate likeness ratings in Part 2c of each experiment provided the most inconsistent results between the three experiments. It was predicted that composite likeness would increase in accuracy throughout the construction procedure but that changes in accuracy between each subsequent stage would reduce during composite construction using more screens. In Experiment 1, composites constructed using more screens were rated higher than those constructed using fewer screens and random composite faces were rated higher than expected in comparison to the *Final Image*. In Experiment 2, composites constructed using fewer screens increased in accuracy more towards the end of the construction procedure compared to those constructed using the typical *Four Screens*. While in Experiment 3, there was little difference in accuracy between composites constructed using *One*, *Two*, *Three* or *Four Screens*. The inconsistent pattern of results for intermediate composite likeness ratings makes the findings from this measure difficult to interpret. Therefore, this measure was carried through to the final *Two* experiments to provide a better understanding of the likeness of facial composites as they move through the stages of construction.

Overall, the findings from Experiments 1-3 demonstrated that reducing the number of screens during the composite construction procedure is beneficial for composite naming and likeness. This result was particularly clear in Experiment 3, which most closely resembled composite construction with the police due to the introduction of the H-CI (Frowd et al., 2019), indicating that composite construction using *One* screen to select the face *Shape* and *Texture* at the beginning of the procedure is optimal.

Across all three experiments, the number of screens used to select face *Shape* and *Texture* was reduced at the same rate. Yet, the literature indicates that face recognition based on a perfect representation of the face, for example, a photograph, differs depending on familiarity with the face (Barton et al., 2006). Put simply, familiar face recognition is more reliant on face *Texture*, and unfamiliar face recognition is more reliant on face *Shape* (Johnston & Edmonds, 2009). During the creation of a facial composite, it is difficult to know whether an eyewitness would be considered to be familiar or unfamiliar with the target, as they have only viewed the face for a short period of time. The individual identifying the facial composite is familiar with the target, yet a facial composite is an imperfect representation of the target, making recognition different than it would be for a perfect face photograph. Consequently, Experiments 4 and 5 reduced the number of screens individually for the selection of the face *Shape* and *Texture* to understand the importance of *Shape* and *Texture* during composite construction and potentially further enhance the composite construction procedure.

Exploring the Importance of Face Shape and Texture during EvoFIT Construction

The importance of face *Shape* and *Texture* for familiar and unfamiliar face recognition has been debated in the literature (Kaufmann et al., 2013; Rogers et al., 2022; Russell & Sinha, 2007), yet the importance of these factors in eyewitness facial composite construction has not yet been tested. These novel experiments reduced the number of face images viewed during the selection of face *Shape* and face *Texture* screens independently in order to understand the importance of face *Shape* and

Texture for facial composite construction and potentially further optimise the construction procedure.

Reducing the Number of Screens during EvoFIT Construction for Face Shape and Texture Individually

In Experiment 4, the number of screens used during EvoFIT composite construction was reduced for the selection of the face *Shape* and the face *Texture* individually, so participants viewed either *Two* or *Four Screens* for the selection of face *Shape* and *Two* or *Four Screens* for the selection of face *Texture*. In this experiment, composites constructed using *Two Screens* to select the face *Shape* were significantly more accurate than those constructed using *Four Screens* to select the face *Shape*, regardless of the number of screens used to select the face *Texture*, supporting the hypothesis. There was a significant difference between composites constructed using *Two* and *Four Screens* to select the face *Shape* but a non-significant difference between composites constructed using *Two* and *Four Screens* to select the face *Texture*. This pattern of results highlights the importance of face *Shape* over face *Texture* for unfamiliar face recognition (see Bruce et al., 1991; Kaufmann et al., 2013; Limbach et al., 2022) as manipulating the number of faces for selection of the face *Shape* had a significant impact on the accuracy of the composite, whereas manipulating the number of screens for *Texture* did not.

The results from composite naming in this experiment are supported by composite likeness ratings. In Part 2b, participants were invited to rate the likeness of composite face shapes, *Texture*s or whole composite images compared to the target. The result indicated that there was a significant three-way interaction between the number of screens used to select face *Shape*, the number of screens used to select face

Texture, and the task (whether participants were asked to rate the face *Shape*, *Texture* or whole face). When rated based on the *Shape*, composites constructed using fewer screens to select the face *Shape* were the most accurate. However, when rated based on the face *Texture*, composites constructed using *Four Screens* to select the face *Texture* were the most accurate. This pattern of results indicates that fewer screens are needed to accurately select the face *Shape*, but more screens are needed to select the face *Texture*.

In Part 2c of this experiment, all composites increased in likeness between each stage of the construction procedure. Moreover, mean likeness ratings in this part of the experiment indicated that composites created using *Two Screens* for *Shape* and *Texture* were the most accurate, followed by those created using *Two Screens* for *Shape* and *Four Screens* for *Texture, Four Screens* for *Texture and Two Screens* for *Shape* and finally *Four Screens* for *Shape* and *Texture.* This pattern of results supports previous findings that reducing the number of screens for the selection of the face *Shape* is particularly beneficial for accurate composite construction.

Further analysis in this part of the experiment focused on the difference in composite likeness rating between *Random Faces* and *Final Images* as composites at each stage of construction were displayed together. Therefore, the difference in likeness ratings between the lowest-rated composite image (*Random Face*) and the highest-rated composite image (*Final Image*) provides information about how much composites increase in likeness throughout construction. Hence, the larger the difference in rating between a *Random Face* and the *Final Image*, the more optimal the construction process was.

The analysis demonstrated that the largest difference in composite likeness between a *Random Face* and the *Final Image* was for composites constructed using

Four Screens to select the face *Shape* and *Two Screens* to select the face *Texture*. However, composites in this condition were rated as the least accurate in the *Final Image*. The analysis assumed that the difference in composite likeness between a *Random Face* and the *Final Image* would inform the most optimal construction procedure. Yet, composites constructed using *Four Screens* for the selection of the face *Shape* and *Two Screens* for the selection of the face *Texture* were rated as the least accurate in the *Final Image*. This calls into question whether this procedure is optimal or whether this measure is a reliable indication of composite likeness. In future, it may be more appropriate to display each face image on a different screen, as opposed to displaying five images side-by-side.

When face images are displayed simultaneously, participants are likely to make subconscious comparisons between the images, which may affect their ratings or may subconsciously make decisions based on the order of the face images, despite being informed that the order is random (Nyman et al., 2020). If the face images were displayed sequentially, the likelihood of comparison between images would be lower (Kaesler et al., 2020), and so ratings would be more reliable.

Further Reducing the Number of Screens during EvoFIT Construction for Face Shape and Texture Individually

The fifth and final experiment replicated Experiment 4, with one crucial difference: the number of screens used for the selection of face *Shape* and *Texture* was reduced from *Two* and *Four Screens* to *One* and *Two Screens*. Participants viewed either *One* screen for the selection of the face *Shape* and *Texture*, *Two Screens* for face *Shape* and *Texture*, *One* screen for face *Shape* and *Two Screens* for *Texture*, or *Two Screens* for face *Shape* and *One* screen for *Texture*.

The hypothesis that face *Shape* is more important than face *Texture* for facial composite construction was supported in Part 2a of this experiment. Composite naming demonstrated that composites constructed using *One Screen* to select the face *Shape* were more accurate than those constructed using *Two Screens* to select the face *Shape*, despite the number of screens used to select the face *Texture*. Composites constructed using *One* screen to select the face *Texture* were also more accurate than those constructed using *Two Screens*. However, the difference in composite naming between the two levels of *Texture* was not significant, whereas the difference between the two levels of *Shape* was significant. This pattern of results indicates that the impact of manipulating the number of screens for face *Shape* is larger than the impact of manipulating the number of screens for *Texture*. As EvoFIT construction relied on unfamiliar facial recognition, this result supports the literature which states that face *Shape* is more important for unfamiliar face recognition than *Texture* (Bruce et al., 1991; Lee & Perret, 1997; Rogers et al., 2022; Russell & Sinha, 2007).

Composite naming in Part 2a is somewhat supported by composite likeness ratings. Overall mean likeness ratings demonstrated that composites constructed using *One* screen to select the face *Shape* were more accurate than composites constructed using *Two Screens* to select the face *Shape*. However, this difference was small and did not indicate a noteworthy improvement in composite likeness. The similarity in likeness rating scores between the levels of *Shape* and *Texture* is highlighted by the lack of significant differences in this part of the experiment. This pattern of results indicates that the difference in composite likeness between composites constructed using *One* or *Two Screens* to select the face *Shape* and *Texture* was too small to attain a significant result when assessed using likeness ratings of composite face *Shape*, *Texture* and *Whole Faces*.

Following the pattern of results from Experiments 2-4, Part 2c of Experiment 5 demonstrated that composites generally increase in accuracy throughout the stages of composite construction. More specifically, composites constructed using *One Screen* for *Shape* and *Texture*, and those constructed using *Two Screens* for *Shape* and *Texture*, increased in accuracy throughout the construction procedure. However, composites constructed using *One Screen* for *Shape* and *Two Screens* for *Texture*, or those constructed using *Two Screens* for *Shape* and *One Screen* for *Texture*, reduced in accuracy at the end of the construction procedure, between the stages *After Holistic Tools* and *Final Image.* This pattern of results may indicate that composite construction using a different number of screens during the selection of the *Shape* and face *Texture* reduced the composite accuracy at the end of the construction procedure. However, composites created using *Two* or *Four Screens* in Experiment 4 continued to increase in accuracy throughout the construction procedure. Therefore, this result may indicate that participants constructing a composite using a different number of screens to select face *Shape* and *Texture* when the number of screens used is so low may be unable to enhance the composite likeness towards the end of the construction procedure.

Comparison of Experiments 4 and 5

Experiments 4 and 5 both demonstrate that reducing the number of screens used to select the face *Shape* during EvoFIT composite construction is beneficial for producing a good likeness. Composite naming is arguably the most reliable measure of composite likeness as it closely reflects the process of facial recognition that takes place when an individual attempts to identify an individual from a facial composite image produced by the police. In Experiment 4, composite naming demonstrated that

composites constructed using *Two Screens* to select the face *Shape* and *Four Screens* to select the face *Texture* were more accurate than those constructed using *Four Screens* for *Shape* and *Two Screens* for *Texture* as well as those constructed using *Four Screens* for *Shape* and *Texture*. This result was supported by composite naming in Experiment 5, whereby composites constructed using *One* screen to select the face *Shape* and *Two Screens* to select *Texture* were more accurate than those constructed using *Two Screens* for *Shape* and *One* for *Texture* as well as those constructed using *Two Screens* for *Shape* and *Texture*.

However, in both experiments, composites constructed using the lowest number of screens were the most accurate. Experiments 1-3 clearly demonstrated that reducing the number of screens during EvoFIT construction was beneficial for the construction of an accurate composite image. This finding was supported by the results in Experiments 4 and 5. In Experiment 4, composites constructed using *Two Screens* for the selection of the face *Shape* and *Texture* were the most accurate, and in Experiment 5, composites constructed using *One Screen* for the selection of face *Shape* and *Texture* were the most accurate. Moreover, in Experiments 4 and 5, composites constructed using the highest number of screens for the selection of face *Shape* and *Texture* were the least accurate: *Four Screens* in Experiment 4 and *Two Screens* in Experiment 5.

Overall, the two final experiments in this thesis support the hypothesis that composites constructed using fewer screens to select the face *Shape* are more accurate than composites constructed using more screens to select the face *Shape*. The pattern of results in these experiments indicates that face *Shape* is more important for composite construction than face *Texture*, as manipulating the face *Shape* had more of an impact on the composite image constructed. Furthermore, these experiments

support the findings from Experiments 1-3, demonstrating that composites constructed using fewer screens are more accurate than those constructed using more screens.

