The advantage of low and medium attractiveness for facial composite production from modern forensic systems

Beth Richardson (1*)

Charity Brown (2) psccbr@leeds.ac.uk

Priscilla Heard (3) priscilla.heard@uwe.ac.uk

Melanie Pitchford (4) melanie.pitchford@beds.ac.uk

Emma Portch (5) eportch@bournemouth.ac.uk

Karen Lander (6) karen.lander@manchester.ac.uk

John E. Marsh (1, 7) jemarsh@uclan.ac.uk

Raoul Bell (8) raoul.bell@hhu.de

Cristina Fodarella (1) cfodarella3@uclan.ac.uk

Ashley Taylor (9) colleen.taylor1@me.com

Mikaela Worthington (1) maworthington@uclan.ac.uk

Lauren Ellison (1) leellison@uclan.ac.uk

Philippa Charters (1) pippa_12@hotmail.com

Dannii Green (10) danniigreen1@hotmail.co.uk

Simra Minahil (1) simraminahil@hotmail.co.uk

Charlie D. Frowd (1) CFrowd1@uclan.ac.uk

(1) School of Psychology, University of Central Lancashire PR1 2HE UK

(2) School of Psychology, University of Leeds, Leeds LS2 9JT UK

(3) School of Psychology, University of the West of England, BS16 1QY UK

(4) Department of Psychology, University of Bedfordshire, Luton LU1 3JU

(5) Department of Psychology, Bournemouth University, Bournemouth, BH12 5BB, UK

(6) School of Psychological Sciences, University of Manchester, Manchester M13 9PL UK

(7) Department of Environmental Psychology, University of Gävle, 802 67, Gävle, Sweden

(8) Institut für Experimentelle Psychologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

(9) School of Humanities, Language and Global Studies, University of Central Lancashire PR1 2HE UK

(10) Department of Psychology, University of Winchester, Winchester SO22 4NR, UK

Recognition following long delays is superior for highly attractive and highly unattractive faces (cf. medium-attractive faces). In the current work, we investigated participants' ability to recreate from memory faces of low-, medium- and high physical attractiveness. In Experiment 1, participants constructed composites of familiar (celebrity) faces using the holistic EvoFIT system. When controlling for other variables that may influence face recognition (memorability, familiarity, likeability and age), correct naming and ratings of likeness were superior for composites of low attractiveness targets. Experiment 2 replicated this design using the feature based PRO-fit system, revealing superiority (by composite naming and ratings of likeness) for medium attractiveness. In Experiment 3, participants constructed composites of unfamiliar faces after a forensically-relevant delay of 1 day. Using ratings of likeness as a measure of composite effectiveness, these same effects were observed for EvoFIT and PRO-fit. The work demonstrates the importance of attractiveness for method of composite face construction.

Keywords: facial composite; facial attractiveness; witness; victim; EvoFIT; PRO-fit

A large body of work demonstrates how facial attractiveness influences recognition. Cross, Cross, and Daly (1971) show that when identifying unfamiliar faces, faces with higher attractiveness are recognised more successfully. Shepherd and Ellis (1973) incorporated an intermediate (medium) level of attractiveness, with results showing unfamiliar-face recognition did not vary by categorical attractiveness when tested immediately or following a six-day retention interval. However, following a five-week delay, recognition was worse for medium (vs. low/high) categories, suggesting an effect of encoding time. They argue that, due to the faces being unusual and/or memorable, participants experience heightened arousal upon viewing low and high (relative to medium) attractive faces. Vokey and Read (1992) also investigated the impact of attractiveness on face recognition and find two factors emerge as positive predictors of recognition; typicality which involves familiarity, likeability and attractiveness (i.e., 'context free' familiarity); and, memorability, a result supported elsewhere (e.g., Hancock, Burton, & Bruce, 1996; MacLin & MacLin, 2004; O'Toole, Deffenbacher, Valentin, & Abdi, 1993; Vokey & Read, 1995 but see Morris & Wickham, 2001).

These observations present an interesting prediction for identifications resulting from composites. In a forensic setting, composites are created by witnesses and victims to resemble a person (an offender) and are circulated in the hope that someone will offer a name, thereby helping to solve the crime. Composites are usually constructed using a 'feature' system (e.g., E-FIT, PRO-fit, FACES 4.0), whereby witnesses select individual facial features (eye, nose, etc.), or a 'holistic' system (e.g., EFIT-V, EvoFIT, ID), whereby they repeatedly select whole faces or face regions from arrays of alternatives[1]. For both, witnesses' ability to construct the face is reliant on their ability to remember and recall it, and later, to recognise that the constructed image has reached a good level of visual likeness (Frowd, Bruce, Smith, & Hancock, 2008). Until recently (see, Frowd, Skelton et al., 2013), images were recognised infrequently, and research suggested that the construction process may interfere with subsequent line up identification (Wells, Charman, & Olson, 2005). The Shepherd and Ellis (1973) study indicates that, under challenging conditions (i.e., stress, delay), recognition is inferior for people with medium-attractiveness, thus a similar effect would be expected for composites, especially when controlling for other properties of the face (cf. Vokey & Read, 1992).

Attractiveness and Holistic Processing

Attractiveness judgements are generally rapid and involve a range of cues including facial symmetry, shape and averageness (Little, Jones, & DeBruine, 2011). Faces possess *first-order relational properties*—referring to the basic configuration of features within the face; and *second-*

1 There are also manual (hand-drawn) sketch-based systems used by forensic practitioners that involve face construction based on selection of individual facial features (for more details, see Frowd, 2017).

order relational properties-referring to variations in spacing between and positioning of features (Diamond & Carey, 1986). First-order configuration is considered important for identifying a face, while second-order configuration is important for discriminating between individual faces (Diamond & Carey, 1986; Tsao & Livingstone, 2008). Normally, second-order configuration refers to the holistic nature of face processing (i.e., when making social judgements about the face). Indeed, research suggests that social judgements of faces, concerning their attractiveness for example, depends on holistic processing (Rhodes, 1988), and experimental methods that impair configural processing, such as scrambling top and bottom halves of faces (Abbas & Duchaine, 2008) and facial inversion (Santos & Young, 2008), directly impact judgements of attractiveness.

Conversely, it has been suggested that facial attractiveness positively relates to the mathematical averageness of the face (Piepers & Robins, 2012). Thus, perceptions of facial attractiveness significantly increase the more similar the face is to its group prototype. This leads to the notion of a 'face space' for more attractive faces, making it harder to differentiate individual faces and suggests a potential disadvantage for attractive faces created in holistic systems—as that process involves selection of whole faces or whole-face regions. Consequently, there may be an *advantage* for less attractive faces, since such identities are not bound by a group prototype.

In the current research, over the course of three experiments, using two modern production systems (holistic and feature-based), participants constructed composites of target faces with low, medium and high rated attractiveness, while controlling for factors such as memorability. It was anticipated that composites would be constructed more effectively (with higher correct naming and higher ratings of likeness) when created from faces with low- and high- relative to medium-level attractiveness. In particular, the holistic focus of EvoFIT construction might detract from specific facial features associated with attractiveness, and this effect could lead to a weaker overall benefit for composites constructed of attractive faces.

Method Experiment 1: Detailed investigation of an attractiveness effect for holistic face construction

Participants

Participants were 88 undergraduates, 62 female, from the University of Central Lancashire (UCLan), Preston. They received course credit for participating. Different participants were recruited for the four main stages of the experiment: 12 gave ratings to allow selection of target photographs, 24 constructed composites, 40 named composites and 12 rated composites. All of these participants were recruited on the basis of being familiar with famous faces. Participants involved in face

construction were allocated in three equal groups to the between-subjects factor, attractiveness type, with eight participants in each group. Twelve further participants, three male, were recruited from staff working at a small retail outlet in Winchester; they provided ratings of attractiveness for the target photographs.

Selection of Target Photographs

Materials

Targets were famous faces from a wide range of occupations. While famous faces are not usually involved in forensic practice, there is evidence that they produce composites with correct naming that is similar to non-famous stimuli in composite research (e.g., Brace, Pike, & Kemp, 2000; Bruce, Ness, Hancock, Newman, & Rarity, 2002; Frowd, Bruce, Ness et al., 2007; Frowd, Carson, Ness, Richardson et al., 2005; Frowd et al., 2015; Frowd, Jones et al., 2013). Since we planned to select targets by attractiveness category, while controlling for other attributes of the face (memorability, likeability and familiarity), sampling stimuli from a large pool size was necessary. Furthermore, to use composite naming as a primary dependent variable, we needed to ensure that stimuli were familiar to our participants. Such a design is difficult to achieve using other sets of faces.

A set of 42 good-quality colour photographs of male famous faces was located on the Internet on the basis that these identities would be familiar to undergraduates. Photographs depicted the person in a front-facing pose with a neutral expression, minimal facial hair, and without jewellery or glasses. Images were cropped fairly close to the head and were printed individually in colour to dimensions 8cm (wide) x 10cm (high).