Theoretical Contribution

Cognitive Load

This thesis explored the benefit of reducing cognitive load during composite construction, demonstrating that reducing the number of screens viewed during the face selection stage of the construction procedure will result in a more recognisable composite image. This was expected in line with the theory that reducing the number of interactive elements in a task, therefore reducing the intrinsic cognitive load (Sweller, 1988), lessens the likelihood of participants experiencing cognitive overload. If participants experience cognitive overload during the construction of a facial composite, their memory capacity and their decision-making ability may become impaired, restricting their ability to create an identifiable composite image (Szulewski et al., 2020). This thesis is the first to investigate the impact of cognitive load on facial composite construction. Cognitive load is typically used in the field of education to design tasks or learning materials efficiently to reduce the likelihood of cognitive overload (see, Leppink & van den Heuvel, 2015; van Merriënboer & Sweller, 2010; Paas & Ayres, 2014). Yet, this thesis demonstrated that reducing the number of screens during the construction procedure, therefore reducing the cognitive load, is beneficial to composite accuracy.

The impact of cognitive load on facial composite construction has not previously been explored in the literature or in practice. As such, there was no

theoretical understanding of the impact of viewing many face images on the construction of an identifiable facial composite image may have. This thesis presented clear evidence to demonstrate the positive impact that reducing the number of faces compared during the construction procedure can have on the final facial composite image. Yet, the literature has demonstrated the negative impact of viewing many faces during a police line-up (Hinz & Pezdek, 2001; Lin et al., 2019). Based on the findings in this thesis, an explanation for the negative impact of viewing many faces during a police line-up may be cognitive overload. It may therefore be important to reduce the cognitive load during the line-up procedure to optimise this process, perhaps increasing the number of perpetrators identified and reducing the number of innocent individuals identified in such line-ups.

This thesis expands our understanding of cognitive load and supports the Cognitive Load Theory, acknowledging its impact on complex tasks requiring our attention, such as facial composite construction. Alternative theories of attention in Chapter 1 (the Theory of Visual Working Memory and the Perceptual Load Theory: Jackson & Raymond, 2010; Lavie, 2010) and are unable to explain the increase in composite likeness as the number of face arrays viewed reduces.

The Theory of Visual Working Memory states that individuals can only process two faces at one time (Jackson & Raymond, 2010), yet participants creating facial composites in each experiment viewed 18 faces on each screen. When applying this theory, EvoFIT composite construction should be too difficult to create an accurate composite image in any condition, yet, in all experiments except the first, some composites were named correctly in each condition. Therefore, it is unlikely that the Theory of Visual Working Memory can explain the increase in composite likeness as the number of face arrays during the construction procedure is reduced.

According to the Perceptual Load Theory, participants creating facial composites under high perceptual load (those viewing more face arrays) are more likely to process irrelevant distractors and are more open to suggestion (Murphey & Greene, 2016). During the construction procedure, the researcher was unaware of the target face, so suggestions could not be made consciously or subconsciously to influence the participants' choices. Therefore, it is unlikely that participants viewing more face images (therefore under high perceptual load) were impacted by an increased openness to suggestion.

Moreover, the EvoFIT App is designed to minimise the number of distractors on screen, so when applying the Perceptual Load Theory in this thesis, participants would be more likely to process distractors in their environment than on screen. As participation in all five experiments took place online, with participants creating the facial composite images in their own homes, it is possible that participants with a high perceptual load were more likely to process irrelevant distractors in their environment. Yet, even when under high levels of perceptual load, participants are typically able to remember key details and are more likely to forget peripheral, seemingly unimportant information (Lavie, 2010). Therefore, even with a high perceptual load, participants may not be so distracted that they are unable to create an identifiable composite image. To further rule out the Perceptual Load Theory as an explanation for the pattern of results in this thesis, a future experiment should replicate Experiment 3 (composites created using *One, Two, Three* or *Four Screens* after an H-CI) in a laboratory to reduce the environmental distractors available to participants.

In comparison to the theories discussed above, the Cognitive Load Theory (Sweller, 1988, 2010) sensibly explains the finding whereby composite identification increases as the number of screens is reduced. This theory states that cognitive load

increases with the number of elements that must be processed in the working memory simultaneously during the task. Therefore, when participants must view more face images during the construction procedure, the number of elements is increased, and so is the cognitive load. However, when the number of face images is reduced, the number of elements is also reduced, and the cognitive load of the task is lower, allowing participants to complete the task more successfully.

Face Shape and Texture

There is a lack of literature exploring the importance of face *Shape* and face *Texture* for the recognition of a facial composite image. Although face *Shape* is deemed to be the most important for unfamiliar face recognition based on a photograph (Lee & Perrett, 1997) or head model (Bruce et al., 1991), there was no evidence to indicate whether this is also true for unfamiliar face recognition of a facial composite image during the construction process. Head models and face photographs typically display a perfect representation of a target's face; however, facial composite images created during research or with the police are imperfect representations of the target. Therefore, although there is some understanding of the importance of face *Shape* and *Texture* for the recognition of accurate faces, this research cannot be directly applied to the recognition of imperfect faces during the construction procedure or while identifying a target from a facial composite image.

To bridge this gap in the literature, this thesis explored the importance of face *Shape* and face *Texture* in the creation of an EvoFIT facial composite. As eyewitnesses create facial composites of faces that are unfamiliar to them, it was predicted that the face *Shape* would be more important than face *Texture* during composite construction (Benson & Perrett, 1991; Butcher et al., 2011; Knight &

Johnston, 1997). In support of this hypothesis, the current thesis demonstrates that reducing the number of screens used to select the face *Shape* during composite construction largely improves the composite likeness. This finding supports literature demonstrating the importance of face *Shape* for unfamiliar face recognition and indicates that literature utilising recognition of perfect face representations can, at least in this instance, be applied to the imperfect face images created during facial composite construction.

Overall, this thesis identifies a disparity in the importance of facial *Shape* and *Texture* during the creation of a facial composite, determining that manipulating the face *Shape* has a larger impact than manipulating the face *Texture*. This thesis, therefore, highlights the importance of face *Shape* over face *Texture* for unfamiliar face recognition.

Practical Contribution

This thesis significantly improved the likeness of facial composite images produced using the EvoFIT facial composite system by reducing the number of screens viewed during the face selection stage in the construction procedure. Using this newly developed procedure, eyewitnesses were less likely to experience cognitive overload and were better able to construct a more recognisable composite image. By implementing this optimised procedure, police forces may aid eyewitnesses in producing more recognisable composite images, increasing the likelihood of perpetrators of crime being identified. Currently, based on the findings from this research, this new procedure is being field trailed by all police forces currently using the EvoFIT system. The new procedure was used in the case of a robbery at

knifepoint in February 2023. In this case, two offenders (White, Caucasian males aged 23-28) attempted to steal a bike at knifepoint. A facial composite was created of each assailant by the victim five days after the incident, and they are currently being used to help police with the enquiry. After a 12-month field trial, comments and recommendations for any adaptions to the procedure will be collected and addressed.

A further practical contribution of the new construction procedure is that it may save time, as eyewitnesses only view *One Screen* to select the face *Shape* and *Texture* as opposed to viewing *Four Screens*. As time is a very valuable police resource, adapting the procedure to save time is beneficial and may encourage the police to utilise facial composites in cases where they may have previously been unable to. In the February 2023 robbery mentioned above, the victim was able to create a facial composite of the two assailants in the time it would have previously taken to produce one image, which demonstrates the time-saving ability of the new procedure.

The new construction procedure is not only beneficial for the police practitioners leading the composite construction but, as it reduces the likelihood of cognitive overload, it is also an easier, less demanding process for eyewitnesses. Creating a facial composite as an eyewitness to crime is likely to be a stressful experience, particularly as many crimes that require a police facial composite are in close range, often involving violence (see, Frowd et al., 2015). By reducing the number of face arrays viewed during the construction procedure, the process of composite construction should be faster, meaning that eyewitnesses can spend less time creating the facial composite image.

Although the increased speed with which a facial composite can be created using the new procedure is a great benefit, the greatest advantage is the ability to

create a more identifiable facial composite image. With the increased likeness, facial composites created using *One Screen* should be more likely to result in the criminal perpetrator being identified, leading to more convictions and more reliable convictions.

Additionally, the finding that face *Shape* was more important than *Texture* is particularly important for EvoFIT facial composite construction as the face *Shape* and *Texture* are selected individually, so future developments in this system may decide to concentrate on improving the face *Shape*. However, many other facial composite systems, such as EFIT6 and FACES, display the face *Shape* and *Texture* together as a global face model (George et al., 2008). The new understanding that face *Shape* is more important than face *Texture* for the creation of an accurate facial composite may prompt other facial composite system developers to focus on improving the face *Shape* over the face *Texture*, perhaps even separating the selection of the face *Shape* and *Texture* during the face selection process.

Limitations and Future Research

In this thesis, all experimentation was conducted online using video conferencing platforms such as Microsoft Teams and Skype. Aside from Experiment 1, where composites were created using the self-assessment system EvoFIT Online, facial composites were constructed over a conference call. Participants creating the facial composites could see the researcher's screen using the 'screen share' feature available on both platforms, and the researcher had control of the mouse as they would during typical, face-to-face composite construction. Every step was taken to ensure that the

construction procedure replicated that used by the police; however, it is important to recognise that there may be some limitations to online composite construction.

Online Data Collection

One limitation that may arise during any online experimentation, for example, is connection problems. There may be a delay between the two individuals on the conference call, or there may be instances in which the screen freezes for one or both individuals. Facial composite construction in the current thesis was not a timed task, so any connection problems that occurred did not directly impact the results.

Such problems may have caused issues when the researcher was trying to explain the procedure to the participant, particularly because it was important to communicate clearly and to ensure that the participant understood the procedure. Connection difficulties during the construction process were otherwise unproblematic as the process did not rely on the construction proceeding at a set pace, and the researcher or the participant was able to repeat themselves if they were not heard.

Importantly, having to repeat yourself during a conversation online due to connection difficulties may impair rapport (Weller, 2017). Furthermore, it is well known that building rapport during an eyewitness interview is important for obtaining an accurate description of the perpetrator (Abbe & Brandon, 2013) as it has been shown to facilitate the accuracy of details collected from the witness. Rapport building is more difficult via video conferencing than it is face-to-face (Fullwood, 2008) as it is more difficult to see an individual's body language via video conferencing, and therefore more difficult to mirror body language, which is considered an effective method of increasing rapport (Nancarrow & Penn, 1998). As composites were constructed online via video conferencing, it may be more difficult

for the researcher to build a rapport with participants. Previous research demonstrates that a lack of rapport during a police interview may reduce accurate information (Collins et al., 2002) and increase inaccurate or misinformation (Vallano & Compo, 2011). Therefore, to build rapport with the participants to a level that would be typical in a police investigation, the researcher spent time asking open-ended, friendly questions, as in Nash et al. (2014).

A further limitation of the online data collection in this thesis is that participants did not complete the experiment in a controlled environment, such as a laboratory. This limitation may be most prevalent for participants creating facial composites using EvoFIT Online in Experiment 1, as there was no researcher supervision during the procedure. Consequently, there was no assurance that participants remained focused on the task throughout the construction procedure without distractions, as an eyewitness would be during composite construction with the police. Therefore, it is important for the police to trial the new procedure developed through this thesis before this is introduced to all police forces, and these trials are currently underway.

Population Size

This thesis clearly demonstrated the benefits that reducing the number of screens during EvoFIT composite construction had on EvoFIT composite construction. It was theorised that as the number of screens reduced, the number of interacting elements (i.e., face images) decreased, which lessened the intrinsic cognitive load of the task. When reducing the number of face images available for eyewitnesses to view, the variation in faces is also reduced. It is unlikely that 18 faces (the number of faces displayed on one screen) represent the variability in the population. Therefore, it is

possible that some eyewitnesses are unable to select a face that closely resembles the perpetrator when creating a composite using one screen. Yet, composites created in this condition were typically named more frequently. This finding indicates that identifiable composite images can be created despite a lack of variability in the faces displayed. One explanation for this finding may be that image enhancement tools (holistic tools and the shape tool) and more important for the creation of an identifiable composite image than the face that is selected at the beginning of the procedure.

The reduction in population size throughout this PhD project is rather coarse and does not identify the exact optimal number of faces. Therefore, future experiments should continue to reduce the population size during composite construction in a more refined manner, reducing the number of face images on *One Screen*, for example. If reducing the number of faces on *One Screen* from the 18 faces used in this thesis further increases the composite likeness, it is likely that the optimum population size is lower than that used in this thesis, and cognitive load may still reduce composite likeness during construction using *One Screen*. However, if the composite likeness is reduced when fewer than 18 faces are used during the construction procedure, the optimum population size is likely between 18 and 36 faces (between *One* and *Two Screens* in this thesis), and the cognitive load in this condition does not reduce composite likeness.

Face Shape and Texture

The importance of face *Shape* for identifiable composite construction was clearly demonstrated in the final two experiments of this thesis. However, it is still unknown whether the face *Shape* was important for facial recognition during the creation of the

composite or for facial recognition during composite naming. Perhaps a future experiment could use similar techniques to those implemented in research manipulating face *Shape* and *Texture* information of face photographs (see, Benson & Perrett, 1991; Knight & Johnston, 1997; Lee & Perret, 1997; Rogers et al., 2022; Russell & Sinha, 2007) to assess the importance of face *Shape* and *Texture* during the recognition of a facial composite image. In such an experiment, facial composites would be created in Part 1 and would be manipulated to enhance or reduce the *Shape* or *Texture* information for composite naming in Part 2. Such an experiment would provide information about the importance of face *Shape* and *Texture* during the recognition of a facial composite image during composite naming and would extend the literature in this area by determining whether changes during the construction procedure benefit the eyewitness or the recogniser.

Internal and External Features

As composite *Internal Features* were rated as more accurate than *Whole Composites*, future research should display only *Internal Features* for composite naming and likeness rating. Although the literature indicates that whole faces are easier to recognise than their isolated parts (Tanaka & Farrah, 2007), facial composites are imperfect images of faces. Therefore, reducing the facial information displayed also reduced the number of imperfections displayed and, in this thesis, increased the composite likeness ratings. Future research should aim to explore the impact of displaying only composite *Internal Features* for composite likeness ratings but also for composite naming. If composite *Internal Features* continue to be more recognisable than whole face images, facial composites constructed with the police

could be displayed alongside an image of the *Internal Features* to increase the number of criminal perpetrators recognised.

Experimental Power

The number of participants per group for each experiment was determined using historical data. However, the practice of using historical data in this way is now a relatively outdated approach and has resulted in low-powered research in the past that cannot be replicated. Therefore, future research should include more evidence-based methods for calculating the number of participants, such as power calculations using G*power or Bayesian methods. The use of these methods in the future will ensure the research is fully up to date with current approaches and practices.

Dependent Variables

Each experiment included three DVs: composite naming, final composite rating, and intermediate composite rating. However, there were further measurements that were not taken, for example, cognitive load and time on task, and several limitations of the measurements used, such as the ecological validity of likeness ratings.

Measurement of Cognitive Load

Although cognitive load is manipulated in this thesis, no direct measurement of cognitive load was utilised. One subjective measure of cognitive load is the questionnaire designed by Paas (1992), whereby participants are asked to rate the mental effort invested in a task on a 9-point Likert scale. However, the use of Paas' (1992) questionnaire has faced heavy criticism due to the inconsistency between the use of the phrases "task difficulty" and "mental effort" as well as the inconsistency in

timing and frequency of measurement (Sweller et al., 2011). Alternatively, physiological measures of stress such as electrodermal activity (EDA: Posada-Quintero and Chon, 2020), electroencephalography (EEG: Antoneko et al., 2010) and electrooculography (EOG: Belkhiria & Peysakhovich., 2021). Although physiological measures of stress may provide a more reliable measure of cognitive overload (Ayres et al., 2021), conducting such tests during EvoFIT composite construction may induce stress in participants (Ayres et al., 2021), and would not replicate composite construction with the police. Therefore, it was decided that no measure of cognitive load would be taken, and the research would focus on optimising the construction procedure by reducing the number of face arrays. Yet, further research would benefit from implementing a measure of cognitive load to understand the reasoning for the increased likeness of composites created using fewer face arrays.

Specifically, future research may identify a more reliable, suitable measure of cognitive load for use during EvoFIT composite construction, for example, pupillometry, as used in Bafna (2021) as a measure of mental fatigue, which can be caused by heavy cognitive load (Mizuno et al., 2011). Alternatively, future research may devise a short survey for participants to complete at the end of the construction procedure to measure mental fatigue. Unfortunately, there is no current measure to differentiate between the different types of load, which makes it difficult to confirm if it is intrinsic cognitive load being manipulated, as predicted, or whether it is a different type of load: extraneous or germane (Leppink, 2013). Although cognitive load was not directly measured in this thesis, the pattern of results does suggest that reducing the number of screens during composite construction makes the task easier for participants, which supports a theory of cognitive load (Sweller, 1988).

In the knowledge that reducing cognitive load during EvoFIT construction is beneficial to composite likeness, it may also be advantageous to consider the cognitive load that is involved in composite construction using other composite systems, such as E-FIT-V. Although eyewitnesses do not view a pre-determined number of face images during composite construction using E-FIT-V, restricting the number of face images viewed may increase composite identification, as it did when reducing the face arrays viewed during EvoFIT construction.

Measurement of Time on Task

In this PhD research, EvoFIT facial composites were created using *One, Two, Three* or *Four* screens to select the face Shape or Texture. It is sensible to predict that the number of screens used during composite construction has an impact on the length of time spent on the task, with fewer screens resulting in a quicker procedure for participants overall. Reducing the time spent on composite construction would benefit the police, as this would lessen the time spent on composite construction and may increase the number of composites that can be created. Consequently, it would be beneficial to know whether composite construction using *One Screen*, a condition that produced the most identifiable composite images, takes less time when compared to the *Four Screen* baseline condition. In addition, it would be useful to determine whether this time reduction could, in part, explain the increased accuracy rates in the *One Screen* condition. Unfortunately, this measurement was not taken in any of the experiments. In the future, it may be beneficial to replicate Experiment 3 (as the procedure in this experiment most accurately replicates that during composite construction with the police) in a laboratory setting and measure the time taken for composite construction in each condition. If the time taken for composite construction

using *One Screen* is much quicker than that in the other conditions, communicating this information with police forces may encourage the use of EvoFIT to assist in the identification of criminal perpetrators.

Composite Likeness Rating

In each experiment in this thesis, two groups of participants rated the likeness of facial composite images in comparison to the original target photographs. Participants in these groups were unfamiliar with the identities of the targets and were recruited because they did not watch football (in Experiments 1 and 2), did not watch EastEnders (in Experiment 3), did not watch Coronation Street (Experiment 4), or did not watch Emmerdale (Experiment 5).

Likeness ratings are not without limitations, which is important to discuss in the context of this thesis. For example, in previous research, these types of composite likeness ratings have not been found to align with naming rates (see, Frowd et al., 2004; Lam, 2016). This is perhaps due to theoretical differences between familiar (composite naming) and unfamiliar (likeness ratings) face recognition. In fact, one important difference between the recognition of familiar and unfamiliar faces is the facial region utilised for recognition. The internal features of a face are often more important for familiar face recognition; however, there is little difference between the reliance on internal and external features during unfamiliar face recognition (Ellis et al., 1979; Young et al., 1986). One explanation for this difference in reliance on internal and external features may be that people change their external features more frequently, as compared to their internal features, for example, changing the colour, length, or style of their hair. Therefore, people rely on internal features for reliable familiar face recognition. In addition to the importance of internal and external

features, Johnston and Edmonds (2009) highlight further differences between familiar and unfamiliar face recognition, namely the impact of inversion, lighting, movement and caricaturing.

In this research, composite naming data is considered to be the most ecologically valid measure of composite accuracy because it replicates the process of recognising and naming a perpetrator of crime from a facial composite image. Individuals viewing a composite in a case will not be asked to rate its likeness in comparison to a photograph of the perpetrator, so the likeness rating data lacks ecological validity. Despite the lack of ecological validity, likeness ratings are important in research, particularly when facial composites are not identifiable enough for composite naming, for example, in Experiment 1 and composites at the beginning of the construction procedure in Intermediate Composite Rating.

In future research, it may be more ecologically valid to recruit participants familiar with the target identities and invite them to select a target, for example, a footballer, from a line-up of composite images. In addition to composite naming, this would provide a second indication of composite identifiability, using a more ecologically valid procedure, as eyewitnesses are often invited to select a perpetrator from a line-up once they have been identified via a composite image.

The literature indicates that composite construction does not have a significant impact on the accuracy of line-up identification (Davis et al., 2014; Pike et al., 2019; Tredoux et al., 2021; Tsourrai & Davis., 2020). Therefore, it may be predicted that the number of screens used during EvoFIT construction would also have little effect on line-up identification. Holistic composite systems used to explore this effect include EFIT V (Pike et al., 2019) and EFIT 6 (Tsourrai & Davis., 2020), both of which have some flexibility in the number of face images viewed. Using both of these systems,

participants view nine face images on each screen and as many or as few screens as they would like. Alternatively, EvoFIT displays 18 face images on each screen and includes a pre-determined number of screens. Therefore, negative consequences of viewing many face images are less likely to affect participants using EFIT V and EFIT 6, which may explain the lack of impact that composite construction had on line-up identification in this research. However, as EvoFIT invites participants to view a predetermined number of face images and has not been used in line-up research, it is beneficial for future research to explore this effect.

Intermediate Composite Likeness Rating

In all five experiments in this thesis, a measure of likeness was taken for composites at each stage in the construction procedure: after the *First* and *Second Generation*, use of *Holistic Tools* and use of the *Shape Tool*. Typically, the findings for this measure demonstrated that composites increased in likeness throughout the composite procedure. However, in some conditions, composite likeness reduced after the use of certain tools. Yet, this finding was not consistent throughout the experiments, indicating that this measure may not be reliable.

Before this thesis, the change in composite accuracy between the stages of the EvoFIT construction procedure was unknown. However, as this measure has not previously been conducted, the best practice was unknown during the experimentation. In this thesis, all four composite face images (and one random face for comparison) were displayed on screen with the target photograph. Yet, based on the inconsistency of findings between each experiment, it may be more suitable to display composite faces sequentially rather than simultaneously. Measuring

composite likeness in this way in the future may provide a better understanding of the change in composite likeness between the stages of composite construction without the potential influence of other face images on the screen.

Concluding Remarks

This thesis demonstrated that reducing the number of screens viewed during EvoFIT composite construction was beneficial for composite likeness, with composites constructed using only *One Screen* for selection of the face *Shape* and *One Screen* for the selection of the face *Texture* resulting in the most recognisable composites. As the number of face images viewed during the construction procedure was reduced, the number of interacting elements decreased, alluding to a reduction in intrinsic cognitive load and, therefore, a lower likelihood of cognitive overload. This thesis applied the Theory Of Cognitive Load (Sweller, 1988) in a novel way and found that reducing cognitive load supported eyewitness memory, allowing them to construct a more accurate composite.

Moreover, this thesis also demonstrated the importance of face *Shape* for the construction of a recognisable facial composite image. Manipulating the number of screens during EvoFIT composite construction for selection of the face *Shape* had a larger impact than manipulating the number of screens for selection of the face *Texture*. This finding indicates the increased importance of face *Shape* over face *Texture* during composite construction, so future research may concentrate on enhancing the likeness of composite face shapes rather than face *Texture* to further increase composite likeness.

The state-of-the-art face construction procedure developed and tested in this thesis has now been field trailed by EvoFIT customers since the summer of 2022, including police in Lancashire and Greater Manchester. This new procedure has many improvements over its predecessor, including the speed with which a facial composite is produced and the increased rates of identification. Crucially, this new procedure should increase the likelihood of a criminal perpetrator being correctly identified based on a facial composite image, assisting in the arrest of dangerous criminals and reducing crime rates.

Reference List

Abbe, A., & Brandon, S. E. (2013). The role of rapport in investigative interviewing: A review. *Journal of investigative psychology and offender profiling*, *10*(3), 237-249.