Procedure

Twelve participants were presented sequentially with these photographs to rate, using Vokey and Read's (1992) scale definition (see above) and rating (1 = "sure"… 4 = "sure not"), first for how familiar they were with the face, and then in a different block order for ratings of memorability, likeability and attractiveness (with order of scale and photograph permutated randomly across participants). Participants were tested individually and worked at their own pace.

Participant ratings were reverse-coded to give a positive sense (now: 1 = low ... 4 = high). By attractiveness, the overall mean occurred approximately midway along the scale (M = 2.4), and there was good variability (SD = 0.7) and range ($1.0 < M < 3.6$) [(2)] of items. There was a slight negative skew (S = -0.3), indicating a tendency towards ratings of more attractive faces, appropriate for celebrities in general. Kurtosis was negative (K = -1.3) indicating a somewhat flat response curve. The set appeared to be a sensible basis for selecting stimuli.

2 For readers unfamiliar with this notation, it indicates a range of means (M)—here, from 1.0 to 3.6.

Three identities were not well-known (a familiarity rating less than the maximum for more than 25% of participants), and were discarded. For the remaining items, familiarity was very high (*M* = 3.8, *SD* = 0.2). Three groups of eight famous faces were selected by attractiveness category (low, medium and high) with the aim of each group maintaining rated memorability, familiarity and likeability. Based on previous research (e.g., Frowd et al., 2014), the number of items, taking into account the planned composite naming and likeness tasks, should provide sufficient experimental power to be able to detect a practically-useful medium effect size, should one exist. Example identities are shown later in Figure 1. Mean ratings by target (see Table 1) largely spanned the attractiveness scale ($1.2 < M < 3.6$). Target age ranged from 24 to 59 (*M* = 42.5, *SD* = 9.5) years; while not attempting to control for this variable, mean age was largely consistent across attractiveness categories ($39.8 < M(years) < 45.0$). As would be expected in general for well-known famous faces, mean rating was high for memorability (*M* = 3.8) and likeability (*M* = 3.4).

Table 1 about here

Individual rating scores (*N* = 1152) were analysed using Generalized Estimating Equations (GEE), a regression-type approach that is statistically more powerful than the popular ANOVA and provides a combined by-subjects and by-items model appropriate for the repeated observations of four sets of Likert ratings for each target photograph (Barnett, Koper, Dobson, Schmiegelow, & Manseau, 2009). The model used was full-factorial with an ordinal logistic response function and, based on homogenous correlations between elements, an Exchangeable Working Correlation Matrix. The subject variable was coded by participant number, and item number was a within-subject variable. All models were checked for multicollinearity. In this case, it was unlikely to be an issue as the predictors involved were not too highly correlated with each other: the only reliable inter-correlations ($p < .005$) were between memorability and familiarity [$r(22) = .63$], and between likeability and attractiveness [$r(22) = .59$]—these positive correlations have been reported in the literature (e.g., Rule, Slepian, & Ambady, 2012; Vokey & Read, 1992). Once built, Beta (*B*) and *SE(B)* values were checked to be within sensible bounds (not too low or too high) that might otherwise indicate a poor fit of the model. As the emerging *SE(B)* values indicated a stable fit of Beta values, a Model-based estimator was used for the Covariance Matrix.

The GEE model (see Table 1, *Note*) was significant for the three predictors contained in the model: rating scale [model fit $X^2(3) = 793.2$, $p < .001$], attractiveness category [$X^2(2) = 88.5$, $p < .001$] and rating scale x attractiveness category [$X^2(6) = 215.8$, $p < .001$]. Separate analyses using

parameter estimates revealed the following reliable differences by category with a medium effect size for the Odds Ratio [*Exp(B)* ≈ 2.5] (Sporer & Martschuk, 2014) for: (i) attractiveness, as planned, with high > medium [$p < .001$, *Exp(B)* = 2.0] and medium > low [$p < .001$, *Exp(B)* = 2.8], and (ii) likeability, with medium > low [$p < .001$, *Exp(B)* = 2.0] and high > low [$p < .001$, *Exp(B)* = 2.2]. While familiarity and memorability were successfully controlled for in the target set, likeability was not; this was presumably as people with faces rated as highly attractive tend to also be perceived as more likeable (see previous paragraph), and so likeability was taken into account (as a covariate) in the analysis for composite naming (see Results).[3]

Face Construction

Materials

EvoFIT was selected for Experiment 1 as this holistic system produces composites that are correctly named well after a forensically-relevant retention interval (Frowd, Pitchford et al., 2010). EvoFIT has been subject to considerable research and development (see Frowd, 2017 for a review) and is in current police use (e.g., Brown, Portch, Nelson, & Frowd, in press; Frowd et al., 2012).

Procedure

Each of the 24 participants were presented with a target photograph to construct, randomly selected. When asked, participants confirmed that the identity depicted was familiar[4], and then inspected the photograph for 60 seconds in the knowledge that they would later construct a composite of the target image. Next, each person worked with the experimenter, who was blind to target identity, with the aim of creating the best possible likeness. The procedure used to construct a face with EvoFIT is detailed (described in Fodarella, Kuivaniemi-Smith, Gawrylowicz, & Frowd, 2015); for the sake of brevity, we provide a summary here. First, participants received a short overview of the face-construction procedure. They then recalled the appearance of their target face in a free recall format (i.e., without interruption from the experimenter) and the experimenter recorded this information in written form. EvoFIT was started, and participants indicated the closest white-male database to match their target by age (20, 30, 40 or 50 years). They then selected a single item from about 300 alternatives for hair, ears and neck. These external features were presented blurred in

[3]We also collected data to compare categorical attractiveness ratings of these 24 targets with photographs of 12 persons, white males aged from about 18 to 60 years, who had been convicted of violent crimes. Analysis suggested that stimuli in the low-attractiveness category of the experiment were representative of convicted persons.

4 Another target face would have been randomly selected had any participants reported that the face was unfamiliar to them.

subsequent face arrays, to help participants select by internal features (the important central region of the face for composite naming, see below). Next, participants were shown four screens of 18 'smooth' faces, faces that changed by facial shape, and selected a total of six which appeared to match the target face best; they similarly selected six for facial texture, followed by a single item with the best overall likeness. Characteristics of these selected faces were combined ('bred' together) and participants selected again in the same way for smooth and texture faces. Participants then worked to improve the likeness using software scales for age, weight, attractiveness and seven other whole-face characteristics. Finally, participants were given the opportunity to manipulate the shape and placement of facial features. The final face was saved to disk. The procedure took about 45 minutes to complete, including the time for debriefing. See Figure 1 for example composites.

Figure 1 about here

Composite Naming and Rating

Materials

The 24 complete composites (Figure 1, top row, for examples) were printed individually in greyscale for participants to name and assign ratings. Composites can be difficult to name, leading to few correct names for some items and a reduction in experimental power. This issue was overcome by boosting experimental power through the use of additional data collected from a separate group of participants who were first shown the celebrity targets to name, to facilitate recognition for the set, and who then attempted to name the composites from their internal features (Figure 1, bottom row). This region of the face provides important cues when naming a photograph (Ellis, Shepherd, & Davies, 1979; Endo, Takahashi & Maruyama, 1984; Ge et al., 2008) or a facial composite (Frowd, Bruce, McIntyre, & Hancock, 2007; Frowd, Skelton, Butt, Hassan, & Fields, 2011).

Procedure

Forty participants, recruited from UCLan, evaluated the 24 composites using two naming tasks and a rating task. One group of participants (N = 22) named complete composites and then target photographs (to check that they were familiar with the relevant identities), while another group (N = 18) named target photographs first and then composites that comprised just internal-features. In both cases, participants inspected all composites (approx. 6cm wide x 8cm high; 300 x 300 pixels), and so attractiveness category (low, medium and high) was a within-participants' factor. To check the extent to which properties of the targets had been incorporated in the composites, further participants (N = 12) rated the composites for memorability, familiarity, attractiveness and likeability; composite ratings were also collected for likeness in the presence of the relevant target

photograph, itself a fairly-good proxy to correct naming of complete composites (e.g., Frowd, Bruce, Smith et al., 2008; Valentine, Davis, Thorner, Solomon, & Gibson, 2010). For the rating task, the design was within participants for both category (low, medium and high) and rating scale (memorability, familiarity, likeability, attractiveness and likeness).

In the main composite-naming task, 24 complete composites were presented. Participants were asked to name composites of well-known celebrities; they were told to guess if unsure but could also opt to not give a name. Next, the 24 target photographs were presented for naming. Composites and target photographs were presented in a different random sequential order for each participant. For the internal-features naming task, they named target photographs first followed by internal-features composites. A final set of participants gave ratings (1 = low … 4 = high), blocked by rating scale (familiarity, memorability, likeability and attractiveness) with a random order (in a Latin-Square design with equal sampling). Afterwards, participants rated the composites for likeness alongside the relevant target photograph, and then named the target photographs. Items within each block were presented in a different random order for each person.

Results

Composite Naming

Responses to target photographs were scored for accuracy: a value of 1 was assigned if the correct name of the famous person was given, and 0 for either a mistaken name or no name given. Out of a possible 960 responses, 830 targets were correctly named, suggesting that participants were very familiar with the relevant identities (M = 86.4%). When a target was not named correctly, this suggests that the participant could not have correctly named the associated composite; these cases were handled in the following analyses by scoring them as missing data (see Table 2).