Andrews, T. J., Davies-Thompson, J., Kingstone, A., & Young, A. W. (2010). Internal and external features of the face are represented holistically in face-selective regions of visual cortex. *Journal of Neuroscience*, *30*(9), 3544-3552.

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational psychology review*, *22*, 425-438.

Aspley Limited (1993). *E-fit.* Hatfield, UK: Aspley Limited.

Atkinson, R. C., & Shiffrin, R. M. (1968). *Human memory: A proposed system and its control processes*. Academic Press.

Axelrod, V., & Yovel, G. (2010). External facial features modify the representation of internal facial features in the fusiform face area. *Neuroimage*, *52*(2), 720-725.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and instruction*, *16*(5), 389-400.

Ayres, P., & Paas, F. (2009). Interdisciplinary perspectives inspiring a new generation of cognitive load research. *Educational Psychology Review*, *21*, 1-9.

Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.702538

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Baddeley, A. (1998). Recent developments in working memory. *Current opinion in neurobiology*, *8*(2), 234-238.

Baddeley, A. D., & Hitch, G. (2001). *Working memory in perspective*. Psychology Press.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory (Vol. 8). *New York: GA Bower (ed), Recent advances in learning and motivation*.

Baker, B. S. (1985). A new proof for the first-fit decreasing bin-packing algorithm. *Journal of Algorithms*, *6*(1), 49-70.

Baker, E. J. (1999) *The mug-shot search problem: A study of the eigenface metric, search strategies, and interfaces in a system for searching facial image data* [Doctoral thesis, Harvard University]. Harvard University ProQuest Dissertations Publishing. https://www.proquest.com/openview/0f1d7d6509edcc61230cf4602e1257f4/1?pq-origsite=gscholar&cbl=18750&diss=y.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, *12*, 219-228.

Barton, J. J., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations. *Perception*, *35*(8), 1089-1105.

Bartsch, L. M., Singmann, H., & Oberauer, K. (2018). The effects of refreshing and elaboration on working memory performance, and their contributions to long-term memory formation. *Memory & cognition*, *46*, 796-808.

Bayer, F. (2015). *Facial composite production: development of a new technique to identify perpetrators in a more reliable way* (Bachelor's thesis, University of Twente).

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public opinion quarterly*, *71*(2), 287-311.

Belkhiria, C., & Peysakhovich, V. (2021). EOG metrics for cognitive workload detection. *Procedia Computer Science*, *192*, 1875-1884.

Benson, P. J., & Perrett, D. I. (1991). Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology*, *3*(1), 105-135.

Benson, P. J., & Perrett, D. I. (1994). Visual processing of facial distinctiveness. *Perception*, 23(1), 75-93.

Bentley, J. P., & Thacker, P. G. (2004). The influence of risk and monetary payment on the research participation decision making process. *Journal of medical ethics*, *30*(3), 293-298.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American journal of political science*, *58*(3), 739-753.

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 317-333.

Binu, V. S., Mayya, S. S., & Dhar, M. (2014). Some basic aspects of statistical methods and sample size determination in health science research. *Ayu*, *35*(2), 119-123.

Bishara, S. (2021). Psychological availability, mindfulness, and cognitive load in college students with and without learning disabilities. *Cogent Education*, *8*(1). https://doi.org/10.1080/2331186x.2021.1929038.

Bookbinder, J., & Osman, E. (1979). Attentional strategies in dichotic listening. *Memory & Cognition*, *7*, 511-520.

BPS, 2018. *Code of Ethics and Conduct (2018) | BPS*. [online] Bps.org.uk. Available at: <https://www.bps.org.uk/news-and-policy/bps-code-ethics-and-conduct> [Accessed 3 May 2022].

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of vision*, *11*(5), 4-4.

Broadbent, D. E. (1958). Effect of noise on an "intellectual" task. *The Journal of the Acoustical Society of America*, *30*(9), 824-827.

Brown, C., Portch, E., Fodarella, C., Jackson, E., Hancock, P. J. B., Lewis. M. B., Liu, C. H., Marsh, J. E., Tran, L., Wood, E., Damin, E., Robertshaw, L., Date, L., Joyce, S., Brooks, L., & Frowd, C. D. (unpublished). The impact of forensic delay: facilitating facial composite construction using an early-recall retrieval technique.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305-327.

Bruce, V., & Young, A. (1998). *In the eye of the beholder: The science of face perception*. Oxford university press.

Bruce, V., Healey, P., Burton, M., Doyle, T., Coombes, A., & Linney, A. (1991). Recognising facial surfaces. *Perception*, *20*(6), 755-769.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339-360.

Bryant, C. G. (1985). Instrumental positivism in American sociology. *Positivism in Social Theory and Research*, 133-173.

Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments against a configural processing account of familiar face recognition. *Perspectives on Psychological Science*, *10*(4), 482-496.

Butcher, N., Lander, K., Fang, H., & Costen, N. (2011). The effect of motion at encoding and retrieval for same-and other-race face recognition. *British Journal of Psychology*, *102*(4), 931-942.

Byyny, R. L. (2016). Information and cognitive overload. *Pharos*. Available at: https://www.alphaomegaalpha.org/wp-content/uploads/2021/07/2016-4-Byyny.pdf.

Camos, V., & Portrat, S. (2015). The impact of cognitive load on delayed recall. *Psychonomic Bulletin & Review*, *22*, 1029-1034.

Campbell, R., Walker, J., & Baron-Cohen, S. (1995). The development of differential use of inner and outer face features in familiar face identification. *Journal of Experimental Child Psychology*, *59*(2), 196-210.

Chan, J. P., & Ryan, J. D. (2012). Holistic representations of internal and external face features are used to support recognition. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00087

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction*, *8*(4), 293-332.

Chapman, A. F., Hawkins-Elder, H., & Susilo, T. (2018). How robust is familiar face recognition? A repeat detection study of more than 1000 faces. *Royal Society Open Science*, *5*(5). https://doi.org/10.1098/rsos.170634

Clarke, C., Milne, R., & Bull, R. (2011). Interviewing suspects of crime: The impact of PEACE training, supervision and the presence of a legal advisor. *Journal of Investigative Psychology and Offender Profiling*, *8*(2), 149-162.

Collins, R., Lincoln, R., & Frank, M. G. (2002). The effect of rapport in forensic interviewing. *Psychiatry, psychology and law*, *9*(1), 69-78

Comish, S. E. (1987). Recognition of facial stimuli following an intervening task involving the Identi-kit. *Journal of Applied Psychology*, *72*(3), 488-491.

Consumer Insights, Microsoft Canada. (2015). *Attention spans*. Microsoft Attention Spans. Retrieved September 9, 2022, from https://dl.motamem.org/microsoft-attention-spans-research-report.pdf

Conway, M. A. (2009). Episodic memories. *Neuropsychologia*, *47*(11), 2305-2313.

Cootes, T. F., Wheeler, G. V., Walker, K. N., & Taylor, C. J. (2002). View-based active appearance models. *Image and vision computing*, *20*(9-10), 657-664.

Cornish, D. M., & Dukette, D. (2009). *The essential 20: Twenty components of an excellent health care team*. Dorrance Publishing.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Cutler, B. L., Stocklein, C. J., & Penrod, S. D. (1988). Empirical examination of a computerized facial composite production system. *Forensic Reports*.

Danziger, R. (2023). *Positive Social Acts: A Metapragmatic Exploration of the Brighter and Darker Sides of Sociability*. Cambridge University Press.

Davies, G. M., Ellis, H., & Shepherd, J. (1978). Face identification: The influence of delay upon accuracy of Photofit construction. *Journal of Police Science and Administration*, *6*(1), 35-42.

Davies, G., & Oldman, H. (1999). The impact of character attribution on composite production: A real world effect?. *Current Psychology*, *18*, 128-139.

Davies, G., Van der Willik, P., & Morrison, L. J. (2000). Facial composite production: A comparison of mechanical and computer-driven systems. *Journal of Applied Psychology*, *85*(1), 119-124.

Davis, J. P., Gibson, S., & Solomon, C. (2014). The positive influence of creating a holistic facial composite on video line-up identification. *Applied Cognitive Psychology*, *28*(5), 634-639.

Davis, J. P., Sulley, L., Solomon, C., & Gibson, S. (2010, September). A comparison of individual and morphed facial composites created using different systems. In *2010 International Conference on Emerging Security Technologies* (pp. 56-60). IEEE.

Dianiska, R. E., Manley, K. D., & Meissner, C. A. (2021). A process perspective: The importance of theory in eyewitness identification research. *Methods, measures, and theories in eyewitness identification tasks*, 136-168.

Dion Larivière, C., Crough, Q., & Eastwood, J. (2022). The Effects of Rapport Building on Information Disclosure in Virtual Interviews. *Journal of Police and Criminal Psychology*, 1-9.

Dirkx, K. J. H., Skuballa, I., Manastirean-Zijlstra, C. S., & Jarodzka, H. (2021). Designing computer-based tests: Design guidelines from multimedia learning studied with eye tracking. *Instructional Science*, *49*(5), 589-605.

Edward Geiselman, R., & Fisher, R. P. (1988). The cognitive interview: An innovative technique for questioning witnesses of crime. *Journal of Police and Criminal Psychology*, *4*(2), 2-5.

Elliott, E. M., Bell, R., Gorin, S., Robinson, N., & Marsh, J. E. (2022). Auditory distraction can be studied online! A direct comparison between in-person and online experimentation. *Journal of Cognitive Psychology*, *34*(3), 307-324.

Ellis, H. D. (1975). Recognizing faces. *British Journal of Psychology*, *66*(4), 409-426.

Ellis, H. D., Davies, G. M., & Shepherd, J. W. (1978). A critical examination of the Photofit system for recalling faces. *Ergonomics*, *21*(4), 297-307.

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*(4), 431-439.

Erickson, W. B., Brown, C., Portch, E., Lampinen, J. M., Marsh, J. E., Fodarella, C., & Frowd, C. D. (2022). The impact of weapons and unusual objects on the construction of facial composites. *Psychology, Crime & Law*, 1-22

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, *16*(1), 143-149.

FACES Software, Facial Composite Software, FACES 4.0 (n.d.) https://facialcomposites.com/

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149-1160.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Fisher, R. P., & Geiselman, R. E. (2010). The cognitive interview method of conducting police interviews: Eliciting extensive information and promoting therapeutic jurisprudence. *International journal of law and psychiatry*, *33*(5-6), 321-328.

Fodarella, C. (2017). Adjusting the focus of attention: Helping witnesses to evolve a more identifiable composite. *Foresic Research & Criminology International Journal*, *5*(1). https://doi.org/10.15406/frcij.2017.05.00143

Fodarella, C. (2020). *The impact of physical and mental reinstatement of context on the identifiability of facial composites* (Doctoral dissertation, University of Central Lancashire).

Fodarella, C., & Frowd, C. D. (2013). Accuracy of relational and featural information in facial-composite images. In *2013 Fourth International Conference on Emerging Security Technologies* (pp. 16-20). IEEE.

Fodarella, C., Kuivaniemi-Smith, H., Gawrylowicz, J., & Frowd, C. D. (2015). Forensic procedures for facial-composite construction. *Journal of Forensic Practice*.

Fodarella, C., Marsh, J. E., Chu, S., Athwal-Kooner, P., Jones, H. S., Skelton, F. C., Wood, E., Jackson, E., & Frowd, C. D. (2021). The importance of detailed context reinstatement for the production of identifiable composite faces from memory. *Visual Cognition*, *29*(3), 180–200. https://doi.org/10.1080/13506285.2021.1890292

Fortenbaugh, F. C., DeGutis, J., Germine, L., Wilmer, J. B., Grosso, M., Russo, K., & Esterman, M. (2015). Sustained attention across the life span in a sample of 10,000: Dissociating ability and strategy. *Psychological science*, *26*(9), 1497-1510.

Frowd, C. (2015). Facial composites and techniques to improve image recognizability. *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV*, 43-70.