The scoring procedure was applied in the same way to participant responses from complete and internal-features composites. Correct naming of complete composites ranged from 0% to 65%. Mean naming across items was 20.1% (SD = 21.3%), which is comparable to other research (17 < $M(\%)$ < 25) using a similar design with famous or non-famous stimuli (Brace et al., 2000; Davies, van der Willik, & Morrison, 2000; Frowd, Carson, Ness, Richardson et al., 2005; Frowd, Hancock, & Carson, 2004). The mean was much higher for internal-features composites (M = 35.7%, SD = 30.6%), a likely consequence of these images being named after participants had seen the target faces. These means were much lower than the mean for target pictures, but this is also the usual case as composites are error-prone stimuli and are more difficult to recognise than photographs (Frowd,

Carson, Ness, Richardson et al., 2005). As shown in Table 2, composite naming was clearly superior for both face types in the low- relative to medium- and high-rated attractiveness categories.

Table 2 about here

The impact of attractiveness category on composite accuracy was analysed together using naming data from both complete and internal-features. These data were combined because both dependent variables are useful indicators of the quality of the composites (Frowd & Hepton, 2009). The predictors were face type (between subjects: 1 = complete face and 2 = internal-features) and attractiveness category (within subjects: 1 = Low, 2 = Medium and 3 = High). GEE were run in the same way as before, although a binary-logistic function was used as the DV is dichotomous (0 = incorrect or no name, and 1 = correct name). As the role played by properties of the targets was potentially relevant, we included item means collected during target selection for familiarity, memorability and likeability as covariates in the model. Target age was also included, as this variable has been found to (negatively) correlate with facial attractiveness (Wickham & Morris, 2003)[5]. The backward approach was used, which commenced with a saturated model and removed variables sequentially (either predictor or covariate) with least contribution to naming (for $p > .1$, lowest X^2 is subject to removal). This process resulted in one predictor being removed, the interaction term, attractiveness x face type ($p = .55$).

The final model is summarised in Table 2. Attractiveness category was significant [$X^2(2) = 56.9$, $p < .001$] and parameter estimates indicted that composites were named more successfully for low than medium [$B = 1.3$, $SE(B) = 0.2$, $p < .001$, $Exp(B) = 3.7$, (95%) CI (2.5, 5.4)] and high categorical attractiveness [$B = 1.7$, $SE(B) = 0.2$, $p < .001$, $Exp(B) = 5.4$, CI (3.4, 8.5)]; a third contrast indicated that medium was superior to the high category [$B = 0.4$, $SE(B) = 0.2$, $p = .033$, $Exp(B) = 1.5$, CI (1.0, 2.1)]. Face type was also significant [$X^2(1) = 17.1$, $p < .001$], with naming higher for internal-features than complete composites [$B = 0.8$, $SE(B) = 0.2$, $p < .001$, $Exp(B) = 2.2$, CI (1.5, 3.2)]. Note that the non-significant interaction between attractiveness and face type indicates consistency of naming for complete and internal-features composites across attractiveness categories (and vice versa). Covariates ($p < .05$) led to a positive influence on naming for memorability [$B = 2.3$, $SE(B) = 0.7$, $Exp(B) = 9.9$], likeability [$B = 0.5$, $SE(B) = 0.2$, $Exp(B) = 1.6$] and age [$B = 0.05$, $SE(B) = 0.01$, $Exp(B) =$

5 These two variables were also negatively correlated in our data [$r(40) = -.34$, $p = .028$].

1.06], and negative for familiarity [*B* = -3.2, *SE(B)* = 0.9, 1/*Exp(B)* = 24.3]. We also scored participant responses for incorrect names given for complete composites, as these data can provide an indication of guessing, or response bias. Incorrect names were infrequent in total (*N* = 16) and were distributed roughly equally by condition. Due to low cell frequencies, these data were not subjected to inferential statistics.

Composite Rating

Mean ratings were calculated for each composite for the remaining rating scales. As composites should represent properties of target faces, mean ratings should be positively correlated between target and composite faces for attractiveness, memorability and likeability. These correlations were indeed positive, and reliable for attractiveness [$r(22) = .73$, $p < .001$] and likeability [$r(22) = .45$, $p = .027$], but not for memorability [$r(22) = .16$, $p = .46$]. Similarly, inter-correlations were carried out for composite ratings between familiarity, attractiveness, memorability and likeability. Two correlations were significant (and of similar magnitude and sign to those found for targets): between (i) memorability and familiarity [$r(22) = .44$, $p = .031$] and (ii) likeability and attractiveness [$r(22) = .53$, $p = .007$].

Lastly, we assessed the suitability of using composite likeness ratings as a proxy to correct naming (in part as we were planning to use likeness ratings as the DV for one of the experiments, see Experiment 3). This exercise was an attempt to replicate the current experiment using likeness ratings (cf. naming) as a measure of composite effectiveness. GEE were run again, this time with composite likeness ratings as DV. The pattern of significant results was the same as for correct naming: categorical attractiveness was significant [$X^2(2) = 26.5$, $p < .001$], with attractiveness category low > high [*B* = 1.0, *SE(B)* = 0.1, $p < .001$, *Exp(B)* = 2.7, CI (2.2, 3.4)], low > medium [*B* = 0.4, *SE(B)* = 0.1, $p < .001$, *Exp(B)* = 1.5, CI (1.2, 1.8)] and medium > high [*B* = 0.6, *SE(B)* = 0.1, $p < .001$, *Exp(B)* = 1.8, CI (1.5, 2.3)]. The previous four covariates (mean ratings of composites) were included in the model and all emerged significant (with a positive influence for familiarity, memorability and likeability, and negative for age)[6].

Discussion: Experiment 1

Composites were named more successfully when constructed of low relative to medium-attractiveness targets, and (to a lesser extent) medium relative to high-attractiveness targets. There

6 We also ran the GEE with likeness as DV without the presence of covariates (as categorical differences thereof can skew overall results), with the same significant outcome for attractiveness category.

was an advantage of naming for targets rated as more memorable, and a small advantage for targets rated as more likeable. Likeness ratings of composites were positively associated with composite naming for both complete and internal-features, suggesting more accurate-looking composites enjoyed higher correct naming (Frowd, Bruce, Smith, & Hancock, 2008). Memorability was a strong predictor of composite naming, replicating Frowd et al. (2005) for distinctiveness (an analogous measure to memorability) using celebrity faces. The same pattern was found between complete and internal-features composites, suggesting that naming is driven by accurate construction of internal features (e.g., Ellis et al., 1979; Frowd et al., 2011).

Other research using the archaic Photofit (Shepherd, Ellis, McMurran, & Davies, 1978) and the modern E-FIT feature-systems (Davies & Oldman, 1999) suggest more effective composites are created when constructors dislike the appearance of a target face; thus, we anticipated that likeability would be negatively related to correct naming. The opposite was found: a positive covariate exerting a small but reliable influence. This discrepancy seems related to attractiveness, since removing categorical attractiveness as a predictor produced the same pattern of significant results. It seems lower levels of attractiveness, rather than lower levels of likeability, are important for face construction. Following the same procedure as Experiment 1, Experiment 2 explored the impact of attractiveness with a more traditional 'feature' construction method. For simplicity, we did not select targets by likeability (as this measure proved difficult to equate across attractiveness groupings). To improve generalisability of results, we included university students as well as participants who were sampled more widely. As such, we increased the number of celebrity faces from which to select stimuli for face construction, given the likelihood that a more diverse sample of constructors would need a larger pool to select well-known faces by attractiveness category. We also increased the participant pool size by at least 50% for target selection and likeness ratings, anticipating these tasks would likewise require more experimental power.

Method Experiment 2: Assessment of an attractiveness effect for feature-based face construction

Participants

Eighty-six adults (59 female) participated. None of these participants were involved in Experiment 1. Of these, 20 gave ratings to allow selection of the target photographs, 24 constructed composites, 24 named composites and 18 rated composites. All participants were recruited on the basis of being familiar with the famous faces and were sampled widely from students at UCLan and

residents living in Manchester, UK. Participants involved in face construction were allocated to three equal groups for attractiveness type, with eight participants in each group.

Materials

A new set of celebrity faces was used as familiarity changes over time and we wanted to obtain identities for face construction that would be well-known. A set of 70 good-quality colour photographs of male famous faces was located on the Internet on the basis that these identities would be familiar to participants. Characteristics of these facial photographs matched those in Experiment 1.

Procedure

The procedure was also very similar, with 20 participants presented sequentially with photographs to rate (using the actual scale: 1 = low ... 4 = high), in a different block order for ratings of memorability and attractiveness (with order permutated randomly across participants), and for how familiar they were with the face. As before, participants were presented with photographs in a different random order, were tested individually and worked at their own pace.

As in Experiment 1, three groups of eight famous faces were selected by attractiveness category (low, medium and high). Mean ratings by target (Table 3) largely spanned the scale for attractiveness ($1.0 < M < 3.8$, Overall $M = 2.7$) and memorability ($1.7 < M < 3.5$, Overall $M = 2.7$), and were overall high for familiarity ($2.8 < M < 4.0$, Overall $M = 3.5$).

Table 3 about here

Individual rating scores (N = 1,440) were analysed using GEE (see, Experiment 1). The model was significant for the three predictors contained in the model: rating scale [$X^2(2) = 220.8$, $p < .001$], attractiveness category [$X^2(2) = 64.5$, $p < .001$] and rating scale x attractiveness category [$X^2(4) = 196.0$, $p < .001$]. Separate analyses revealed that attractiveness was reliable [$X^2(2) = 152.8$, $p < .001$], with high > medium [$p < .001$, *Exp(B)* = 8.9] and medium > low [$p < .001$, *Exp(B)* = 15.9]; neither memorability [$X^2(2) = 1.5$, $p = .47$] nor familiarity [$X^2(2) = 1.8$, $p = .41$] were reliable predictors.

Face Construction

Procedure

The procedure was the same as in Experiment 1 except for use of PRO-fit to construct the facial composites (Fodarella et al., 2015). Twenty-four participants received a short overview of the face-construction procedure and were asked to freely recall the appearance of their target face, with the experimenter writing down this information. PRO-fit was started, the white male database

selected, and the participant's description of the face was entered into PRO-fit, to provide appropriately 20 matching facial features for the participant to select, size and position on the face. The resulting face was saved to disk. The procedure took about 45 minutes to complete, including the time for debriefing. See Figure 2 for example composites constructed by attractiveness category.

Figure 2 about here

Composite naming and rating

Materials

The 24 complete composites were printed individually in greyscale for separate groups of participants to name and assign ratings of likeness (see *Participants*, and *Design and Materials*). In Experiment 1, composites were named from their internal-features region, as we anticipated that correct naming would be low; however, given that this DV turned out to be satisfactory, only complete composites were used for naming in Experiment 2. As both groups of participants inspected all composites, attractiveness category (low, medium and high) was a within-participants factor for both tasks.

Procedure

In the composite-naming task, 24 participants were asked to name composites of well-known celebrities; as before, they were asked to guess if unsure, or not to give a name. The 24 (complete) composites were presented and participants offered a name (or not). Next, the 24 target photographs were presented for naming. A second set of participants gave ratings (1 = low … 4 = high) for likeness alongside the photograph of the relevant identity, and then named the target photographs (the same as in the composite naming task itself). For both tasks, composites and target photographs were shown in a different random order of sequential presentation. Participants were tested individually and worked at their own pace.

Results

Composite Naming

A total of 429 out of 480 targets were named correctly, suggesting participants were familiar with the relevant identities (M = 89.4%). Once again, cases were removed where the target had not been named correctly. Correct naming of composites ranged from 0% to 78%, and mean naming across items was 18.8% (SD = 22.0%), comparable with that found in Experiment 1 (M = 20.1%). The

mean was fairly good for composites in the low-attractiveness category (M = 22.9, SD = 25.1), somewhat higher in medium (M = 28.6%, SD = 26.9), and both of these categories were much greater than composites in the high category (M = 10.4%, SD = 10.4).

Using the same model settings as Experiment 1, attractiveness was significant [$X^2(2) = 14.2$, $p = .001$], and parameter estimates indicted that composites were named more successfully for low than high-attractiveness [B = 0.9, $SE(B)$ = 0.3, $p < .001$, $Exp(B)$ = 2.5, CI (1.5, 4.2)] and for medium than high-attractiveness [B = 1.2, $SE(B)$ = 0.4, $p = .001$, $Exp(B)$ = 3.4, CI (1.7, 6.8)]; a third estimate revealed that low and medium were equivalent [B = 0.3, $SE(B)$ = 0.2, $p = .21$, $Exp(B)$ = 1.3, CI (0.8, 2.1)], suggesting an advantage for composites of low and medium relative to high categorical attractiveness. The advantage for medium-attractiveness did not emerge as significant but the low-attractiveness category appears to be roughly midway between medium and high, and so we ran GEE again, with attractiveness analysed by polynomial contrasts with categories entered in the order medium, low and high. This analysis was significant for linear [$X^2(1) = 15.4$, $p < .001$] but not quadratic [$X^2(1) = 1.3$, $p = .25$], indicating that composite naming was indeed best for medium-attractiveness, then low and then high.

GEE was run again including covariates for properties of the targets: age, and mean rated memorability and familiarity. Note that none of these covariates were strongly correlated with each other ($|r| < .25$). Attractiveness was significant [$X^2(2) = 22.3$, $p < .001$] and covariates were reliable for age [B = -0.03, $SE(B)$ = 0.01, $p = .022$, $Exp(B)$ = 1.03] but not for memorability [B = -0.01, $SE(B)$ = 0.1, $p = .51$, $1/Exp(B)$ = 1.1] and familiarity [B = 0.3, $SE(B)$ = 0.2, $p = .21$, $Exp(B)$ = 1.4]. This indicates that as target age increased, correct composite naming for PRO-fit reduced, the opposite to that found in Experiment 1—although for EvoFIT, results also indicated that memorability and familiarity *reliably* influenced correct naming[7].

Composite likeness ratings

[7] Mistaken names were also analysed. In Experiment 1, mistaken names were infrequent. For composites constructed here using PRO-fit, there were 77 mistaken names. The analysis proceeded by removing responses to composites (i) where the associated target was named incorrectly, as before, and (ii) where the composite itself was named correctly. The number of possible responses was 339, giving mean mistaken naming of 22.7% overall. GEE was conducted in the same way as for correct naming, with responses (coded here as 1 if mistaken name given, and 0 otherwise) but attractiveness category was not significant [$X^2(2) = 4.4$, $p = .11$]. This suggests that while participants mistakenly named composites about as frequently as they correctly named them, attractiveness category did not influence production of mistaken names.

As in Experiment 1, likeness ratings attributed to composites were analysed for targets that were named correctly (*M* = 91.7%). Mean rating was again best by categorical attractiveness for medium (*M* = 2.3, *SD* = 0.6), then low (*M* = 2.1, *SD* = 0.6) and high (*M* = 1.8, *SD* = 0.4). GEE were conducted as before, yielding the same significant outcome as for correct naming: categorical attractiveness was significant [$X^2(2) = 16.5$, $p < .001$], attractiveness category low > high [*B* = 0.5, *SE(B)* = 0.2, *p* = .013, *Exp(B)* = 1.7, CI (1.1, 2.6)] and (with a greater effect size) medium > high [*B* = 0.8, *SE(B)* = 0.2, *p* < .001, *Exp(B)* = 2.2, CI (1.5, 3.3)]; low and medium did not differ [*B* = 0.3, *SE(B)* = 0.2, *p* = .18, *Exp(B)* = 1.3, CI (0.9, 1.9)]. As polynomial contrasts are not available for multinomial DVs, ANOVA was run instead, which was significant by-participants[8] for attractiveness [$F(2,34) = 7.4$, $p = .002$, $\eta_p^2 = .30$; Mauchly's Test of Sphericity, $p = .84$], which was significant for a linear [$F(1,17) = 15.9$, $p = .001$, $\eta_p^2 = .48$] but not for a quadratic trend [$F(1,17) = 0.9$, $p = .37$, $\eta_p^2 = .05$]. So, the result for likeness ratings suggests the same ordinal relationship as for correct naming: medium > low > high categorical attractiveness[9].

Discussion: Experiment 2

Experiment 2 sought to understand further the effect of attractiveness on composite construction, this time using a feature-based system. Consistent with Experiment 1, composites of targets categorised as low and medium-attractiveness were both named more successfully than composites of targets categorised as high-attractiveness. However, unlike Experiment 1, composites of targets categorised as medium-attractiveness were better named than composites of targets categorised as low-attractiveness. The same pattern of results was found for both naming and likeness ratings. Before theorising as to likely reasons for differences in composite effectiveness by system, we attempt a replication in Experiment 3.

Experiments 1 and 2 involved face construction immediately after presentation of a picture of a celebrity. In forensic practice, an offender is sometimes a familiar identity (e.g., in cases of fraud), but usually the face is unfamiliar and construction occurs a day or two after the crime. This

[8] ANOVA, by-items, was not significant for attractiveness category ($F < 1$), a result that was not unexpected as we did not design the experiment to be analysed in this way using this type of analysis.

[9] *Additional analysis*. As in Experiment 1, a bi-variate correlation between correct naming and likeness was positive and very strong, $r(23) = .80$, $p < .001$, indicating once again the close relationship of these two measures for assessing composite effectiveness.

design was modelled in Experiment 3. If effects are largely independent of both identity type (celebrity or unfamiliar) and retention interval to construction (immediate or delayed), then the findings from Experiments 1 and 2 should be replicated in Experiment 3.

Method Experiment 3: Attempted replication plus extension

Participants

Participants comprised an opportunity sample of 82 undergraduate students, 65 female, recruited from the University of Central Lancashire. Ten students gave ratings to allow set up of the targets for face construction. These participants were volunteers, while course credit was offered to another 48 participants who each constructed a composite, and a further 24 participants who provided likeness ratings for the resulting composite images. For face construction, participants were allocated in equal groups to the two between-subjects factors, attractiveness type, eight per group, and composite system, 24 per group.