Frowd, C. D. (2017). Facial composite systems: Production of an identifiable face. In M. Bindemann and A. Megreya (Eds.) Face Processing: Systems, Disorders and Cultural Differences (pp. 55 - 86). Nova Science.

Frowd, C. D. (2021). Forensic facial composites. In *Methods, Measures, and Theories in Eyewitness Identification Tasks* (pp. 34-64). Routledge.

Frowd, C. D., Bruce, V., Smith, A. J., & Hancock, P. J. (2008). Improving the quality of facial composites using a holistic cognitive interview. *Journal of Experimental Psychology: Applied*, *14*(3), 276-287.

Frowd, C. D., Carson, D., Ness, H., McQuiston-Surrett, D., Richardson, J., Baldwin, H., & Hancock, P. (2005). Contemporary composite techniques: The impact of a

forensically-relevant target delay. *Legal and Criminological Psychology*, *10*(1), 63-81.

Frowd, C. D., Carson, D., Ness, H., Richardson, J., Morrison, L., Mclanaghan, S., & Hancock, P. (2005). A forensically valid comparison of facial composite systems. *Psychology, Crime & Law*, *11*(1), 33-52.

Frowd, C. D., Erickson, W. B., Lampinen, J. M., Skelton, F. C., McIntyre, A. H., & Hancock, P. J. (2015). A decade of evolving composites: regression-and meta-analysis. *Journal of Forensic Practice*, *17*(4), 319-334.

Frowd, C. D., & Hancock, P. J. (2008). Evolving human faces. *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, 189-210.

Frowd, C. D., Hancock, P. J., & Carson, D. (2004). EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on applied perception (TAP)*, *1*(1), 19-39.

Frowd, C. D., McQuiston-Surrett, D., Anandaciva, S., Ireland, C. G., & Hancock, P. J. (2007). An evaluation of US systems for facial composite production. *Ergonomics*, *50*(12), 1987-1998.

Frowd, C. D., Pitchford, M., Bruce, V., Jackson, S., Hepton, G., Greenall, M., & Hancock, P. J. (2011). The psychology of face construction: giving evolution a helping hand. *Applied Cognitive Psychology*, *25*(2), 195-203

Frowd, C. D., Pitchford, M., Skelton, F., Petkovic, A., Prosser, C., & Coates, B. (2012). Catching even more offenders with evofit facial composites. *2012 Third International Conference on Emerging Security Technologies*. https://doi.org/10.1109/est.2012.26

Frowd, C. D., Pitchford, M., Skelton, F., Petkovic, A., Prosser, C., & Coates, B. (2012). Catching even more offenders with EvoFIT facial composites. In *2012 Third International Conference on Emerging Security Technologies* (pp. 20-26). IEEE.

Frowd, C. D., Portch, E., Killeen, A., Mullen, L., Martin, A. J., & Hancock, P. J. (2019). EvoFIT facial composite images: a detailed assessment of impact on forensic practitioners, police investigators, victims, witnesses, offenders and the media. In *2019 eighth international conference on emerging security technologies (EST)* (pp. 1-7). IEEE.

Frowd, C. D., Ramsay, S., & Hancock, P. J. (2011). The influence of holistic interviewing on hair perception for the production of facial composites. *International Journal of Bio-Science and Bio-Technology*, *3*(3), 55-64.

Frowd, C. D., Skelton, F., Atherton, C., Pitchford, M., Hepton, G., Holden, L., & Hancock, P. J. (2012). Recovering faces from memory: The distracting influence of external facial features. *Journal of Experimental Psychology: Applied*, *18*(2), 224.

Frowd, C. D., Skelton, F., Hepton, G., Holden, L., Minahil, S., Pitchford, M., & Hancock, P. J. (2013). Whole-face procedures for recovering facial images from memory. *Science & Justice*, *53*(2), 89-97.

Frowd, C. D., Underwood, S., Athwal, P., Lampinen, J. M., Erickson, W. B., Mahony, G., & Marsh, J. E. (2015). Facial stereotypes and perceived mental illness. In *2015 Sixth International Conference on Emerging Security Technologies (EST)* (pp. 62-68). IEEE.

Frowd, C., & Hepton, G. (2009). The benefit of hair for the construction of facial composite images. *The British Journal of Forensic Practice*.

Frowd, C., Bruce, V., McIntyre, A., & Hancock, P. (2007). The relative importance of external and internal features of facial composites. *British journal of psychology*, *98*(1), 61-77.

Gallard, R. H., & Esquivel, S. C. (2001). Enhancing evolutionary algorithms through recombination and parallelism. *Journal of Computer Science & Technology*, *1*.

Gambarota, F., & Sessa, P. (2019). Visual working memory for faces and facial expressions as a useful "tool" for understanding social and affective cognition. *Frontiers in psychology*, *10*, 2392. https://spiral.imperial.ac.uk/bitstream/10044/1/20424/2/Garneau-MJ-1973-PhD-Thesis.pdf.

Garcia-Solley, E. (2019). *The Use of Facial Composites in Person Identification*. University of Kent (United Kingdom).

Garneau, M. J. P. (1973). The perception of facial images. [Doctoral thesis, Imperial College, University of London].

Gawrylowicz, J., Gabbert, F., Carson, D., Lindsay, W. R., & Hancock, P. J. (2012). Holistic versus featural facial composite systems for people with mild intellectual disabilities. *Applied Cognitive Psychology*, *26*(5), 716-720.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1986). Enhancement of eyewitness memory with the cognitive interview. *The American journal of psychology*, 385-401.

George, B., Gibson, S. J., Maylin, M. I., & Solomon, C. J. (2008). EFIT-V- interactive evolutionary strategy for the construction of photo-realistic facial composites. In *Proceedings of the 10th annual conference on genetic and evolutionary computation* (pp. 1485-1490).

Gerjets, P., Scheiter, K., & Catrambone, R. (2004). Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures. *Instructional Science*, *32*, 33-58.

Giannou, K., Frowd, C. D., Taylor, J. R., & Lander, K. (2021). Mindfulness in face recognition: Embedding mindfulness instructions in the face-composite construction process. *Applied Cognitive Psychology*, *35*(4), 999-1010.

Gibling, F., & Bennett, P. (1994). Artistic enhancement in the production of Photo-FIT likenesses: An examination of its effectiveness in leading to suspect identification. *Psychology, Crime and Law*, *1*(1), 93-100.

Gibson, S. J., Solomon, C. J., Maylin, M. I., & Clark, C. (2009). New methodology in facial composite construction: From theory to practice. *International Journal of Electronic Security and Digital Forensics*, *2*(2), 156-168.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of verbal learning and verbal behavior*, *5*(4), 351-360.

Goldberg, D. E. (1989). Sizing populations for serial and parallel genetic algorithms. In *Proceedings of 3rd international conference on genetic algorithms* (pp. 70-79).

Hacking, I. (1981). Scientific revolutions. *Oxford University Press.*

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in cognitive sciences*, *4*(9), 330-337.

Haridat, S. H. (2016). *Facial composites: a comparison between a traditional and new technique for conducting facial composite techniques used in the field of investigation* (Master's thesis, University of Twente).

Hartson, R., & Pyla, P. S. (2012). Rigorous Empirical Evaluation: Preparation. *The UX Book*, 503-536.

Hasel, L. E., & Wells, G. L. (2007). Catching the bad guy: Morphing composite faces helps. *Law and human Behavior*, *31*, 193-207.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in cognitive sciences*, *11*(7), 299-306.

Havard, C. (2021). The importance of internal and external features in matching own and other race faces. *Perception*, *50*(10), 861-875.

Heidtman, J., Wysienska, K., & Szmatka, J. (2000). Positivism and types of theories in sociology. *Sociological Focus*, *33*(1), 1-26.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431-440.

Hinz, T., & Pezdek, K. (2001). The effect of exposure to multiple lineups on face identification accuracy. *Law and Human Behavior*, *25*(2), 185-198.

Hjelmas, E., & Wroldsen, J. (1999, June). Recognizing faces from the eyes only. In *Proceedings of the Scandinavian Conference on Image Analysis* (Vol. 2, pp. 659-664).

Holland John, H. (1975). Adaptation in natural and artificial systems. *Ann Arbor: University of Michigan Press*.

Holland, J. H. (2000). Building blocks, cohort genetic algorithms, and hyperplane-defined functions. *Evolutionary computation*, *8*(4), 373-391.

Holland, M. K., & Tarlow, G. (1972). Blinking and mental load. *Psychological Reports*, *31*(1), 119-127.

Homa, G. (1983). The law enforcement composite sketch artist. *Berlin, NJ: Community*.

IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Innocence Project. (n.d.). Innocence Project. Retrieved July 20, 2023, from https://innocenceproject.org/

Itz, M. L., Golle, J., Luttmann, S., Schweinberger, S. R., & Kaufmann, J. M. (2017). Dominance of texture over shape in facial identity processing is modulated by individual abilities. *British Journal of Psychology*, *108*(2), 369-396.

Jackson, M. C., & Raymond, J. E. (2008). Familiarity enhances visual working memory for faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 556.

Jeung, H. J., Chandler, P., & Sweller, J. (1997). The role of visual indicators in dual sensory mode instruction. *Educational Psychology*, *17*(3), 329-345.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577-596.

Jones, T., & Forrest, S. (1995, July). Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *ICGA* (Vol. 95, pp. 184-192).

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional science*, *38*(2), 105-134.

Jongerius, C., Hessels, R. S., Romijn, J. A., Smets, E. M., & Hillen, M. A. (2020). The measurement of eye contact in human interactions: A scoping review. *Journal of Nonverbal Behavior*, *44*, 363-389

Kaesler, M., Dunn, J. C., Ransom, K., & Semmler, C. (2020). Do sequential lineups impair underlying discriminability?. *Cognitive research: principles and implications*, *5*, 1-21.

Kalyuga, S. (2009). Knowledge elaboration: A cognitive load perspective. *Learning and Instruction*, *19*(5), 402-410.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need?. *Educational Psychology Review*, *23*, 1-19.

Kaufmann, J. M., Schulz, C., & Schweinberger, S. R. (2013). High and low performers differ in the use of shape information for face recognition. *Neuropsychologia*, *51*(7), 1310-1319.

Kellogg, R. T. (2007). Are written and spoken recall of text equivalent?. *The American Journal of Psychology*, *120*(3), 415-428.

Kempen, K., & Tredoux, C. G. (2012). 'Seeing is believing': the effect of viewing and constructing a composite on identification performance. *South African Journal of Psychology*, *42*(3), 434-444

Kieckhaefer, J. M., Vallano, J. P., & Schreiber Compo, N. (2014). Examining the positive effects of rapport building: When and why does rapport building benefit adult eyewitness memory?. *Memory*, *22*(8), 1010-1023.

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological bulletin*, *100*(1), 78-100.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01997

Knight, B., & Johnston, A. (1997). The role of movement in face recognition. *Visual cognition*, *4*(3), 265-273.

Koehn, C. E., & Fisher, R. P. (1997). Constructing facial composites with the Mac-a-Mug Pro system. *Psychology, Crime and Law*, *3*(3), 209-218.

Kovera, M. B., Penrod, S. D., Pappas, C., & Thill, D. L. (1997). Identification of computer-generated facial composites. *Journal of Applied Psychology*, *82*(2), 235-246.

Kramer, R. S. (2021). Forgetting faces over a week: investigating self-reported face recognition ability and personality. *PeerJ*, *9*, e11828.

Kramer, R. S., Manesi, Z., Towler, A., Reynolds, M. G., & Burton, A. M. (2018). Familiarity and within-person facial variability: The importance of the internal and external features. *Perception*, *47*(1), 3-15.

Kuivaniemi-Smith, H. J., Nash, R. A., Brodie, E. R., Mahoney, G., & Rynn, C. (2014). Producing facial composite sketches in remote Cognitive Interviews: A preliminary investigation. *Psychology, Crime & Law*, *20*(4), 389-406.