Selection of Target Faces

Materials

A set of 50 Caucasian faces was extracted from the Psychological Image Collection at Stirling (pics.stir.ac.uk). This image set comprises young adult males photographed in a front view pose depicting a neutral expression. We verified with participants that these faces were unfamiliar to them. The rating procedure was very similar to that used in the first two experiments. As before, ratings were collected from participants for attractiveness and memorability. However, as Experiment 1 revealed that likeability exerted only a weak effect on face construction at best, this variable was not included. Instead, another forensically-relevant variable was considered, trustworthiness. This property is known to be positively related to attractiveness (e.g., Todorov, Baron, & Oosterhof, 2008), and is interesting in the current context as confidence crimes tend to rely on trust: we let persons who appear to be trustworthy into our homes—even though it becomes apparent later that such a judgement was inaccurate. Trustworthiness has also been positively associated with a criminal stereotype (e.g., Flowe, 2012), and is negatively associated with memory performance (Rule et al., 2012) and other forensically relevant variables such as attribution of harsh sentencing (Wilson & Rule, 2015). This variable does not appear to have been formally investigated for composite face production. We anticipated that lower perceived trust would result in a better memory for the face and so a more effective composite.

Procedure

Ten participants rated the target faces. We again followed the same design and procedure for selection of target faces, resulting in eight items per category (see Table 4). We do note that

mean attractiveness emerged as slightly lower in each category than before, which is presumably the result of familiarity being lower for targets here relative to celebrity targets in Experiments 1 and 2. Using the same analysis as before with GEE, attractiveness emerged reliable as designed by category ($p < .001$, *Exp(B)* > 4.4). Target memorability was not significant by category ($p > .5$), indicating that this variable had been suitably controlled. However, higher levels of perceived trustworthiness are usually associated with higher levels of attractiveness (e.g., Todorov et al., 2008), as observed here. Accordingly, ratings emerged reliably higher by increasing category ($p < .001$, *Exp(B)* > 3.3). There was one reliable correlation, between attractiveness and trustworthiness [$r(23) = .62$, $p = .001$].

As before, the overall mean occurred approximately midway along the attractiveness scale ($M = 1.9$), and there was good variability ($SD = 0.6$) and range ($1.4 < M < 2.9$) of items; this time, there was a slight positive skew ($S = 0.4$), indicating a tendency towards ratings of less (vs. more) attractive faces, a sensible outcome for non-celebrities. Kurtosis was negative ($K = 1.5$) and indicates a somewhat flat response curve, similar to the celebrity set.

Table 4 about here

Composite Construction

Materials

Targets were printed in the same way as before; two sets were required, one set for construction using EvoFIT and PRO-fit composite systems and (later) another for rating.

Procedure

Forty-eight participants took part in this stage. Participants were allocated in equal groups to the two between-subjects factors, attractiveness type, eight per group, and composite system, 24 per group. The basic method of the first two experiments was followed except for two main differences. First, the impact of facial attractiveness was investigated for both EvoFIT and PRO-fit. Second, given evidence that likeness ratings are a fairly good proxy to correct naming of composites (Frowd, Bruce, Smith et al., 2008; Valentine et al., 2010; here, Experiments 1 and 2), we evaluated the resulting composites using a likeness-rating task. Please note, though, that while Experiment 3 involves contrasting face-production systems, we have avoided making comparisons between the two. This is due to concern that ratings of likeness may be influenced by properties of the image rather than by identity. For example, due to blending processes involved in the production of an EvoFIT face, this type of image has a tendency to look more “face like” compared with an image

from another system, potentially inflating ratings of likeness. This issue is unlikely to be involved in naming tasks, and readers interested in system comparisons using this measure are directed to other work (Frowd et al., 2015).

Participants inspected a randomly-selected target photograph for 60 seconds and, after 20 to 28 hours, described the face in a free-recall format and constructed a single composite using EvoFIT or PRO-fit. Participants were randomly assigned to composite system (EvoFIT or PRO-fit) and attractiveness type (low, medium and high) with the constraint that each target was constructed once for each system, to produce a total of 24 EvoFIT and 24 PRO-fit composites, respectively. Face recall and construction sessions took about 45 min, including debrief.

Composite Rating

The 48 composite faces constructed in the experiment were presented alongside the relevant target picture for participants to rate for overall likeness (1 = low … 4 = high), as before. Each of the 24 participants received a different random order of sequential presentation, and the task took approximately 20 minutes to complete, including debrief.

Results

Mean participant ratings for composite likeness are shown in Table 5. By composite system, mean likeness ratings (by items) spanned the range 1.3 to 3.3 (M = 2.1, SE = 0.1) for EvoFIT, and 1.4 to 3.6 (M = 2.4, SE = 0.2) for PRO-fit.

Table 5 about here

Participant rating scores were analysed as the DV using GEE. The two within-subject predictors were categorical attractiveness (1 = Low, 2 = Medium and 3 = High) and system (1 = EvoFIT and 2 = PRO-fit), and both were entered as a full-factorial model (incl. participant number as a subject variable, and item code as a within-subject variable). When the model was run, it was apparent that mean trustworthiness ratings, when included as a covariate, were substantially inflating differences between attractiveness categories. We did not observe such an effect in Experiment 1 for likeability, since the effect size was small, but it is larger in the current experiment for trustworthiness. The issue was addressed in two ways. First, a model was built using both system

and attractiveness as categorical variables but without covariates. This indicated superiority for low-attractiveness targets for EvoFIT and medium-attractiveness targets for PRO-fit. A second model involved system as a categorical variable but covariates (i.e., using mean target item values) for attractiveness, memorability and trustworthiness.

For the first model, as before, predictors were subject to sequential backward elimination (lowest X^2 for removal for $p > .1$). Categorical attractiveness was removed at Step 1 ($p = .47$). For the resulting model, system [$X^2(1) = 57.0$, $p < .001$] and system x attractiveness [$X^2(4) = 26.9$, $p < .001$] were reliable. For EvoFIT, we found the same benefit for the low-attractiveness category as in Experiment 1: low > medium [*B* = 0.5, *SE(B)* = 0.2, *Exp(B)* = 1.7, CI (1.2, 2.5)] and low > high [*B* = 0.3, *SE(B)* = 0.2, *Exp(B)* = 1.4, CI (1.0, 2.0)]; however, medium and high were equivalent [*B* = 0.2, *SE(B)* = 0.2, *Exp(B)* = 1.2, CI (0.9, 1.8)]. For PRO-fit, we found the same benefit for the medium-attractiveness category: medium > low [*B* = 0.8, *SE(B)* = 0.2, *Exp(B)* = 2.2, CI (1.6, 3.3)] and medium > high [*B* = 0.5, *SE(B)* = 0.2, *Exp(B)* = 1.7, CI (1.2, 2.4)]; low and high categories were equivalent [*B* = 0.3, *SE(B)* = 0.2, *Exp(B)* = 1.4, CI (0.9, 1.9)].

The second model involved system as a categorical variable, but now included covariates for mean target attractiveness, memorability and trustworthiness. Inter-correlations for these covariates as well as for mean rated likeness revealed positive correlations between ratings of trustworthiness and both attractiveness [$r(23) = .62$, $p = .001$] and likeness [$r(23) = .45$, $p = .024$]. The strength of these associations suggested that there should not be an issue of multicollinearity in the GEE. In addition, interaction terms were included for system and these three covariates. Memorability was removed in Step 1 ($p = .80$), memorability x system in Step 2 ($p = .61$) and system in Step 3 ($p = .30$). For the final model, attractiveness was a reliable covariate [$X^2(1) = 38.8$, $p < .001$], with an overall negative effect on likeness ratings [*B* = -1.0, *SE(B)* = 0.2, 1/*Exp(B)* = 2.0], as before, as was trustworthiness [$X^2(1) = 90.4$, $p < .001$], with an overall positive effect on likeness ratings [*B* = 1.4, *SE(B)* = 0.1, *Exp(B)* = 3.9]. However, system interacted with attractiveness and trustworthiness. For both systems, the slope (Beta coefficient) of the individual covariate was in the same direction as the main covariate but was stronger for PRO-fit than for EvoFIT [attractiveness: *B* = -1.0 vs. -0.5; trustworthiness: *B* = 1.4 vs. 0.5].

Discussion: Experiment 3

Experiment 3 replicated the attractiveness/system interaction; the lowest attractiveness category was best for EvoFIT, leading to a partial replication (since medium and high categories were

equivalent here[10]). Similarly, for PRO-fit, the medium-attractiveness category was best (and low and high categories did not differ reliably). Trustworthiness was positively related to composite likeness.

General Discussion

In Experiment 1, EvoFIT composites were more identifiable when constructed of targets from low relative to medium, and medium relative to high-attractiveness, even when controlling for other factors. In Experiment 2, composites from PRO-fit were more identifiable when constructed of targets from medium relative to low, and low relative to high-attractiveness. This pattern, indicating superiority of low-attractiveness targets for EvoFIT and superiority of medium-attractive targets for PRO-fit, was replicated in Experiment 3 using a more ecologically-valid design. Experiment 3 also revealed that targets with higher rated trustworthiness produced more effective composites for EvoFIT and, to a lesser extent, PRO-fit.