Lange, C., Costley, J., & Han, S. L. (2017). The effects of extraneous load on the relationship between self-regulated effort and germane load within an e-learning environment. *International Review of Research in Open and Distributed Learning*, *18*(5), 64-83.

Latif, M., & Moulson, M. (2021). The Importance of Internal and External Features in Face Recognition. *Journal of Vision*, *21*(9), 2190-2190.

Laughery, K. R., & Fowler, R. H. (1980). Sketch artist and Identi-kit procedures for recalling faces. *Journal of Applied Psychology*, *65*(3), 307-316.

Laughery, K. R., & Wogalter, M. S. (1989). Forensic applications of facial memory research. In A. W. Young & H. D. Ellis (Eds.), *Handbook of research on face processing* (pp. 519-555). Amsterdam: North Holland.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human perception and performance*, *21*(3), 451.

Lavie, N. (2001). Capacity limits in selective attention: Behavioral evidence and implications for neural activity. *Visual attention and cortical circuits*, 49-68.

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current directions in psychological science*, *19*(3), 143-148.

Lavie, N., Ro, T., & Russell, C. (2003). The role of perceptual load in processing distractor faces. *Psychological science*, *14*(5), 510-515.

Lee, K. J., & Perrett, D. (1997). Presentation-time measures of the effects of manipulations in colour space on discrimination of famous faces. *Perception*, *26*(6), 733-752.

Lee, K. J., & Perrett, D. I. (2000). Manipulation of colour and shape information and its consequence upon recognition and best-likeness judgments. *Perception*, *29*(11), 1291-1312.

Leppink, J., & van den Heuvel, A. (2015). The evolution of cognitive load theory and its application to medical education. *Perspectives on medical education*, *4*, 119-127.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology.*

Limbach, K., Itz, M. L., Schweinberger, S. R., Jentsch, A. D., Romanova, L., & Kaufmann, J. M. (2022). Neurocognitive effects of a training program for poor face recognizers using shape and texture caricatures: A pilot investigation. *Neuropsychologia*, *165*, 108133. https://doi.org/10.1016/j.neuropsychologia.2021.108133

Lin, W., Strube, M. J., & Roediger, H. L. (2019). The effects of repeated lineups and delay on eyewitness identification. *Cognitive research: principles and implications*, *4*, 1-19.

Lindsay, R. C. L., Nosworthy, G. J., Martin, R., & Martynuck, C. (1994). Using mug shots to find suspects. *Journal of Applied Psychology*, *79*(1), 121

Lindsay, R. C., Ross, D. F., Read, J. D., & Toglia, M. P. (Eds.). (2013). *The handbook of eyewitness psychology: volume ii: memory for people* (Vol. 2). Psychology Press.

Lobo, F. G., & Goldberg, D. E. (2004). The parameter-less genetic algorithm in practice. *Information Sciences*, *167*(1-4), 217-232.

Lobo, F. G., & Lima, C. F. (2005, June). A review of adaptive population sizing schemes in genetic algorithms. In *Proceedings of the 7th annual workshop on Genetic and evolutionary computation* (pp. 228-234).

Loftus, E. F., & Cahill, L. (2007). Memory distortion: From misinformation to rich false memory. *The foundations of remembering: Essays in honor of Henry L. Roediger III*, 413-425.

Louw, A. (2021). Cognitive load theory in simulations to facilitate critical thinking in radiography students. *African Journal of Health Professions Education*, *13*(1), 41-46.

Low, R., & Sweller, J. (2005). The modality principle in multimedia learning. *The Cambridge handbook of multimedia learning*, *147*, 158.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, *17*(8), 391-400.

Luria, S. M., & Strauss, M. S. (2013). Comparison of eye movements over faces in photographic positives and negatives. *Perception*, *42*(11), 1134-1143.

Marr, C., Otgaar, H., Sauerland, M., Quaedflieg, C. W., & Hope, L. (2021). The effects of stress on eyewitness memory: A survey of memory experts and laypeople. *Memory & Cognition*, *49*, 401-421.

Marsh, J. E., Demaine, J., Bell, R., Skelton, F. C., Frowd, C. D., Röer, J. P., & Buchner, A. (2015). The impact of irrelevant auditory facial descriptions on memory for target faces: Implications for eyewitness memory. *Journal of Forensic Practice*, *17*(4), 271-280.

Martin, A. J., Peter, J. H., Frowd, C. D., Heard, P., Gaskin, E., Ford, C., & Hewett, T. (2018). EvoFIT composite face construction via practitioner interviewing and a witness-administered protocol. In *2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)* (pp. 311-316). IEEE.

Masip, J., Garrido, E., Herrero, C., Ullán, A. M., & Conde, J. (2012). Teaching students about facial composites using the FACES software. *Teaching of Psychology*, *39*(2), 137-141.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of educational psychology*, *93*(1), 187.

McQuiston-Surrett, D., Topp, L. D., & Malpass, R. S. (2006). Use of facial composite systems in US law enforcement agencies. *Psychology, Crime & Law*, *12*(5), 505-517

Meijer, G. T. (2019). More than meets the eye: Processing of visual and auditory information in the sensory cortex.

Michalos, A. C. (2014). Encyclopedia of quality of life and well-being research. Springer.

Miller, G. (1956). Human memory and the storage of information. *IRE Transactions on Information Theory*, *2*(3), 129-137.

Min, J. (2017). Effects of the use of social network sites on task performance: Toward a sustainable performance in a distracting work environment. *Sustainability*, *9*(12), 2270.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT.

Mizuno, K., Tanaka, M., Yamaguti, K., Kajimoto, O., Kuratsune, H., & Watanabe, Y. (2011). Mental fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity. *Behavioral and brain functions*, *7*(1), 1-7.

Mora-Melià, D., Martínez-Solano, F. J., Iglesias-Rey, P. L., & Gutiérrez-Bahamondes, J. H. (2017). Population size influence on the efficiency of evolutionary algorithms to design water networks. *Procedia Engineering*, *186*, 341-348.

Moreno, R. (2010). Cognitive load theory: More food for thought.

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, *64*(5), 482-488.

Murphy, G., & Greene, C. M. (2016). Perceptual load affects eyewitness accuracy and susceptibility to leading questions. *Frontiers in psychology*, *7*, 1322.

Nakabayashi, K., Lloyd-Jones, T. J., Butcher, N., & Liu, C. H. (2012). Independent influences of verbalization and race on the configural and featural processing of faces: a behavioral and eye movement study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 61-77.

Nancarrow, C., & Penn, S. (1998). Rapport in telemarketing-mirror, mirror on the call?. *Marketing Intelligence & Planning*, *16*(1), 12-21.

Nash, R. A., Houston, K. A., Ryan, K., & Woodger, N. (2014). Remembering remotely: would video-mediation impair witnesses' memory reports?. *Psychology, Crime & Law*, *20*(8), 756-768

Nash, R. A., Nash, A., Morris, A., & Smith, S. L. (2016). Does rapport-building boost the eyewitness eyeclosure effect in closed questioning?. *Legal and Criminological Psychology*, *21*(2), 305-318.

Nejati, H., Sim, T., & Martinez-Marroquin, E. (2011, October). Do you see what i see? A more realistic eyewitness sketch recognition. In *2011 International Joint Conference on Biometrics (IJCB)* (pp. 1-8). IEEE.

Nelson, N., & Mondloch, C. (2014). Is He Afraid or Looking at a Spider? Visual Attention to Facial Expressions Varies With the Task. *Journal of Vision*, *14*(10), 1389-1389.

Nishida, S., Shibata, T., & Ikeda, K. (2009). Prediction of human eye movements in facial discrimination tasks. *Artificial Life and Robotics*, *14*, 348-351.

Nyman, T. J., Antfolk, J., Lampinen, J. M., Korkman, J., & Santtila, P. (2020). Line-up image position in simultaneous and sequential line-ups: The effects of age and viewing distance on selection patterns. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01349

O'Donnell, C., & Bruce, V. (2001). Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception*, *30*(6), 755-764.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of educational psychology*, *84*(4), 429.

Paas, F., & Ayres, P. (2014). Cognitive load theory: A broader view on the role of memory in learning and education. *Educational Psychology Review*, *26*, 191-195.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, *38*(1), 1-4.

Perkins, D. (1975). A definition of caricature and caricature and recognition. *Studies in Visual Communication*, *2*(1), 1-24.

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology*, *58*(3), 193.

Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences*, *109*(48), E3314-E3323.

Peterson, M., Cox, I., & Eckstein, M. (2008). The use of the eyes for human face recognition explained through information distribution analysis. *Journal of Vision*, *8*(6), 894-894.

Pike, G. E., Brace, N. A., Turner, J., Ness, H., & Vredeveldt, A. (2019). Advances in facial composite technology, utilizing holistic construction, do not lead to an increase in eyewitness misidentifications compared to older feature-based systems. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01962.

Pitchford, M., Green, D., & Frowd, C. D. (2017). The impact of misleading information on the identifiability of feature-based facial composites. In *2017 Seventh International Conference on Emerging Security Technologies (EST)* (pp. 185-190).

Pohl, C., Kiesel, A., Kunde, W., & Hoffmann, J. (2010). Early and late selection in unconscious information processing. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(2), 268-285.

Portch, E., Logan, K., & Frowd, C. D. (2017). Interviewing and visualisation techniques: attempting to further improve EvoFIT facial composites. In *2017 Seventh International Conference on Emerging Security Technologies (EST)* (pp. 97-102). IEEE.

Posada-Quintero, H. F., Florian, J. P., Orjuela-Cañón, A. D., Aljama-Corrales, T., Charleston-Villalobos, S., & Chon, K. H. (2016). Power spectral density analysis of electrodermal activity for sympathetic function assessment. *Annals of biomedical engineering*, *44*, 3124-3135.

Pozzulo, J., Pica, E., & Sheahan, C. (2019). *Familiarity and conviction in the criminal justice system: Definitions, theory, and eyewitness research*. Oxford University Press.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive psychology*, *3*(3), 382-407.

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive psychology*, *19*(4), 473-497.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological science*, *22*(4), 464-471.

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2009). Holistic processing of faces happens at a glance. *Vision research*, *49*(23), 2856-2861.

Ricker, T. J., Vergauwe, E., & Cowan, N. (2016). Decay theory of immediate memory: From Brown (1958) to today (2014). *Quarterly Journal of Experimental Psychology*, *69*(10), 1969-1995.

Rizzo, M., Hurtig, R., & Damasio, A. R. (1987). The role of scanpaths in facial recognition and learning. *Annals of neurology*, *22*(1), 41-45.

Rodrigues, P. F., & Pandeirada, J. N. (2015). Attention and working memory in elderly: the influence of a distracting environment. *Cognitive processing*, *16*, 97-109.

Rogers, D., Baseler, H., Young, A. W., Jenkins, R., & Andrews, T. J. (2022). The roles of shape and texture in the recognition of familiar faces. *Vision Research*, *194*, 108013.

Royer, J., Blais, C., Charbonneau, I., Déry, K., Tardif, J., Duchaine, B., & Fiset, D. (2018). Greater reliance on the eye region predicts better face recognition ability. *Cognition*, *181*, 12-20.

Russell, R., & Sinha, P. (2007). Real-world face recognition: The importance of surface reflectance properties. *Perception*, *36*(9), 1368-1374.

Schiewe, M. H., Landahl, J. T., Myers, M. S., Plesha, P. D., Jacques, F. J., Stein, J. E., McCain, B. B., Weber, D, D., Chan, S., & Varanasi, U. (1988). Relating field and laboratory studies: Cause-and-effect research. In *Proceedings of the First Annual Meeting on Puget Sound Research* (Vol. 1, pp. 577-584).

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, *20*(1), 11.

Sewell, J. L., Santhosh, L., & O'Sullivan, P. S. (2020). How do attending physicians describe cognitive overload among their workplace learners?. *Medical Education*, *54*(12), 1129-1136.

Shallice, T., & Warrington, E. K. (1970). Independent functioning of verbal memory stores: A neuropsychological study. *The Quarterly journal of experimental psychology*, *22*(2), 261-273.

Shepherd, J. W., Ellis, H. D., McMurran, M., & Davies, G. M. (1978). Effect of character attribution on Photofit construction of a face. *European journal of social psychology*.

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Josa a*, *4*(3), 519-524.

Skelton, F. C., Frowd, C. D., & Andrews, S. (n.d.). *Witness Interviews: Does recall of relational information improve identifiability of a facial composite?*. Napier Repository. Retrieved June 4, 2022, from http://researchrepository.napier.ac.uk/id/eprint/9540.

Skelton, F. C., Frowd, C. D., Hancock, P. J., Jones, H. S., Jones, B. C., Fodarella, C., Battersby, K., & Logan, K. (2020). Constructing identifiable composite faces: The importance of cognitive alignment of interview and construction procedure. *Journal of Experimental Psychology: Applied*, *26*(3), 507–521. https://doi.org/10.1037/xap0000257

Sobel, N. R., Pridgen, D., Vogelman, L. A., & Ruoff, D. W. (1981). *Eyewitness identification: Legal and practical problems*. Boardman.

Solomon, C., Gibson, S., & Maylin, M. (2012). EFIT-V: Evolutionary algorithms and computer composites. In C. Wilkinson & C. Rynn (Eds.), *Craniofacial Identification.* Cambridge University Press.

Sporer, S. L., Tredoux, C. G., Vredeveldt, A., Kempen, K., & Nortje, A. (2020). Does exposure to facial composites damage eyewitness memory? A comprehensive review. *Applied Cognitive Psychology*, *34*(5), 1166-1179.

Stephan, C. N., Caple, J. M., Guyomarc'h, P., & Claes, P. (2019). An overview of the latest developments in facial imaging. *Forensic sciences research*, *4*(1), 10-28.

Sullivan, T. (2007). *Comparing featural and holistic composite systems with the aid of guided memory techniques* (Master's thesis, University of Cape Town).

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257-285.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, *22*, 123-138.

Sweller, J., Ayres, P., Kalyuga, S., Sweller, J., Ayres, P., & Kalyuga, S. (2011). Altering element interactivity and intrinsic cognitive load. *Cognitive Load Theory*, 203-218.

Szulewski, A., Howes, D., van Merriënboer, J. J., & Sweller, J. (2020). From theory to practice: the application of cognitive load theory to the practice of medicine. *Academic Medicine*, *96*(1), 24-30.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly journal of experimental psychology*, *46*(2), 225-245.

Taylor, D. A. (2012). *Featural and holistic processing in facial composite construction: The role of cognitive style and processing sets* (Doctoral dissertation, University of Westminster).

Taylor, T. A., Kamel-ElSayed, S., Grogan, J. F., Hajj Hussein, I., Lerchenfeldt, S., & Mohiyeddini, C. (2022). Teaching in uncertain times: Expanding the scope of extraneous cognitive load in the cognitive load theory. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.665835

Towse, J. N., Cowan, N., Hitch, G. J., & Horton, N. J. (2008). The recall of information from working memory: Insights from behavioural and chronometric perspectives. *Experimental Psychology*, *55*(6), 371-383.

Tredoux, C. G., Sporer, S. L., Vredeveldt, A., Kempen, K., & Nortje, A. (2021). Does constructing a facial composite affect eyewitness memory? A research synthesis and meta-analysis. *Journal of Experimental Criminology*, *17*, 713-741.

Treisman, A. M. (1964). The effect of irrelevant material on the efficiency of selective listening. *The American journal of psychology*, *77*(4), 533-546.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson, *Organization of memory.* Academic Press.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual review of psychology*, *53*(1), 1-25.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *80*(5), 352-373.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, *3*(1), 71-86.

Turoman, N., & Vergauwe, E. (2023). The effect of multisensory distraction on working memory: A role for task relevance? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.31234/osf.io/fd9xh.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43*(2), 161-204.

Valentine, T. (1999). Face-Space Models of Face Recognition. In M. J. Wenger & T. J. Townsend (Eds.) *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges.* Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc.

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, *15*(5), 525-535.

Valentine, T., & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(2), 151-161.

Valentine, T., Davis, J. P., Thorner, K., Solomon, C., & Gibson, S. (2010). Evolving and combining facial composites: Between-witness and within-witness morphs compared. *Journal of Experimental Psychology: Applied*, *16*(1), 72.

Vallano, J. P., & Compo, N. S. (2011). A comfortable witness is a good witness: Rapport-building and susceptibility to misinformation in an investigative mock-crime interview. *Applied cognitive psychology*, *25*(6), 960-970.

Vallesi, A., Tronelli, V., Lomi, F., & Pezzetta, R. (2021). Age differences in sustained attention tasks: A meta-analysis. *Psychonomic Bulletin & Review*, *28*(6), 1755–1775.

Van Merriënboer, J. J., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical education*, *44*(1), 85-93.

Vredeveldt, A., & Tredoux, C. G. (2022). Composite communication: how dissemination of facial composites in the media affects police investigations. *Memory, Mind & Media*, *1*.

Walsh, D. W., & Milne, R. (2008). Keeping the PEACE? A study of investigative interviewing practices in the public sector. *Legal and Criminological Psychology*, *13*(1), 39-57.

Wells, G. L., & Hasel, L. E. (2007). Facial composite production by eyewitnesses. *Current Directions in Psychological Science*, *16*(1), 6-10.

Wells, G. L., & Hryciw, B. (1984). Memory for faces: Encoding and retrieval operations. *Memory & Cognition*, *12*(4), 338-344.

Wells, G. L., Charman, S. D., & Olson, E. A. (2005). Building face composites can harm lineup identification performance. *Journal of experimental psychology: Applied*, *11*(3), 147.

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological science in the public interest*, *7*(2), 45-75.

White, M. (2001). Effect of photographic negation on matching the expressions and identities of faces. *Perception*, *30*(8), 969-981.

Wittgenstein, L. (1921). *Logical-Philosophical Treatise*

Wogalter, M. S., & Bradley Marwitz, D. (1991). Face composite construction: In-view and from-memory quality and improvement with practice. *Ergonomics*, *34*(4), 459-468.

Wolfs, A. C., Sneyd, D., Vallano, J. P., Schreiber Compo, N., & Reinoso, L. (2022). The effects of building and maintaining rapport on cooperative mock eyewitness recall. *Journal of Investigative Psychology and Offender Profiling*, *19*(3), 151-166.

Wood, E., & Zivcakova, L. (2015). Understanding multimedia multitasking in educational settings. *The Wiley handbook of psychology, technology, and society*, 404-419.

Woodman, G. F., Vogel, E. K., & Luck, S. J. (2012). Flexibility in visual working memory: Accurate change detection in the face of irrelevant variations in position. *Visual cognition*, *20*(1), 1-28.

Xu, Y. (2002). Limitations of object-based feature encoding in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 458-468.

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.

Young, A. W., McWeeny, K. H., Hay, D. C., & Ellis, A. W. (1986). Matching familiar and unfamiliar faces on identity and expression. *Psychological research*, *48*(2), 63-68.

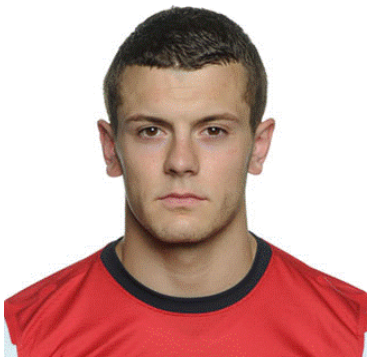
Zahradnikova, B., Duchovicova, S., & Schreiber, P. (2018). Facial composite systems. *Artificial Intelligence Review*, *49*, 131-152.

Zvyagintsev, M., Clemens, B., Chechko, N., Mathiak, K. A., Sack, A. T., & Mathiak, K. (2013). Brain networks underlying mental imagery of auditory and visual information. *European Journal of Neuroscience*, *37*(9), 1421-1434.