We argue that face construction with EvoFIT is based on similar processes to natural face recognition: a focus on the overall appearance of the face. We also argue that face construction is difficult to achieve accurately after long intervals (Frowd et al., 2015), and so predicted a benefit of low (vs. medium) attractiveness. This hypothesis was supported, with composites constructed from familiar targets evaluated by naming and likeness-rating tasks (Experiment 1), and composites constructed of unfamiliar targets evaluated via likeness-ratings (Experiment 3). Our result did not replicate findings for a benefit of medium (vs. high) attractiveness following a delay. Indeed, the high-attractiveness category emerged *less* effective than medium in Experiment 1. In Experiment 3, means were greater for high than medium but not significantly different. Similarly, an advantage of high (vs. low) categorical attractiveness (Cross et al., 1971), was not supported.

In Experiment 2, the effect of attractiveness on familiar face construction was again investigated, this time using the PRO-fit feature-system. Results from naming and likeness ratings indicated that the medium-attractiveness category was superior, a result obviously quite different to EvoFIT. This is important given past literature suggesting a disadvantage for medium-attractive or average faces. Our results suggest the potential for PRO-fit to overcome this disadvantage. These

10 We note that the lack of significant difference is unlikely to be caused by lower categorical target attractiveness (cf. Experiment 1). In Experiment 3, the curve for target attractiveness has essentially shifted to the left, the region of the scale where composites should be more identifiable, and the difference from medium to high target attractiveness is greater (*MD* = 0.7 vs. 0.9). If anything, the impact of medium to high-attractiveness should be greater in this experiment than the previous one.

patterns were replicated in Experiment 3 with unfamiliar-face construction. As the only intended difference in procedure was the construction method, this would appear to explain these results.

There are fundamental differences in construction between systems that may drive disparities in results. With EvoFIT, participants select whole regions from face arrays, a recognition task, then make adjustments to feature size and position, a procedure that benefits from recall of individual features. With PRO-fit, participants select facial features in the context of a complete face, again a recognition task, but one heavily dependent on participants' recall, and participants must suggest changes to the size and placement of features at the start, rather than the end. Evidence suggests that encoding by features (vs. holistically) leads to more identifiable composites (Frowd, Bruce, Ness et al., 2007; Wells & Hryciw, 1984) because it provides useful information to make more-informed choices when selecting individual features or whole-face regions from arrays, and when making adjustments to the shape and placement of features (Frowd, Bruce, Ness et al., 2007).

The mechanism for a low (vs. medium) attractiveness advantage for EvoFIT construction may be based on encoding differences. It is possible (perhaps from an evolutionary survival strategy) that unattractive faces capture more attention (due to perceived threat), thus increase encoding duration and facilitate recognition (Shapiro & Penrod, 1986). Alternatively, unattractive faces may increase arousal (also leading to an increase in encoding focus / effectiveness). Shepherd and Ellis (1973) proposed this mechanism to explain an advantage of extremes of attractiveness, but that account may only be relevant to the low end of the attractiveness scale. Factors that increase the duration and / or effectiveness of encoding may prompt a focus on individual features, a strategy useful for holistic- and feature-based construction.

We argue that a low-attractiveness target is valuable to facilitate encoding, thereby improving construction with EvoFIT. However, loss of facial information following a one-day delay may diminish benefit for feature-based construction. Instead, the feature-system is likely to be biased for more medial levels of attractiveness. Faces of this type are likely to be average (prototypical) in appearance (Potter & Corneille, 2008) so, over-represented in a normal population. Feature-systems like PRO-fit are therefore likely to have many more such exemplars of this type, thus providing a good variety of features for selection, even if the witness cannot describe the face in detail. The suggestion that attractive faces are characterised by their averageness would explain worse performance for high and medium (cf. low) attractive faces in EvoFIT: creating a face space of attractive faces makes it harder to differentiate faces (Pieper & Robins, 2003).

The strong positive correlation in Experiment 1 between target attractiveness ratings and composite naming suggests attractiveness translates to constructed faces. However, it does not pinpoint the mechanism responsible for the relationship between target encoding, long-term

representation of the target face, and subsequent composite construction. These questions could be addressed in a more ecologically-valid design to determine whether initial eyewitness judgements of attractiveness affect encoding and composite accuracy. Future research could also address questions about other aspects of the face that could convey a memory advantage, such as differences in attractiveness for female faces (Wei & Zhang, 2012; Zhang, Wei, Zhao, Zheng, & Zhang, 2016).

We also explored other properties that might be relevant to face production. Memorability was positively related to composite naming in Experiment 1, although not in Experiments 2 or 3. This is likely caused by higher mean memorability in Experiment 1: lower memorability may not be sufficient to produce a measurable effect with face construction, a process insensitive to psychological manipulation (Frowd, 2017). The positive benefit of memorability has been found in other composite research using celebrity faces (Frowd, Carson, Ness, Richardson et al., 2005), so effects appear reliable. Due to the potential relevance of this variable for real crimes, it could be a focus of future work. It is likely, as we show, that attractiveness and distinctiveness ratings are different. In line with the literature, our results suggest differential effects of distinctiveness and attractiveness on recognition. For example, Sarno and Alley (1997) show distinctive faces are better recognised. Attractiveness, in contrast, was a poor predictor of recognition, especially when distinctiveness was controlled for. This supports our benefit for low relative to medium and medium relative to high-attractiveness in Experiments 1 and 2.

Target familiarity was also a strong but negative covariate in Experiment 1 for composite naming. It is likely that high levels of familiarity are associated with better memory given that exposure to faces is the mechanism by which we naturally learn identity (Bruce, 1994; Frowd et al., 2014; Longmore, Liu, & Young, 2008). The knock-on effect means memories might interfere with one another during production of the face, reducing performance. Indeed, some participants reported that facial images other than their previously-seen celebrity came to mind during construction: in contrast, less well-known faces should involve less influence from competing memories. Familiarity emerged as a *positive* covariate in the analysis for likeness ratings, and so effects may not be as clear-cut—although assessing composites using likeness ratings is only a proxy to composite naming.

Experiment 3 considered the influence of trustworthiness, another forensically-relevant variable. For confidence crimes, a victim may assume that a person is honest, in part due to perceptions of trustworthiness. Indeed, such persons are likely to be successful if they possess an attractive face and appear trustworthy (Todorov et al., 2008). As categorical trustworthiness could not be controlled in the target stimuli, we ran a separate GEE using attractiveness as a covariate. While attractiveness remained negatively related to composite likeness across both composite

systems, trustworthiness was positively related, and so offenders with a trusted visage produce more effective composites.

For EvoFIT, there is evidence for an advantage of trustworthiness in police data of real crimes: Frowd et al. (2012) reported that EvoFIT composites were particularly effective at identifying offenders of distraction burglary. Experiment 3 also revealed a relatively stronger benefit of trustworthiness (as well as attractiveness) for PRO-fit (cf. EvoFIT). It would seem sensible to conclude that the feature-based method, while tending to create unidentifiable faces in forensic practice (Frowd, Hancock et al., 2010) and the laboratory (Frowd et al., 2015), may be more responsive to properties of a target individual. More generally, it is worth mentioning that the existing literature indicates a benefit to recognition memory for faces perceived as untrustworthy (Rule et al, 2012). While this finding relates to non-error prone stimuli such as facial photographs, mechanisms for faces constructed from memory may be more complex and warrant further investigation, particularly given differing effects of trustworthiness using different face production methods, as found here, and knowledge that perceived trustworthiness appears to be modulated by a person's behaviour (Suzuki & Suga, 2010).

In practical terms, the research not only allows police practitioners to more accurately gauge the effectiveness of their composites based on target attractiveness, but also provide a theoretically grounded understanding of factors affecting composite images constructed and assessed using simulated real-life procedures. The feature-system is likely to be biased for more medial levels of attractiveness. Faces of this type are likely to be average (prototypical) in appearance (Potter & Corneille, 2008) and so over-represented in a normal population of faces (Valentine, 1991). PRO-fit is therefore likely to have many more such exemplars, thus providing a good variety of features for a participant to select, even when the face cannot be described in detail (as is generally the case following an overnight delay). By further establishing biases implicit in composite systems, it may be possible to use inbuilt tools to control for these—such as implementing holistic tools in PRO-fit for use with non-average faces, ideally to manipulate faces away from the average. In other words, the effects of attractiveness may not relate to preferential encoding/processing of the faces, but rather by the features/faces offered by a composite system biased towards certain levels of attractiveness. The work also revealed that perceived trustworthiness of a target was positively related to a composite's effectiveness, although the effect seemed less robust for PRO-fit, the feature-system. This is good news for the criminal justice system as offenders who commit confidence crimes are likely to be more readily identified from faces constructed using a holistic system such as EvoFIT.

References

Barnett, A.G., Koper, N., Dobson, A.J., Schmiegelow, F., & Manseau, M. (2009). Selecting the correct variance–covariance structure for longitudinal data in ecology: A comparison of the Akaike, quasi-information and deviance information criteria, *http://eprints.qut.edu.au/19195/.*

Brace, N., Pike, G., & Kemp, R. (2000). Investigating E-FIT using famous faces. In A. Czerederecka, T. Jaskiewicz-Obydzinska & J. Wojcikiewicz (Eds.). *Forensic Psychology and Law* (pp. 272-276). Krakow: Institute of Forensic Research Publishers.

Brown, C., Portch, E., Nelson, L., & Frowd, C. D. (in press). Re-evaluating the role of verbalisation of faces for composite production: Descriptions of offenders matter! *Journal of Experimental Psychology: Applied*.

Bruce, V. (1994). Stability from variation: The case of face recognition. The M.D. Vernon Memorial Lecture. *Quarterly Journal of Experimental Psychology, 47A*, 5–28.

Bruce, V., Ness, H., Hancock, P.J.B., Newman, C., & Rarity, J. (2002). Four heads are better than one. Combining face composites yields improvements in face likeness. *Journal of Applied Psychology*, 87, 894-902.

Cash, T.F., & Janda, L.H. (1984). The eye of the beholder: Attractive people are preferred for dates, jobs and friendships: but there are some thorns on the rose of beauty. *Psychology Today*, *18*, 46–52.

Cross, J. F., Cross, J., & Daly, J. (1971). Sex, race, age, and beauty as factors in recognition of faces. *Perception & Psychophysics, 10*, 393-396.

Davies, G.M., & Oldman, H. (1999). The impact of character attribution on composite production: A real world effect? *Current Psychology: Developmental, Learning, Personality, Social*. 18, 128-139.

Davies, G.M., van der Willik, P., & Morrison, L.J. (2000). Facial Composite Production: A Comparison of Mechanical and Computer-Driven Systems. *Journal of Applied Psychology, 85*, 119-124.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*, 285–290.

Eagly, A.H., Ashmore, R.D., Makhijani, M.G., & Longo, L.C. (1991). What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, 110, 109–128.

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

Ellis, H.D., Shepherd, J. W., & Davies, G.M. (1980). The deterioration of verbal descriptions of faces over different delay intervals. *Journal of Police Science and Administration*, *8*, 101-106.

Endo, M., Takahashi, K., & Maruyama, K. (1984). Effects of observer's attitude on the familiarity of faces: Using the difference in cue value between central and peripheral facial elements as an index of familiarity. *Tohoku Psychologica Folia*, *43*, 23-34.

Fink, B., & Penton-Voak, I.S. (2002). Evolutionary psychology of facial attractiveness. *Current Directions in Psychological Science*, *11*, 154-158.

Flowe, H. D. (2012). Do characteristics of faces that convey trustworthiness and dominance underlie perceptions of criminality? *PLoS ONE*, DOI: 10.1371/journal.pone.0037253.

Flowe, H., & Humphries, J.E. (2010). An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology*, *25*, 265-273.

Fodarella, C., Kuivaniemi-Smith, H., Gawrylowicz, J., & Frowd, C.D. (2015). Forensic procedures for facial-composite construction. *Journal of Forensic Practice*, *17*, 259-270.

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: New York.

Frowd, C.D. (2015). Facial composites and techniques to improve image recognisability. In T. Valentine, & J. Davis (Eds.) *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and cctv* (pp. 43 - 70.) Chichester:Wiley-Blackwell.

Frowd, C.D., Bruce, V., & Hancock, P.J.B. (2008). Changing the face of criminal identification. *The Psychologist, 21*, 670-672.

Frowd, C.D., Bruce, V., McIntyre, A., & Hancock, P.J.B. (2007). The relative importance of external and internal features of facial composites. *British Journal of Psychology, 98*, 61-77.

Frowd, C.D., Bruce, V., Ness, H., Bowie, L., Thomson-Bogner, C., Paterson, J., McIntyre, A., & Hancock, P.J.B. (2007). Parallel approaches to composite production. *Ergonomics, 50*, 562-585.

Frowd, C.D., Bruce, V., Smith, A., & Hancock, P.J.B. (2008). Improving the quality of facial composites using a holistic cognitive interview. *Journal of Experimental Psychology: Applied, 14*, 276 - 287.

Frowd, C.D., Carson, D., Ness, H., McQuiston, D., Richardson, J., Baldwin, H., & Hancock, P.J.B. (2005). Contemporary Composite Techniques: The impact of a forensically-relevant target delay. *Legal & Criminological Psychology, 10*, 63-81.

Frowd, C.D., Carson, D., Ness, H., Richardson, J., Morrison, L., McLanaghan, S., & Hancock, P.J.B. (2005). A forensically valid comparison of facial composite systems. *Psychology, Crime & Law, 11*, 33-52.

Frowd, C.D., Erickson, W.B., Lampinen, J.L., Skelton, F.C., McIntyre, A.H., & Hancock, P.J.B. (2015). A decade of evolving composite techniques: regression- and meta-analysis. *Journal of Forensic Practice, 17*, 319-334.

Frowd, C.D., Hancock, P.J.B., Bruce, V., McIntyre, A., Pitchford, M., Atkins, R., et al. (2010). Giving crime the 'evo': catching criminals using EvoFIT facial composites. In G. Howells, K. Sirlantzis, A. Stoica, T. Huntsberger and A.T. Arslan (Eds.) *2010 IEEE International Conference on Emerging Security Technologies* (pp. 36 - 43). ISBN 978-0-7695-4175-4.

Frowd, C.D., Hancock, P.J.B., & Carson, D. (2004). EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on Applied Psychology (TAP)*, *1*, 1-21.

Frowd, C.D., & Hepton, G. (2009). The benefit of hair for the construction of facial composite images. *British Journal of Forensic Practice, 11*, 15-25.

Frowd, C.D., Jones, S., Forarella, C., Skelton, F.C., Fields, S., Williams, A., Marsh, J., Thorley, R., Nelson, L., Greenwood, L., Date, L., Kearley, K., McIntyre, A., & Hancock, P.J.B. (2013). Configural and featural information in facial-composite images. *Science & Justice, DOI: 10.1016/j.scijus.2013.11.001*.

Frowd, C.D., Pitchford, M., Bruce, V., Jackson, S., Hepton, G., Greenall, M., McIntyre, A., & Hancock, P.J.B. (2010). The psychology of face construction: Giving evolution a helping hand. Applied *Cognitive Psychology, 25*, 195–203.

Frowd, C.D., Pitchford, M., Skelton, F., Petkovic, A., Prosser, C., & Coates, B. (2012). Catching Even More Offenders with EvoFIT Facial Composites. In A. Stoica, D. Zarzhitsky, G. Howells, C. Frowd, K. McDonald-Maier, A. Erdogan, and T. Arslan (Eds.) *IEEE Proceedings of 2012 Third International Conference on Emerging Security Technologies* (pp. 20 - 26). DOI 10.1109/EST.2012.26.

Frowd, C.D., Skelton, F., Butt, N., Hassan, A., & Fields, S. (2011). Familiarity effects in the construction of facial-composite images using modern software systems. *Ergonomics*, *54*, 1147-1158.

Frowd, C.D., Skelton F., Hepton, G., Holden, L., Minahil, S., Pitchford, M., McIntyre, A., Brown, C., & Hancock, P.J.B. (2013). Whole-face procedures for recovering facial images from memory. *Science & Justice, 53*, 89-97.

Frowd, C.D., White, D., Kemp, R.I., Jenkins, R., Nawaz, K., & Herold, K. (2014). Constructing faces from memory: The impact of image likeness and prototypical representations. *Journal of Forensic Practice*, *16*, 243-256

Ge, L., Anzures, G., Wang, Z., Kelly, D.J., Pascalis, O., Quinn, P.C., Slater, A.M., Yang, Z., & Lee, K. (2008). An inner face advantage in children's recognition of familiar peers. *Journal of Experimental Child Psychology*, *101*, 124–136.

Haist, F., Shimamura, A.P., & Squire, L. R. (1992). On the relationship between recall and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 691-702.

Hancock, P.J., Burton, A.M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition, 24*, 26-40.

Hill, M.K., & Lando, H.A. (1976). Physical attractiveness and sex-role stereotypes in Impression formation. *Perceptual & Motor Skills, 43*, 1251-1255.

Jackson, L.A., Hunter, J.E., & Hodge, C.N. (1995). Physical attractiveness and intellectual competence: A meta-analytic review. *Social Psychology Quarterly, 58*, 108–122.

Little, A.C., Jones, B.C., & DeBruine, L.M. (2011). Facial attractiveness: Evolutionary based research. *Phil. Trans. R. Soc. B*, *366*, 1638–1659.

Longmore, C.A., Liu, C.H. & Young, A.W. (2008). Learning Faces From Photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 77–100.

MacLin, O.H., & MacLin, M.K. (2004). The effect of criminality on face attractiveness, typically, memorability and recognition. *North American Journal of Psychology, 6*, 145-154.

Moore, F.R., Filippou, D., & Perrett, D.I., (2011). Intelligence and attractiveness in the face: Beyond the attractiveness halo effect. *Journal of Evolutionary Psychology, 9*, 205-217.