Appendices

Appendix 1: Experiment 1 Targets and Composites

1 Screen | 2 Screens | 3 Screens | 4 Screens



1 Screen | 2 Screens | 3 Screens | 4 Screens

1 Screen 2 Screens 3 Screens 4 Screens

1 Screen 2 Screens 3 Screens 4 Screens

1 Screen 2 Screens 3 Screens 4 Screens

1 Screen 2 Screens 3 Screens 4 Screens

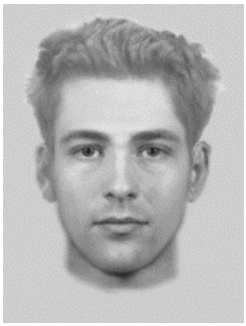
1 Screen 2 Screens 3 Screens 4 Screens

1 Screen 2 Screens 3 Screens 4 Screens

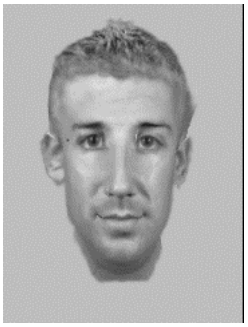
1 Screen 2 Screens 3 Screens 4 Screens

1 Screen 2 Screens 3 Screens 4 Screens

Appendix 2: Experiment 2 Targets and Composites



1 Screen



2 Screens

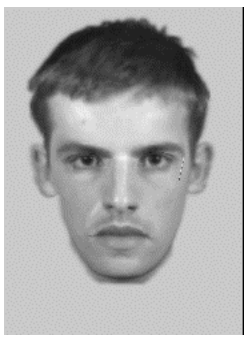
3 Screens

4 Screens

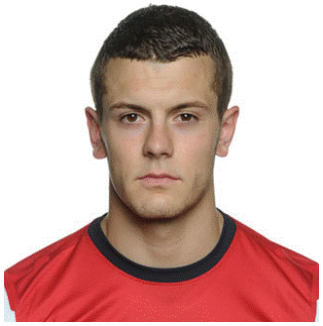
1 Screen

2 Screens

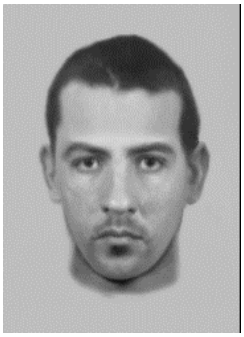
3 Screens



4 Screens

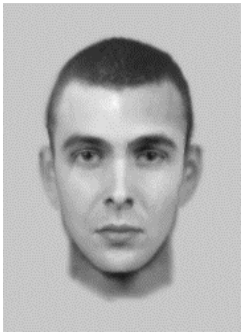


1 Screen

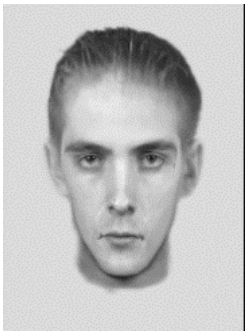


2 Screens

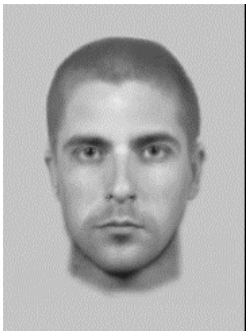
3 Screens



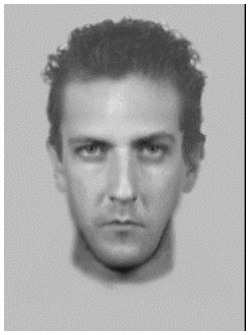
4 Screens



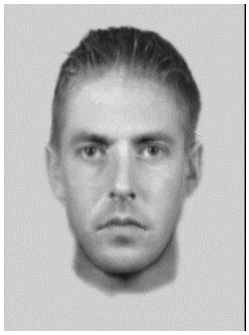
1 Screen



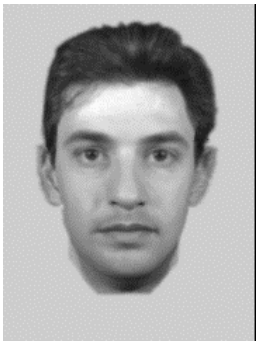
2 Screens



3 Screens

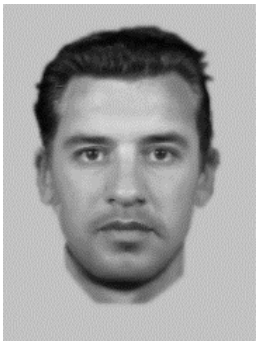


4 Screens

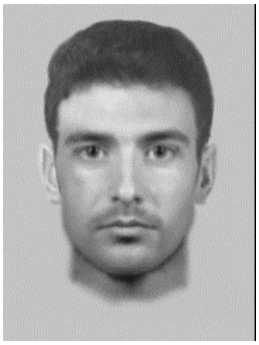


1 Screen

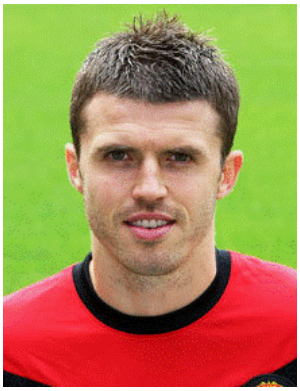
2 Screens

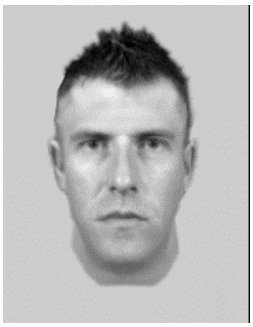


3 Screens



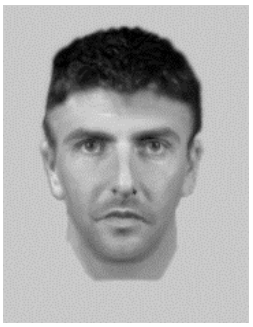
4 Screens



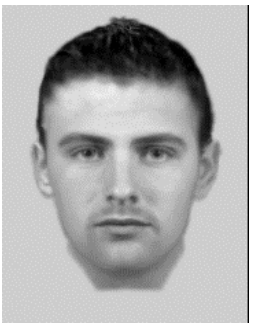


1 Screen

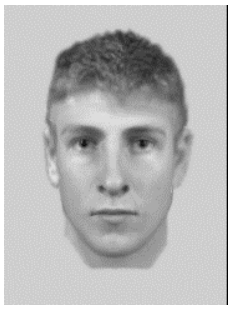
2 Screens



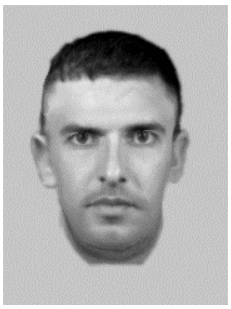
3 Screens



4 Screens



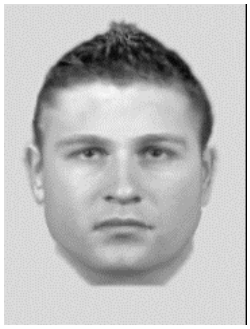
1 Screen



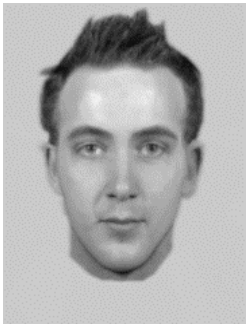
2 Screens

3 Screens

4 Screens



1 Screen

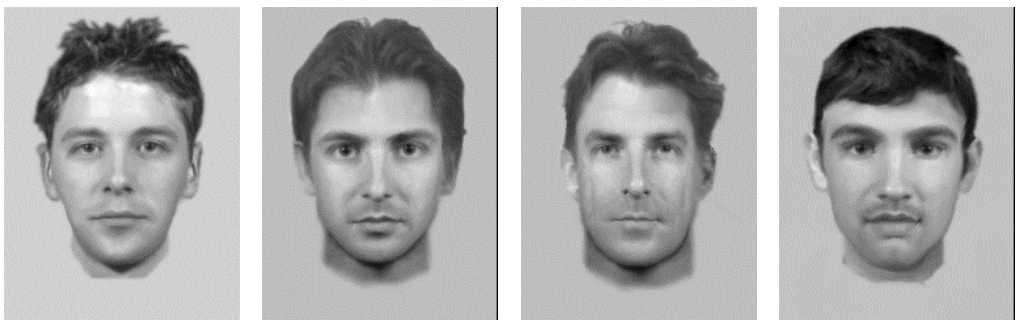


2 Screens

3 Screens

4 Screens

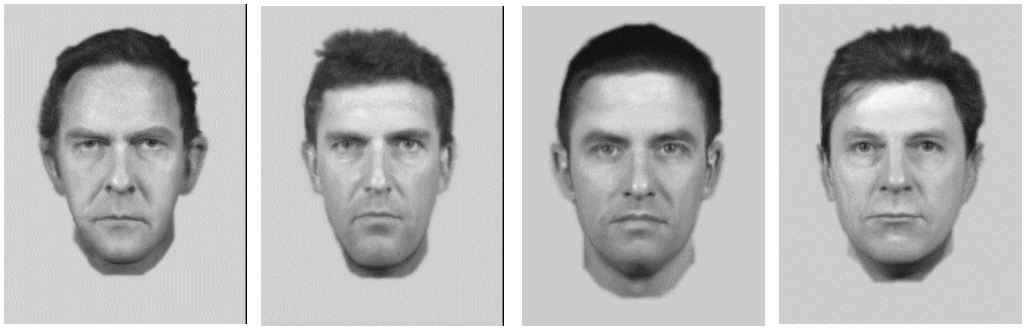
1 Screen | 2 Screens | 3 Screens | 4 Screens



1 Screen | 2 Screens | 3 Screens | 4 Screens

Appendix 3: Experiment 3 Targets and Composites

1 Screen | 2 Screens | 3 Screens | 4 Screens

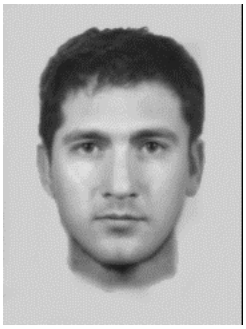


1 Screen | 2 Screens | 3 Screens | 4 Screens

1 Screen | 2 Screens | 3 Screens | 4 Screens

1 Screen | 2 Screens | 3 Screens | 4 Screens

1 Screen

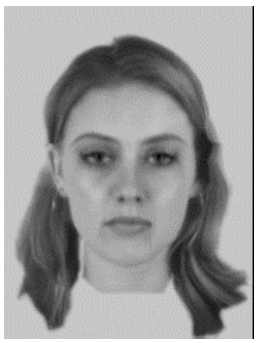


2 Screens

3 Screens

4 Screens

1 Screen



2 Screens

3 Screens

4 Screens

1 Screen

2 Screens

3 Screens

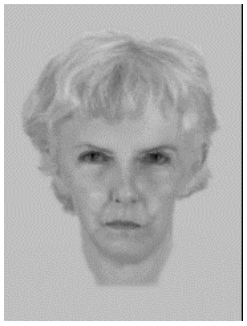
4 Screens

1 Screen

2 Screens

3 Screens

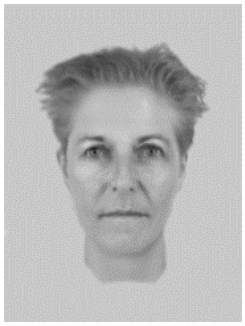
4 Screens



1 Screen

2 Screens

3 Screens



4 Screens

1 Screen

2 Screens

3 Screens

4 Screens

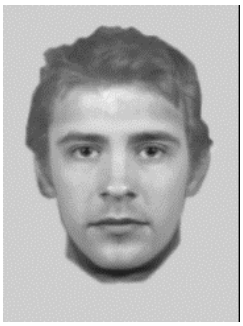
Appendix 4: Experiment 4 Targets and Composites

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

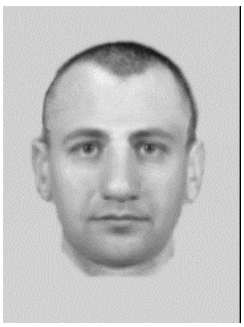


2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

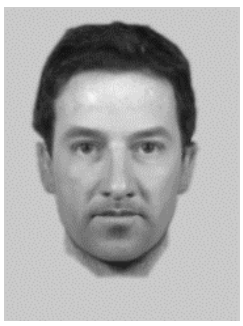


2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

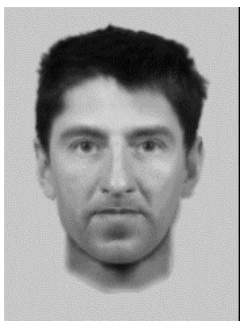
4 Shape
4 Texture



2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

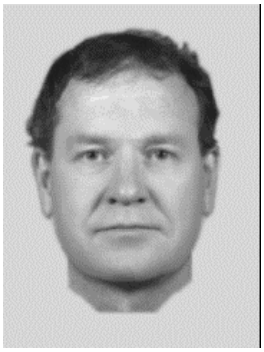


4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture



4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

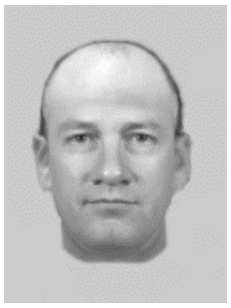
2 Shape
2 Texture

2 Shape
4 Texture

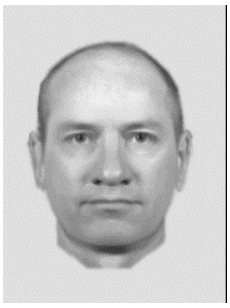
4 Shape
2 Texture

4 Shape
4 Texture

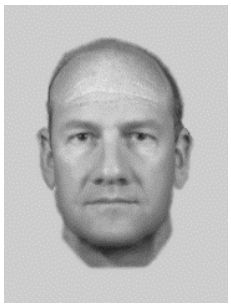
Appendix 5: Experiment 5 Targets and Composites



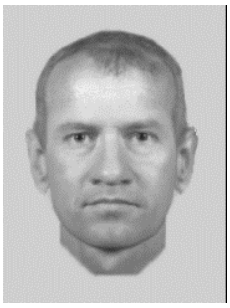
2 Shape
2 Texture



2 Shape
4 Texture



4 Shape
2 Texture



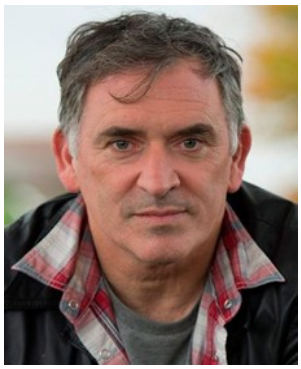
4 Shape
4 Texture

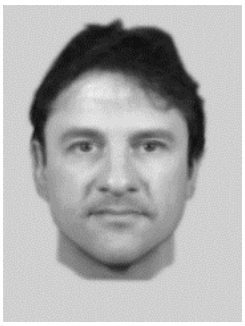
2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

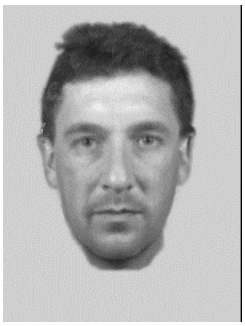
4 Shape
4 Texture





2 Shape
2 Texture

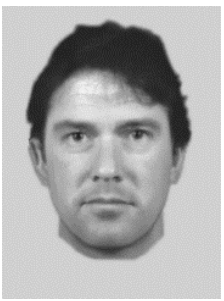
2 Shape
4 Texture



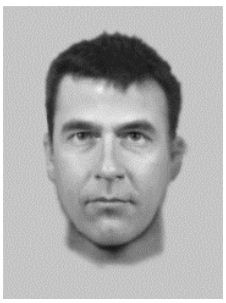
4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture



2 Shape
4 Texture



4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

2 Shape
2 Texture

2 Shape
4 Texture

4 Shape
2 Texture

4 Shape
4 Texture

Appendix 6: Verbal Recall Sheet

Overall

Shape

Hair

Eyebrows

Eyes

Nose

Mouth

Ears