Morris, P.E., & Wickham, L.H.V. (2001). Typicality and face recognition: A critical re-evaluation of the two-factor theory. *Quarterly Journal of Experimental Psychology, 54A*, 863–877.

Morrow, P.C., McElroy, J.C., Stamper, B.G., & Wilson, M.A. (1990). The effects of physical attractiveness and other demographic characteristics on promotion decisions. *Journal of Management, 16*, 723-736.

Nisbett, R.E, & Wilson, T.D. (1977).Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

O'Toole, A.J., Deffenbacher, K.A., Valentin, D., & Abdi, H. (1993). Structural aspects of face recognition and the other race effect. *Memory & Cognition, 22*, 208-224.

Piepers, D., & Robbins, R. (2012). A review and clarification of the terms "holistic,""configural," and "relational" in the face perception literature. *Frontiers in psychology*, *3*, 559.

Potter, T., & Corneille, O. (2008). Locating attractiveness in the face space: Faces are more attractive when closer to their group prototype. *Psychonomic Bulletin & Review, 15*, 615-622.

Rule, N.O., Slepian, M.L., & Ambady, N. (2012). A memory advantage for untrustworthy faces. *Cognition, 125*, 207-218.

Shapiro, P.N., & Penrod, S.D. (1986). Meta-analysis of facial identification rates. *Psychological Bulletin*, *100*, 139-156.

Shepherd, J.W., & Ellis, H.D. (1973). The effect of attractiveness on recognition memory for faces. *American Journal of Psychology, 86*, 627-633.

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*, *8*, 263-8.

Sporer, S.L. & Martschuk, N. (2014). The Reliability of Eyewitness Identifications by the Elderly: An Evidence-based Review. In (Eds.) Michael P. Toglia, David F. Ross, Joanna Pozzulo, Emily Pica. *The Elderly Eyewitness in Court*. Psychology Press: New York.

Suzuki, A.L., & Suga, S. (2010). Enhanced memory for the wolf in sheep's clothing: facial trustworthiness modulates face-trait associative memory. *Cognition, 117*, 224-229.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social Cognitive and Affective Neuroscience*, *3*, 119–127.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology, 43A*, 161-204.

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, 72 – 86.

Vokey, J.R., & Read, J.D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20*, 291-302.

Vokey, J.R., & Read, J.D. (1995). Memorability, familiarity, and categorical structure in the recognition of faces. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition: Explorations in face-space*. London: Routledge.

Wei, B., & Zhang, Y. (2012). Event-related potentials (ERPs) to gender differences in encoding processing for female facial attractiveness. *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2675–2678.

Wells, G.L., & Hryciw, B. (1984). Memory for faces: Encoding and retrieval operations. *Memory & Cognition*, *12*, 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.

Wickham, L.H.V., & Morris, P.E. (2003). Attractiveness, distinctiveness, and recognition of faces: Attractive faces can be typical or distinctive but are not better recognized. *American Journal of Psychology*, *116*, 455-468.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal sentencing outcomes. *Psychological Science, 26*, 1325-1331.

Yarmey, A.D. (2004). Eyewitness recall and photo identification: A field experiment. *Psychology, Crime & Law, 10*, 53-68.

Zhang, Y., Wei, B., Zhao, P., Zheng, M., & Zhang, L. (2016). Gender differences in memory processing of female facial attractiveness: evidence from event related potentials, *Neurocase*, *22*, 317-323.

List of figures and tables

Figure 1. Example composites constructed in Experiment 1 by face-constructor participants using EvoFIT; along with 21 other composites, these images were given to further participants to name and rate for likeness. From left to right, they are of comedian Dara O'Briain (low-rated attractiveness category), actor Colin Firth (medium-) and singer/songwriter Peter Andre (high-). Naming involved complete composites (top row), as constructed by participants, and composites of internal features (bottom row).

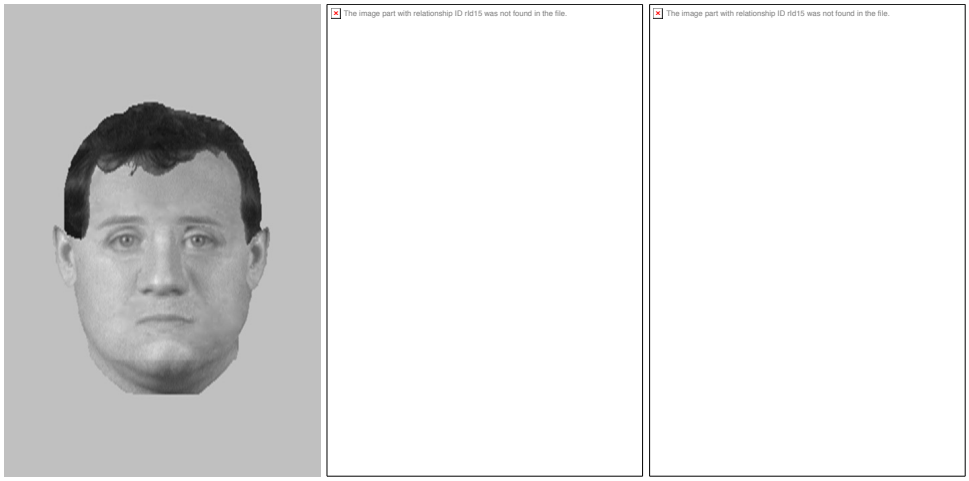

Figure 2. Example composites constructed in Experiment 2 by constructor participants using the PRO-fit system. From left to right, they are of comedian Peter Kay (low-rated attractiveness category), musician Ed Sheeran (medium-) and former footballer David Beckham (high).

Table 1. Mean Likert ratings (1 = low … 4 = high) of selected targets by rating scale. There are eight different target faces in each category (low, medium and high)

	Rating scale			
Category	Attractiveness	Familiarity	Memorability	Likeability
Low	1.5* (0.2)	3.9 (0.1)	3.7 (0.2)	2.7† (0.6)
Medium	2.6* (0.2)	3.9 (0.1)	3.7 (0.2)	3.3 (0.2)
High	3.3* (0.2)	3.8 (0.2)	3.7 (0.2)	3.4 (0.3)

Note. Figures in parentheses are (by-item) standard deviations. Model details for these data: Generalized Estimating Equations' Goodness of Fit, *QIC* = 592.4, and intercept [Beta (gradient) coefficient, *B* = 1.3; standard error of *B*, *SE(B)* = 0.06; Odds ratio (Effect Size), *Exp(B)* = 3.7]. *Significantly different to each other, $p < .001$. †Significantly different to the two other categories for likeability, $p < .001$.

Table 2. Correct naming of composites by face type and target attractiveness category.

Face type	*Attractiveness category*		
	Low*	Medium*	High*
Complete[§]	31.2 (49 / 157)	18.2 (25 / 137)	15.8 (21 / 133)
Internal features[§]	49.6 (68 / 137)	26.9 (35 / 130)	27.9 (38 / 136)

Note. Figures are percentage-correct accuracy calculated from responses in parentheses: summed correct responses (numerator) and total (correct and incorrect) responses (denominator). These data relate to composites for which participants correctly named the relevant target (*N* = 830 out of 960). GEE model parameters for these data: *QIC* = 916.4 and intercept [coefficient *B* = -1.3, *SE(B)* = 1.0, *p* = .037, *Exp(B)* = 0.11]. **p* < .05. [§]*p* < .001. See text for more details.

Table 3. Mean Likert ratings (low = 1 … high = 4) by category (low, medium and high) of selected targets by attractiveness, memorability and familiarity.

Category	*Rating scale*		
	Attractiveness	Memorability	Familiarity
Low	1.1* (0.1)	2.8 (0.7)	3.5 (0.5)
Medium	2.6* (0.6)	2.8 (0.7)	3.5 (0.4)
High	3.5* (0.5)	2.7 (0.3)	3.6 (0.3)

Note. Figures in parentheses are (by-item) standard deviations. **p* < .001.

Table 4. Mean Likert ratings (low = 1 … high = 4) of selected targets by attractiveness, memorability and trustworthiness rating scale and category (low, medium and high).

	Rating scale		
Category	Attractiveness	Memorability	Trustworthiness
Low	1.3* (0.03)	2.0 (0.2)	1.8§ (0.3)
Medium	1.8* (0.04)	2.0 (0.2)	2.2§ (0.2)
High	2.7* (0.04)	2.0 (0.2)	2.9§ (0.2)

Note. Figures in parentheses are (by-item) standard deviations. *$p < .001$. §$p < .001$. See text for details.

Table 5. Mean composite likeness ratings (1 = low … 4 = high) by target attractiveness category and facial composite system.

	Composite system	
Attractiveness category	EvoFIT	PRO-fit
Low	2.28[a,b] (0.25)	2.38[c] (0.28)
Medium	2.01[a] (0.19)	2.77[c,d] (0.23)
High	2.10[b] (0.17)	2.51[d] (0.30)

Note. Figures in parentheses are standard errors of item means. Generalized Estimating Equations' Goodness of Fit Thresholds for these data [R = 1.0, B = -1.9, R = 2.0, B = -0.1, R = 3.0, B = 1.8]. [a,b,c,d] p < .1. See text for details.